



This work is protected by copyright and other intellectual property rights and duplication or sale of all or part is not permitted, except that material may be duplicated by you for research, private study, criticism/review or educational purposes. Electronic or print copies are for your own personal, non-commercial use and shall not be passed to any other individual. No quotation may be published without proper acknowledgement. For any other use, or to quote extensively from the work, permission must be obtained from the copyright holder/s.

Application of data science to inform surface  
engineering for *in vitro* neural stem cell control

Georghios Joseph

Doctor of Philosophy in Regenerative Medicine

October 2018

Keele University

## ABSTRACT

---

The interest in the clinical use of stem cell therapies is increasing rapidly, with a need for more control over cell populations cultured/expanded *in vitro*. This is particularly relevant for the treatment of neurological disorders such as Parkinson's disease where positive outcome measures of clinical trials will be limited by the number of derived neurons and their specific sub-types. The aim is to generate enhanced neural cell populations from stem cells through the design of the cell-material interface.

The niche micro-environment is complex, being responsible for cell attachment, proliferation and differentiation. Material engineering approaches to better control cell responses have looked towards surface chemical, topographical and mechanical cues. The many permutations of these factors pose a major challenge in the optimisation of biomaterial design. Machine learning techniques will be used to assess the impact of surface properties on the biological micro-environment.

Cell interaction/response provides computational outputs, with input variables being derived from material properties such as surface chemical characteristics (logP, charge, density, wettability, etc.) and topography (nano- and micro-scale, aspect ratio, etc). The aim is to unravel the relationship between cells and biomaterial surface of *in vitro* cell culture. *In vitro* experiments and *in silico* modelling will continually inform each other towards the optimisation of neural cell characteristic responses.

## KEYWORDS

---

Neural stem cells, nerve tissue engineering, silanes, machine-learning, predictive modelling, mathematical optimisation

# TABLE OF CONTENTS

---

Abstract.....	ii
Keywords.....	ii
Table of Contents.....	iii
Acknowledgements .....	vii
1 Introduction .....	1
1.1 Neurodegenerative diseases.....	1
1.2 The need for tissue engineering .....	3
1.3 Regenerative medicine: biomaterials .....	4
1.3.1 Stem cell therapies .....	5
1.3.2 Cell differentiation complexity .....	9
1.3.3 Optimal cell culture methodologies .....	11
1.3.4 Biomaterials for <i>in vitro</i> cell culture .....	14
1.3.5 Surface-cell interaction.....	19
1.3.6 Surface-protein interaction .....	23
1.3.7 Presenting chemistry of biomaterials.....	25
1.4 Computational approaches.....	27
1.4.1 Data science and overlapping sub-fields .....	27
1.4.2 Relevant literature .....	29
1.5 Problem definition .....	35
1.6 Aims and objectives .....	37
2 Materials and methods.....	39
2.1 Modifying presenting chemistry of surfaces .....	39
2.2 Surface characterisation .....	43
2.2.1 Contact angle measurements (CAMs).....	43
2.2.2 Surface Enhanced Raman Spectroscopy (SERS) .....	44
2.2.3 X-ray Photoelectron Spectroscopy (XPS).....	45
2.3 Cell culture on modified surfaces .....	46
2.3.1 Neural tissue dissection.....	46
2.3.2 Neurosphere expansion.....	48
2.3.3 Neurosphere passage .....	49
2.3.4 Neurospheres micro-culture.....	50
2.4 Fixation and immunocytochemistry .....	51

2.5	Microscopy .....	53
2.5.1	Epi-fluorescence .....	53
2.5.2	Morphological cell performance measurements .....	54
2.6	Statistical Analyses .....	55
2.6.1	Variance tests.....	55
2.6.2	Data distribution .....	56
2.6.3	Correlation .....	58
2.7	Data mining and machine learning .....	64
2.7.1	Data collection and aggregation .....	64
2.7.2	Dataset selection, cleaning, and pre-processing .....	65
2.8	Model selection.....	66
2.8.1	Model performance .....	68
2.8.2	Attribute evaluation and selection .....	69
2.8.3	Cell cluster area .....	70
2.8.4	Neuron proportion.....	72
2.8.5	Type I astrocyte proportion .....	74
2.8.6	Type II astrocyte proportion .....	76
2.8.7	Proportion of unknown type cells .....	79
2.8.8	Neurite length .....	80
2.8.9	Type I astrocyte area .....	80
2.8.10	Astrocyte fibre length .....	83
2.9	Computational cell culture experiments .....	85
2.9.1	Generating test cases.....	86
2.9.2	Ranking method.....	87
2.9.3	Storing results .....	89
2.9.4	Numerical chemistry conversion .....	91
2.9.5	Reassessing converted chemistries .....	91
3	Discovering relationships computationally .....	92
3.1	Introduction .....	92
3.1.1	Previous work data .....	93
3.2	Results .....	95
3.2.1	Variance tests.....	95
3.2.2	Distribution .....	97
3.2.3	Correlation and visual relationships .....	106

3.2.4	LogP correlations .....	137
3.3	Discussion.....	148
3.3.1	Cell cluster area (CCA).....	148
3.3.2	Neuron density (ND) .....	149
3.3.3	Astrocyte density (AD) .....	151
3.3.4	Neuron/Astrocyte ratio (NAR) .....	153
3.3.5	Neuron axon length (NAL) .....	154
3.3.6	Astrocyte fibre length (AFL) .....	155
3.4	Novelty .....	157
4	Describing relationships computationally .....	158
4.1	Introduction .....	158
4.1.1	Model performance and cross-validation .....	160
4.1.2	Linear regression.....	162
4.1.3	Chapter related literature.....	165
4.2	Results .....	168
4.2.1	Surface characterisation .....	168
4.2.2	Cell images and measurements.....	181
4.2.3	Computational cell models .....	212
4.2.4	Sensitivity analysis .....	249
4.3	Discussion.....	259
4.3.1	Cell performance .....	260
4.3.2	Computational cell models.....	274
4.4	Novelty .....	279
4.4.1	Cell culture experiments.....	279
4.4.2	Chemical inputs used by computational models.....	280
5	Screening surface chemistries computationally.....	281
5.1	Introduction .....	281
5.1.1	Aims & Objectives .....	283
5.2	Results .....	284
5.2.1	Model testing.....	284
5.2.2	Better synthetic environments.....	290
5.3	Discussion.....	295
5.3.1	Model testing.....	295
5.3.2	Reassessing discovered surface chemistries .....	298

5.4	Novelty .....	301
6	General discussion .....	302
6.1	Surface chemistry of <i>in vitro</i> cell culture environments.....	302
6.1.1	Lipophobicity.....	302
6.1.2	Lipophilicity .....	303
6.1.3	Molecular mass/volume .....	303
6.1.4	Surface charge (pKa) .....	304
6.2	Predictive models.....	305
6.2.1	Model testing.....	306
6.2.2	Predicting new surface chemistries.....	307
6.3	Summary .....	308
7	References .....	309
8	Appendices.....	351
8.1	Cell culture solutions.....	351
8.2	Immunocytochemistry solutions .....	352
8.3	Machine learning schemes .....	352
8.4	Neuron proportion model.....	356
8.5	Chemical value table .....	357
8.6	Contact angle table .....	358
8.7	Manuscripts in preparation.....	359

## ACKNOWLEDGEMENTS

---

My foremost gratitude goes to my supervisors Dr Paul Roach, Dr Rosemary Fricker, and Dr Theocharis Kyriacou for all their help and support during the challenging years of the PhD. Special thanks to Dr Rupert Right for the cell data for chapter 3 and for the practical help. Special thanks to Matthew Kose-Dunn for being a good friend and for the help throughout the PhD. My gratitude to Dr Munyaradzi Kamudzandu, Dr Síle Griffin, and Dr Folashade Kuforiji for the practical help. Thanks to the NEXUS team at Newcastle for the XPS data. I would like to thank the DTC Regenerative Medicine for recruiting me and the DTC 5 cohort for the fun times. Thank the Keele hardship fund team and Hamza Mashagba/Owida for the help during difficult financial times. For the social times I thank the Harvey labs, Guy Hilton Research centre especially Hati Kose-Dunn, Jessica Bratt, Kaarjel Kauslya, Zaid Younis, Dr Mike Rotherham, Dr Abigail Rutter. My family has been my foundation throughout my life and has been there for me when I needed them the most during the PhD. The people who have most of my gratitude are my parents and my personal mentor, Chris, seeing me through my studies.



# 1 INTRODUCTION

---

This project is about finding artificial environments to grow and mature stem cells more effectively and efficiently for use in therapeutic strategies for neurodegenerative diseases. The focus is on optimising the surface chemistry of artificial cell culture environments.

## 1.1 NEURODEGENERATIVE DISEASES

Neurodegenerative disease is the umbrella term for a range of conditions, which primarily affect neuron cells of the human brain. Neurons are considered the functional component of our nervous system, which includes the brain and spinal cord. Our body's repair system cannot replace dead or damaged neurons well. Examples of neurodegenerative diseases include Parkinson's disease, Alzheimer's disease, and Huntington's disease. Such diseases are incurable and debilitating conditions that result in progressive degeneration of neurons. This causes problems with movement (called ataxias) (Figure 1.1), or mental functioning (called dementias). In the UK alone 1 in every 100 people are affected by neurodegenerative diseases (1,2) and dementias are responsible for the greatest burden of disease with Alzheimer's representing approximately 60-70% of cases (3).

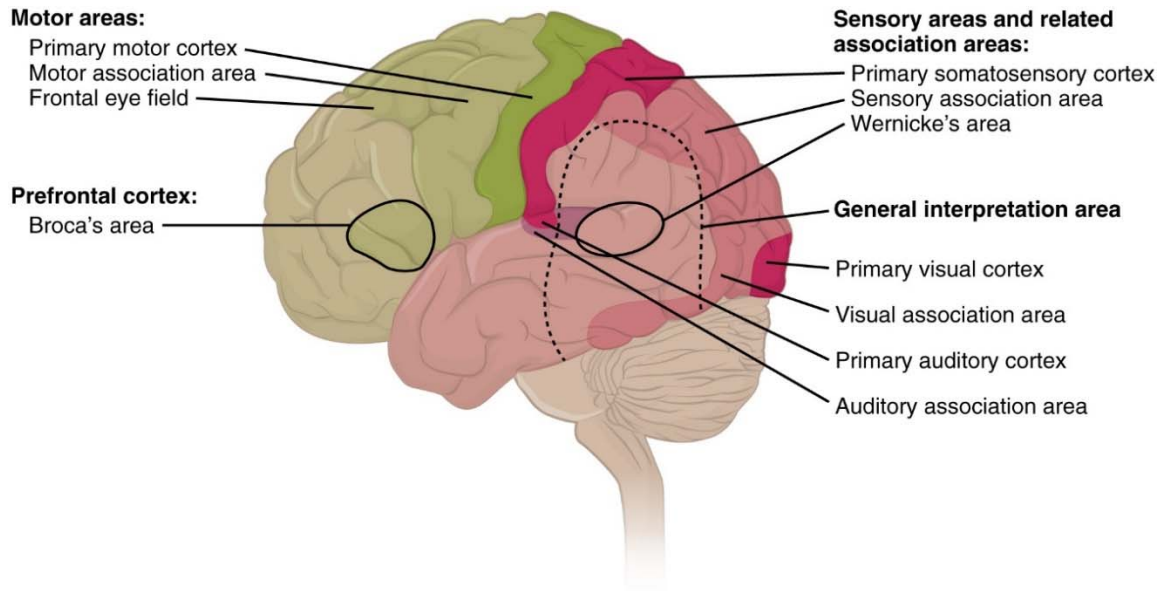


Figure 1.1: Drawing of the human central nervous system. Problems with human movement (ataxia) related with neurodegeneration arise to the motor areas of the brain. Taken with permission from (4).

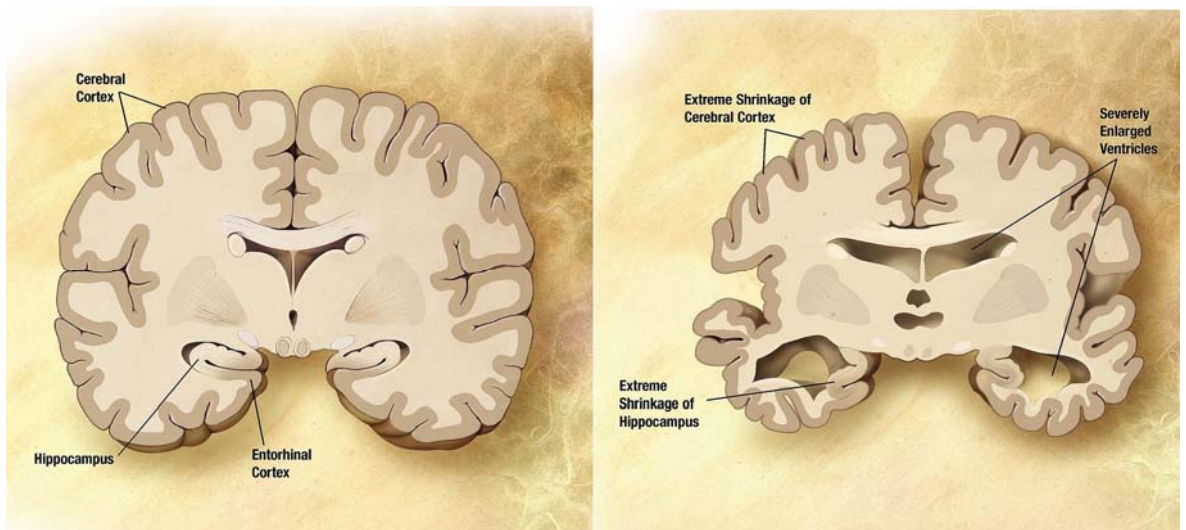


Figure 1.2: Drawing of human brain slice comparing healthy brain (left) with dementia brain with Alzheimer's disease (right). Taken with permission from (5).

Neurons interconnect different parts of the nervous system together with their axons. Injury to the nervous system can have clinical consequences ranging from impairment of musculatory or sensory function, to serious cognitive disruption and death. This is due to cell death and communication interruption along axonal pathways that control neurologic functions. A key issue faced by surgeons repairing the damaged nervous system is scar tissue formation restricting cell communication (6).

Once nerve cells degenerate remaining cells compensate to keep biological function intact as much as possible. At this stage, symptoms are usually mild and can be like other conditions. Once the remaining cells die, the symptoms become profound. With the exception of Huntington's disease, early diagnosis is not yet possible for most neurodegenerative diseases (7–11). Drugs to slow or stop the progression of some of these diseases are in clinical trials (1,12) however it will take years before they are fully developed. Current approaches for nerve repair is graft transplantation from undamaged sites of the same organism (autologous) (13). This method though sacrifices healthy functional tissue and does not result in complete repair (14). Tissue engineering approaches using cells from our endogenous repair system (stem cells) allows the development of cellularised tissue replacement therapies for damaged or degenerated brain tissue from injury or pathology (15).

## 1.2 THE NEED FOR TISSUE ENGINEERING

A key challenge for tissue engineering therapies is neural alignment and specific re-connectivity. The nervous system has very limited self-repair capacity and this explains why clinical outcomes are poor. Post-injury of the central nervous system (CNS) creates glial scars that compromise the ability to regenerate neural circuits that could potentially restore function (6). Preclinical studies for cell replacement therapies to treat neurodegenerative diseases are currently focused on primary neural stem cell-derived populations as the raw tissue material. These are usually induced pluripotent stem cells or are cells harvested from embryonic developing brain tissues due to the abundance of neurons needed to treat a disease (e.g. cholinergic neurons for Alzheimer's disease). Unfortunately, preclinical research has been trivially successful largely due to inefficient differentiation of stem cells to target adult neural populations.

Enhancing regeneration capacity necessitates addressing glial scarring and differentiation potential shortcomings by directing cellular processes. This necessitates the investigation of surface-cell interaction in the search for a synthetic surface material for the cell's microenvironment (16).

### 1.3 REGENERATIVE MEDICINE: BIOMATERIALS

Biomaterials are predominantly used for medical applications such as drug delivery, device-based therapies and cell therapies in tissue engineering. They have been rapidly adopted since the 80's (17) from merely interacting with biological systems to influencing biological processes. The European Biomaterials Society defined biomaterials as *“material intended to interface with biological systems to repair, replace or augment tissue or organ back to normal function”*.

Nowadays, biomaterials are glorified but are perhaps unexploited (18). Some tissues in the human body like skin and liver have excellent regeneration ability after damage. Other tissues such as cardiac muscle and the nervous system have poor regeneration ability. The new paradigm of regenerative medicine is to extend the quality of life span in ageing populations where currently chronic disease is common. This is to reduce some of the burden of the healthcare system. The notion here is to use cells to restore diseased or damaged tissue and/or cure chronic diseases that were previously managed (symptoms treatment). Early in the noughties, cell therapies have been tested for neurodegenerative diseases in pilot clinical trial of small scale. Material-based approaches for tissue engineering has not been used for neurodegenerative diseases. They have been used in

spinal-injury lesion animal models (19) with some functionality returning. The authors used poly(lactic-co-glycolic acid) scaffolds with neurons seeded as the implant solution.

Most translational projects for regenerative therapies have been focused on cells. For example, foetal neural grafts of cells directly dissected from foetal tissue of the central nervous system were used to treat Parkinson's disease (20). The grafts varied from aggravating symptoms to improving diseases progression and reducing the dependency on drug medications. The biggest challenge with cell therapies is cell source scarcity; there is simply not enough foetal neural tissue to meet patient demands. In the future, cell therapies are likely to be most successful treating neurological disorders by replacing discrete cell populations e.g. cholinergic neurons (work with acetylcholine) degenerate in Alzheimer's disease and dopaminergic neurons (work with dopamine) degenerate in Parkinson's disease.

### 1.3.1 Stem cell therapies

Stem cells are the raw materials used for tissue engineering. Stem cells are unspecialised and are tasked to make copies of themselves (self-renewal) and mature (differentiate) to cell types that make up our organs. Stem cells are attractive as a cell source in regenerative medicine for scalability reasons. It is a way to alleviate the need for large numbers of cells for use in e.g. transplants. Stem cells are situated throughout the human body in organs and they have existed since conception and development.

Stem cells are classified in three groups based on their differentiation ability. These classes include pluripotent cells that can differentiate to three primary groups of cells that form an organism. These groups are ectoderm giving rise to skins and nervous system; endoderm

forms gastrointestinal, respiratory tracts, endocrine glands, liver, and pancreas; and mesoderm forms bone, cartilage, most of the circulatory system, muscles, connective tissue, among other organs/tissues. Embryonic cells can come from:

- Unused embryos from donations (ES cells)
- Transferring the nucleus from any somatic cell to an egg cell (ntES cells)
- Unfertilised eggs by tricking them into developing into embryos using chemical treatments

Multipotent cells can differentiate to multiple types but are more restricted compared to pluripotent cells. Multipotent cells can differentiate to cell types within a given cell lineage or small number of lineages, such as white or red blood cell. Multipotent cells can differentiate to oligopotent cells where these are limited to becoming one of a few different cell types. Finally, unipotent cells are fully specialised and can reproduce to its own cell type.

There are two forms of stem cell therapies regarding donor and recipient. In autologous cell therapies, the donor is also the therapy recipient. In this way, the chances of immune rejection for the therapy are reduced. This kind of therapy is possible when there is an abundance of adult stem cells to work with, either where situated or extracted, manipulated and returned. Differentiating cells to desired cell type(s) provides the flexibility to produce a wider range of stem cell therapies.

Allogeneic cell therapies use cells from different donor(s). The benefit here is that stem cells can be derived from more diverse and multiple sources to target more diverse therapy requirements (e.g. mesenchymal stem cells for Crohn's disease (21). In pharmaceutical manufacturing, this approach is promising it form the basis of "off the shelf" products (22).

Allogeneic therapies need to consider address immune-rejection before embryonic stem cell-derived cells or tissues can be used as medicines. The types of stem cells are mentioned next.

#### 1.3.1.1 **Adult stem cells**

Adult (somatic) stem cells are undifferentiated cells found through tissues in the body after development used to replenish dying cells and regenerate damaged tissues. Adult stem cells are multipotent and these include neural, hematopoietic, and mesenchymal lineage (MSCs) among others. MSCs can be derived from adipose and stromal bone marrow tissues. These have been used in the clinic since the noughties because they modulate endogenous tissue and immune cells thus making them ideal for injury healing (23,24).

After birth, we humans possess a limited supply of neural stem cells and this provides the potential to treat neurodegenerative diseases. In lab cell culture (*in vitro*), it is difficult to derive dopaminergic neurons from adult neural stem cells in addition to the difficulty acquiring them in the first place. Because of this, research on treatments for neurodegenerative diseases faces slow progression. Clinically, adult stem cells are the safest to use due the restriction of possible cell fates. This is also one of their biggest drawback – it limits their potential.

#### 1.3.1.2 **Embryonic stem cells**

Unlike adult stem cells, embryonic stem cells (ESCs) have pluripotent differentiation potential. They can form almost any cell type of the developed body. The self-renewal ability of ESCs is excellent and this means large cell populations can be produced if required. This type of cells were first separated from mice back in the 80's (25). Human ESCs (hESCs)

were first isolated late 90's by (26). hESCs are derived from the inner cell mass of day 5-7 blastocysts from residual IVF tissue (27). From small populations of ESCs, one can produce very large volumes of cells needed to develop cell and tissue therapies for neurodegenerative diseases.

#### **1.3.1.3 Induced pluripotent stem cells**

Induced pluripotent stem cells (iPSCs) are derived from mature cells of tissues such as skin and through genetic re-programming they were brought back to the embryonic stem cell lineage (28). These cells can be derived from individuals and therapies involving these cells are autologous. These cells have similar traits as ESCs but since there is no requirement to use embryos there are less ethical issues such as destruction of human embryos, abnormal cell reprogramming due to the induction of human iPSCs, and tumorigenesis in the process of stem cell therapy (29).

#### **1.3.1.4 From stem cells to neurons**

There are several methods developed to differentiate pluripotent stem cells to neurons. Neurons degenerating in Parkinson's disease communicate using dopamine and these cells have been produced from embryonic stem cells (ESCs) using various techniques. Methods providing effective differentiation include the use of co-cultures providing environmental cues (30) and/or the addition of signalling molecules ( $\beta$ FGF,  $N_2$ (31), and Nurr1, LIF, FGF-8, 4 and 2, Shh (32)). These signalling molecules are derived from morphogens, tropic factors, cytokines, or mitogens in complex culture media for ESCs to recapitulate the natural environment for the neural lineage.



#### 1.3.1.5 Foetal neural stem cells

Foetal neural stem cells (FNSCs) are currently used in cell replacements strategies to replace damaged tissue from neurodegenerative diseases. The main reason is FNSC differentiation is easier to control and that gives a clear advantage when choosing a cell source.

Unlike ESCs, foetal neural stem cells have an advantage in transplantation into animal models as they are restricted in differentiating to neural cells and usually they do not divide significantly following transplantation (33). This means recipients of FNSCs are less exposed to the risk of tumours forming (teratomas) post-transplantation compared to ESCs. Teratomas tumours containing cell types from all three germ layers typically attributed to uncontrolled differentiation of rapidly dividing stem cells. Because of their cell type characteristics, these tumours are typically used as an indicator for pluripotency. Our group believes biomaterial approach potential has not been exploited and there is room for improvement to control FNSC responses. There are limitations with human FNSCs – currently, they do not present a practical route for large-scale therapeutic applications due to limited availability and quality of foetal tissue, as well as for ethical reasons (34).

#### 1.3.2 Cell differentiation complexity

For cell and tissue replacement therapies, stem cells need to mature (differentiate) to target cell populations. Directing high efficiency differentiation is a major challenge when creating complex tissues to replace damaged/diseased tissue. This is due to the complexity of stem cell differentiation. Work in this area is shedding some light however. Kirks *et al.* (35) followed what we knew at the time for developmental principles and derived *in vitro* functional dopaminergic neurons from ESCs. Through culture media, different signal

molecules were provided sequentially. The order decided was in line with the up-regulation of molecules during development of dopaminergic neurons in the embryonic midbrain. The idea for this work was to understand the process of stem cell development but the scope is too wide.

Focusing the scope, others investigated different concentrations of signal molecules directing naïve stem cells to different lineages (Figure 1.3: A). Both Wnt and sonic hedgehog proteins are good examples exhibiting the concentration gradient effect on neural tube development (36). Even at the single cell level the situation is still complicated. Dosing signalling molecules as a pulse rather than a steady dose can elicit different responses to cells (37,38). Heterogeneous cell responses can occur to individual cells and neighbouring cells where both were exposed to the same stimuli. This extends to parts of cell populations responding to stimuli and other parts do not (39). These examples are to show the complex dynamics of cell behaviour inferring to the lack of effectiveness in current differentiation protocols. Perhaps the most promising method to understand cellular systems is to combining meticulous experimentation with computation inference in which computational techniques are used to infer biological interaction and experimental techniques used to validate inferred interactions (40). Advances as such should be able to improve stem cell differentiation efficiency, which is a crucial step for clinical therapy translation.

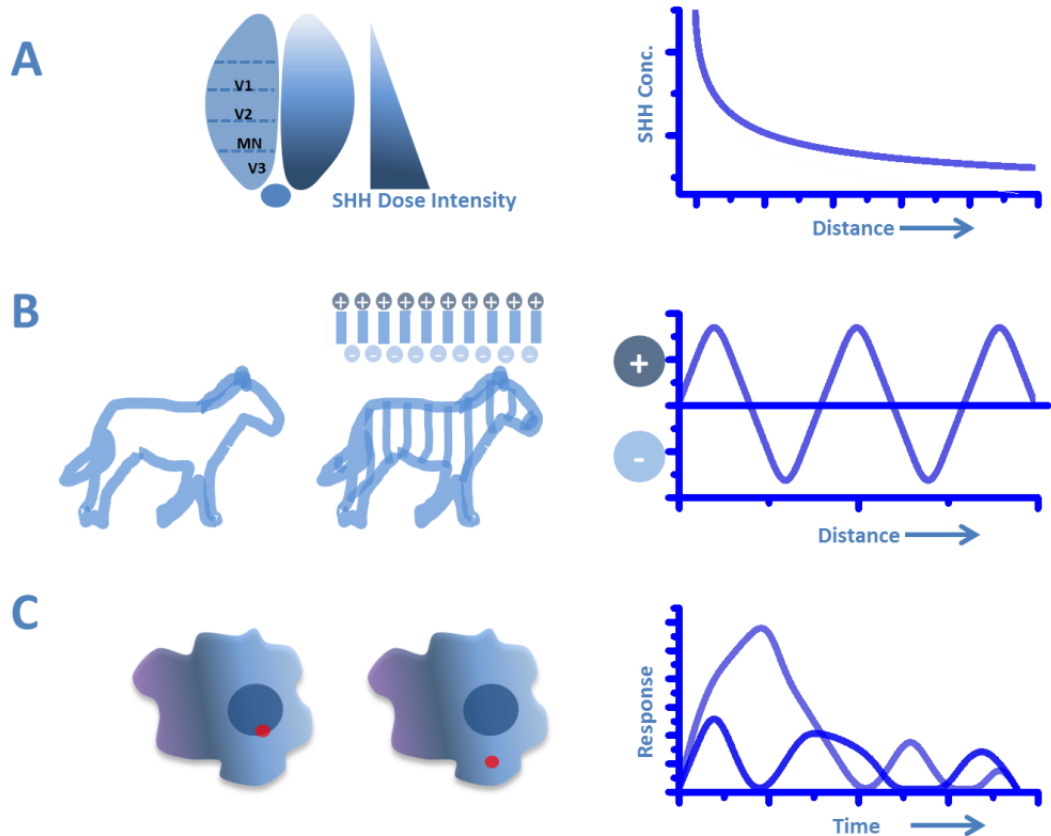


Figure 1.3: Cell morphogenesis related to signalling molecules. **A)** Concentration gradient of a morphogen (sonic hedgehog, SHH) forming brain tissue during development. **B)** Reaction and diffusion of morphogens related to zebra stripes. Natural patterns elaborately formed through spontaneous activator morphogens (pigment, dark colour) and inhibitor morphogens (pigmentless, white colour). **C)** Dose pulsing. Stimuli usually provides heterogeneous responses from cells therefore dosing frequency deserves more consideration. Graph: pulsing doses give temporal effect, which is important where adaptable responses to cells is required (38). Taken with permission from (41).

### 1.3.3 Optimal cell culture methodologies

#### 1.3.3.1 Spheroid culture methods

Spheroid methodologies have been used to produce *in vitro* made organs created from spheroids with their ability to self-organise and differentiate to produce tissue-like structures. The first structure created was cortical neural spheroids (42) from murine and human pluripotent stem cells. The authors used signalling molecules (FGF, Wnt variants) and morphogens (BMP) and low cell adhesive cell environments to achieve neural aggregation. They found spheroids to self-organise to distinguishable cortex structure with relevant markers and positions. The tissue was also found functionally active using calcium imaging – a technique that detects very fast oscillations of ionised calcium ( $\text{Ca}^{2+}$ ) waves

over distances. The same group produced an optic nerve head (optic cup) using the spheroid methodology and a similar approach as their previous work (43). A similar approach to (42) was used to generate cortical organoids from hESCs and iPSCs using ECM embedding and bioreactor culture to scale up the process (44). The authors discovered that iPSCs from small brain disorder (microcephaly) patients exhibited a characteristic of the disorder – premature differentiation in organoids.

### 1.3.3.2 Neurospheres

Expanding FNSCs as neurospheres is a simple solution to address scalability problems. These are multicellular 3D floating spheroids containing neural stem cells and progenitors (45). Murine neurospheres have been characterised and estimated to contain 1-3% oligodendrocytes, 17% neurons, 80% astrocytes and only 0.16% neural stem cells (46) (Figure 1.4). Others believe neurospheres contain different populations of stem cells (47). With FGF2 or EGF, different effects on cells were observed for each additive. In primitive stages of neurospheres in culture, only FGF responsive precursors exist. At low neurosphere density, FGF gave more proliferation and cell responses were different. Another important point with neurospheres is they can only be reformed and passaged a few times before neural stem cell and precursor populations diminish (47).

The location and migration of cells in neurospheres was also investigated. Neural stem cell markers (nestin and sox2) and the majority of dividing cells were found at the periphery of neurospheres (48). Mature cell markers for neurons and glia (Tuj1 and GFAP) are found at the centre of neurospheres. Relevant work where neurospheres were transfected with green fluorescence protein (GFP) using magnetic nanoparticle technology (49) found that cells were migrating throughout the spheroid.

Neurospheres in culture are known to merge (46) and this interferes with clonality as a condition in experimental procedures. Efforts have been made to culture neurospheres from single cells (50). The authors did achieve in making neurospheres from single cells but with a low yield. Neurospheres are believed to be heterogeneous due to cells functioning at biological clocks and pace in their cell cycle (51).

The size of the inherently heterogeneous neurospheres can be controlled to the single (rat) neurosphere in PMMA mirowells with PEG surface (42,52). Linear relationships were discovered between neurosphere and micro-well diameter. Large and small micro-wells (800  $\mu\text{m}$  and 200  $\mu\text{m}$ ) gave accordingly sized neurospheres (225  $\mu\text{m}$  and 50  $\mu\text{m}$ ).

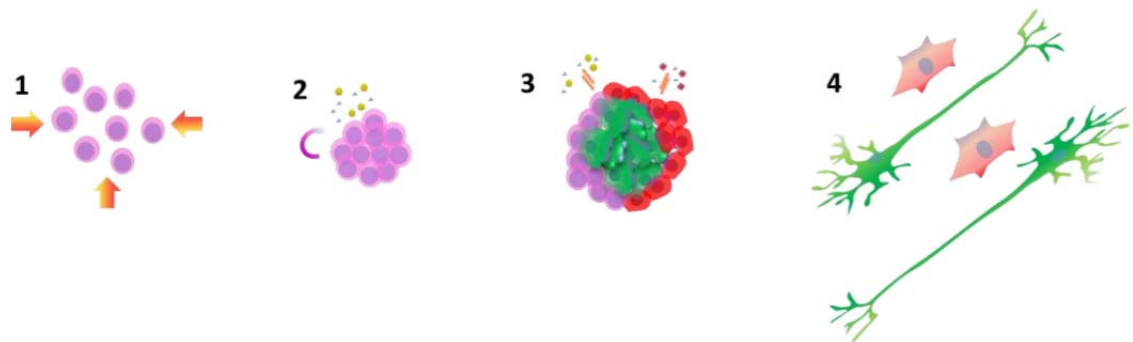


Figure 1.4 Drawing of neural stem cell culture. 1) Aggregation of neural stem cells (NSCs). 2) Neurospheres sustained with NSCs with fibroblast growth factor ( $\beta\text{FGF}$ ) in media. 3) New cell types are formed from proliferation and these tend to group by cell type. 4) Adhesion and differentiation of NSCs and progenitors on a sticky laminin coated surface with  $\beta\text{FGF}$ -free media. Taken with permission from (41).

### 1.3.3.3 Cell niche

Stem cells have their ability to self-organise into proliferative and differentiating niches combined of different cell types. Stem cell niche provides cells with developmental cues such as survival, maintenance, proliferation and activation (53). These are rich in extra cellular matrix, paracrine-signalling regimes among other specific cell signals (54). These signals have a powerful effect on cells as shown by (55) where neural stem cells were reverted to a less differentiated state of cells from all three germ layers (mesoderm,

ectoderm, and mesoderm). Together, the three germ layers will give rise to every organ in the body, from skin and hair to the digestive tract.

Synthetic biomaterials promoting cell self-organisation into proliferative and differentiating niches have not been found. Such biomaterials are envisioned to achieve niche-like cell microenvironments without the use of expensive reagents such as recombinant or highly purified proteins/macromolecules. In fact, it could also be possible to promote the isolation of different differentiating and proliferating niches on the same material. This means next generation high efficiency biotechnology production for cell therapies closer to reach. The key lies in harnessing the stem cell's natural abilities.

#### 1.3.4 Biomaterials for *in vitro* cell culture

Traditional biomaterials do not have the ability to adapt to living tissues during changing pH and body temperature caused by disease. Therefore, biomaterial scientists have been endeavouring to create smart biomaterials (56) that mimic living tissues in the last two decades (57).

Nerve tissue can be engineered and enhanced by modifying the biomaterial properties such as topography, stiffness (compliance), and arguably the most important, the chemistry to improve cell and tissue adhesion (58). A biomaterial provides mechanical support, shape, and hierarchy architecture with surface chemistry for cell attachment, cell-cell communication, as well as proliferation and differentiation for tissue regeneration. To date, most synthetic biomaterials derived for tissue engineering are synthesised either from lactic acid, chitosan, alginate, starch, collagen, hyaluronic acid, cellulose, fibrin, silk, and their derivatives (57).

Stem cells are the raw materials used for tissue engineering. In their natural environment (niche), stem cells can divide (proliferate) while keeping key properties intact. The niche microenvironment controls stem cell populations from growing uncontrollably (54) with two identified switches of senescence p16, p21 proteins signalling cells to stop dividing. In addition, the niche provides feedback and cell signalling which influences the activation, maintenance and differentiation of stem cells. The natural microenvironment and the control it has over neural stem cells is almost impossible to simulate in artificial environments (*in vitro*) (59,60). The closest to such environments are semi-biological which are made of a mixture of artificial bulk materials (e.g. glass) coated with biological materials (biomaterial) such as laminin protein (61). With biological materials however raise pathogenic concerns since they are biologically derived.

Synthetic biomaterials need to display remodelling properties over time to integrate with the ECM formed by the encapsulated cells during tissue maturation, and bridge with the natural ECM of the potential patients (62,63). This can be achieved through the modification of the chemistry with active units that are cleaved under specific biological stimuli (64,65). Mimicking the chemical or physical cues of the extra cellular matrix surrounding cells may not be always necessary for a successful tissue regeneration and integration. In the case of soft connective tissues like intestinal and abdominal walls, rapid prototyped three-dimensional meshes with macrostructural features can restore tissue functionality (66). Biologically modified or “plain” biomaterials have become smarter and more instructive templates for cells, the number of biomaterials that truly promote integration with the host environment is still limited (58). To improve tissue adhesion properties, the chemical design of biomaterials is the most important component. One of

the main strategies for engineering bonding materials is their functionalisation with groups that can connect to the natural tissue.

Self-assembled biomaterials possess both the physical dimensions of micron- and nanoscale ECM fibres and adhesion properties typical of hydrogels. They are generally comprised of functionalised amphiphilic polymers and have been successfully employed for bone, cartilage, and soft tissue regeneration with promising results (67–70).

In 1989, a research group led by Whitesides studied the interfacial properties of organic materials that control chemical properties such as wettability and acidity among others (71). At the time, the relationship between the microscopic structures of the biomaterial and the macroscopic physical properties were poorly understood. They've used long-chain thiols adsorbed on gold surfaces and varied the terminal group. The take home message was that wettability (hydrophobicity/hydrophilicity) is a macroscopic interfacial property was very interesting. In 1992, another revolutionary paper was published titled "How to Make Water Run Uphill" by Chaudhury and Whitesides (72). The group demonstrated how the very same chemical property was responsible for controlling water droplets going against gravity. The authors did this using surfaces that had spatial hydrophobicity gradient (vapor deposited decyltrichlorosilane) over 1 cm and by dropping water droplets on the most hydrophobic part. The water droplets moved towards the hydrophilic part of the gradient due to surface tension acting on the liquid-solid contact line. This inspired biomaterial scientists for the potential to control biological responses using cost-effective methods and materials.

More recently, research groups have been studying the effect of wettability on cell responses (attachment, differentiation, and controlled transfection) using high-throughput



approaches to rapidly screen materials. Alexander's group (2010) (73,74), use combinatorial approaches to synthesise materials as polymers on microarrays to identify optimal compositions for particular biomedical applications. In addition to wettability, the group uses time-of-flight secondary ion mass spectrometry (ToF-SIMS) and X-ray photoelectron spectroscopy (XPS). Others (75,76) use both topographical (parallel grooves and roughness) and chemical (plasma polymerised allylamine (ppAAm) gradients and polystyrene) to study dermal fibroblast and osteosarcoma cell adhesion, morphology, orientation, and spreading. The take home message however for both studies is prefer the 45°-65° water contact angle (WCA) bracket of the surface material with other studies finding WCAs around this bracket helps controlling cell attachment and adherence (77-79). This suggest it is best practise to avoid extreme hydrophobicity/hydrophilicity in the biomaterial design with respect to wettability as this chemical property is a descriptor of many chemical properties. The problem is even with the mid-range WCA bracket, there is a plethora of surface chemistries to test.

Perhaps the paradigm of one surface property such as wettability to explain cell responses needs to shift as it's too generic as a chemical descriptor. In a study in 1996 for controlling neuronal cell attachment, the author (80) found cell attachment was more sensitive to charged functionalities (imidazole and carboxylic groups) rather than hydrophilic/hydrophobic balance of the photoresist surface. In other studies, long chain hydrophobic self-assembled monolayers (SAMs) terminating with amines (-NH<sub>2</sub>) (formed by thiols to gold) sustained the attachment and growth of dorsal ganglia and PC-12h cells however no adhesion was observed on alkene (neutral and hydrophobic) or carboxylic acid surfaces (low charge and pKa value). This suggests a preference for amines (81). In another study, surfaces with a linear increase of amine content with mono, di- or tri-amine terminations were studied. Neuron attachment indicated a preference to di/tri-amines,

whilst, on monoamine surfaces, perinatal rat cerebella and embryonic mouse spinal cells did not attach at all (82). Adding to surface charge evidence – others have shown cell attachment on amino surfaces (83) with some suggesting this is due to the positive charge held by the surface groups providing electrostatic attraction to negatively charged membranes inducing cell adherence (84).

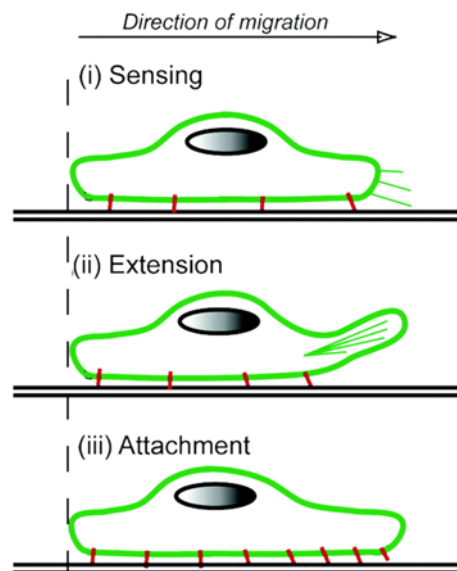
Surfaces with hydrophobic character such as phenyl groups hinder hippocampal neuron attachment (83) and carboxylic acid terminated surfaces can completely prevent neuronal attachment on amino-thiol and lysine coated surfaces (85). When a 1:1 mixture of amino and carboxylic acid thiols was used as presenting surface chemistry, a dramatic decrease in attachment was observed compared to amine terminated surfaces. The authors discussed their findings in terms of electrostatic interactions suggesting a strong relationship between amines and the cell membrane at physiological pH, whereas the mixed chemistry presents a cationic surface hindering interaction.

The logP is the partition coefficient between water and octanol, as a reliable indicator of the hydrophobicity or lipophilicity of (drug) molecules. In an article by Rawsterne *et al* in 2007, it was found that cell spreading correlates better with calculated logP of amino acid-modified surfaces compared to water contact angle (77). Cui *et al.* studied hippocampal cells harvested from 15 day (E15) Sprague-Dawley embryo rats for neuron attachment and axon extension (86). They used surfaces with laminin patterned grids on both positively charged amino groups (PEI) and negatively charged (hydroxyl) underlying surfaces. After 7 days, cell soma attached at cross-points of laminin grid pattern. The interesting part is that on hydroxyl underlying surfaces, neurites only followed the laminin pattern and extended along the grid lines whereas on underlying PEI surfaces, random neurite outgrowth was observed.

The length of the SAM chain was found to influence cell adherence. In a study investigating covalently immobilised adhesive proteins (laminin, collagen, and fibronectin) on surfaces with amino-thiols of varying lengths resulted in tighter connections for subsequently adhering cells (87). In another study, neuronal cell line PC-12 cells anchored more strongly to laminin on more disordered shorter thiols showed by the magnitude and the reproducibility of electrical impedance responses derived from receptor mediated linkages (88).

### 1.3.5 Surface-cell interaction

In laboratory (*in vitro*) cell culture the cells would initially sense the surface for binding sites using protrusions (89). Once found, the cells attach using receptors called integrins and thereafter proceed with survival and development functions such as growth, proliferation and differentiation (90,91). Such cell responses/behaviour depends heavily on their environment and its features such as chemistry and topography (Figure 1.5). Surface topography at the micron level plays an important part in determining cell adhesion and surface-bound characteristics (92) however the focus of this project is on the chemical component of the biomaterial.



Biomaterial success depends largely on the biological/surface interface with a few key molecular properties investigated and assumed “in action”. In previous studies, cell attachment was effected by surface wettability (76,94,95). In one example study, surfaces that repel water (less wettable) were observed to enhance the attachment of osteosarcoma cells (MG63) but had a negative effect on cell spreading. Optimum cell attachment was observed at mid-range contact angles ( $64^\circ$ ) (76). Others have studied another chemical parameter, the measure of solubility (partition coefficient) and its effects on cells (77,94). Engler *et al.* investigated the differentiation of a type of stem cell able to give rise to mesenchymal tissue such as muscle, bone, tendon and ligament among others.

One of the studies investigated partition coefficient estimates of (un-tethered) amino acid functionalised surfaces and cell spreading (77). Amino acids are the building blocks of proteins. It is believed that surface chemical and topographical characteristics influence the protein layer composition that is between the surface and cells (Figure 1.6) (90). The presenting chemistry (functional groups) at the surface has indirect control of cell responses (92,96). Surface chemistry, topography and stiffness (compliance) among other properties branching from them dictate the type, amount, and conformation of proteins adsorbed (92,96) (Figure 1.6). The focus of this project is primarily on cell responses.

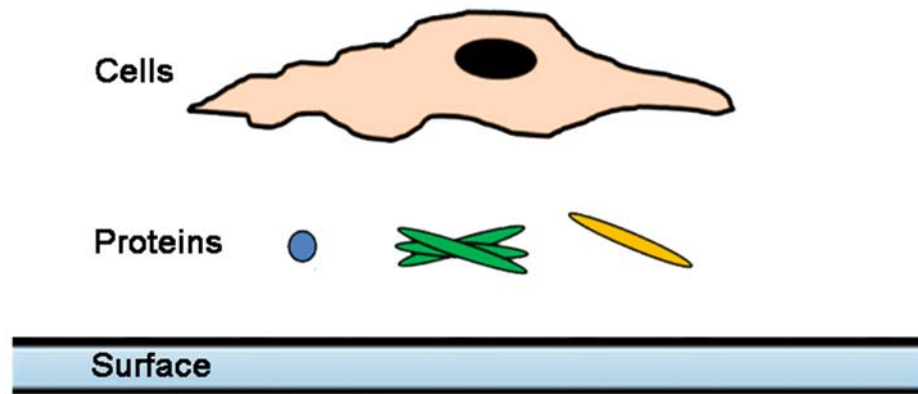


Figure 1.6: The components of the cell culture system in artificial environments (in vitro). Adapted with permission from Dr Paul Roach.

Cell environments with laminin-coated surfaces are a good choice as they mimic the laminin-rich niche and are usually chosen for neural culture protocols (61,97,98). Laminin is an extracellular matrix (ECM) protein and provides structural and biochemical support to the surrounding cells. Laminin promotes axonal outgrowth and cell adhesion through its positive charge followed by integrin binding for enhanced cell attachment (99). Laminin is usually part of complex differentiation protocols with cocktails of factors to direct stem cells or progenitors to a desired lineage. These factors serve as biological signalling with an effort to simulate closer the biological niche. Laminin with its ligands (binding sites), other extra cellular coated surfaces, and matrigel owe their success to adhesion molecules and integrins (such as  $\alpha_5\beta_1$ ,  $\alpha_8\beta_1$ ,  $\alpha_v\beta_3$  (100)) residing on the cell membrane. Long term cultures of neurons (hippocampal) has been achieved for 24 passages on laminin-coated surfaces (101).

Surface engineering techniques can be used to develop cell therapies for neurodegenerative diseases. These therapies can be realised by defining the inputs of manufacturing processes. Perturbations to the cell culture system represent an immense challenge to overcome as they can cause disturbances from any direction so systems requiring fewer interventions are preferred. The idea here is to harness the stem cell's

inherent ability to proliferate, organise and differentiate. Relevant work has aimed to simplify the culture surfaces by using only laminin (102) or synthetic substrates (103). However, their culture media and technique are far from simple and controlling cell behaviour was not the scope of the study. Tailored cell-population specific culture surfaces would have advantages such as reduced cost and incorporation into pre-existing workflows compared to optimising culture conditions. Surface materials are easier to adopt and new procedures involving these can evolve in the therapy development pipeline. Materials should be designed with core cell behaviours to begin with, as it will be difficult to activate biological pathways requiring many steps. The culture surface will be impeded with the protein layer among other factors that affect cell behaviour.

Crude high throughput studies (94) cannot grasp the mechanisms for rational material design. However, a novel drug design method where ligand designs are adapted to multiple-targets beforehand experimentation (a priori) has emerged in 2012 (104). This method is Pareto-based where ligand designs (resources) are allocated in the most efficient manner to target multiple profiles (outputs) and they consider ligand (chemical) design trade-offs. There are other examples of this method (105). Biomaterial research needs development to find where these design trade-offs occur. Stem cell research has been slow to adopt these new biomaterial approaches as often they fail to compare fairly with ascertained individual effects and perhaps due to improper computational model validation (106). For the former issue, in a study where different substances that govern the pattern of tissue development (morphogens) and growth factors were compared for maintained and differentiation of neurospheres. The study had clear inputs and outputs but an assumption was made that different proteins behave the same way in cell culture, which is not true (92,107,108).

### 1.3.6 Surface-protein interaction

Proteins are secreted by cells themselves. Proteomics is the study of protein function and allows the investigation of the sets of proteins expressed in cells to understand cell proliferation and differentiation to specific lineages (109). Secreted proteins comprise vital molecules which are encoded with around 10% of the human genome (110). The secreted molecules mediate intercellular interactions and are involved in maintaining homeostasis at the organ level (111).

The choice of reagents is important to cells in culture. Cells require numerous factors to maintain their development and growth *in vitro*. In their natural environment (*in vivo*), factors are available in biological fluids surrounding cells whereas *in vitro* they are usually added in the form of serum such as foetal calf serum. There are 30 to 40k signalling molecules in these serums, many of which are poorly understood as to their interaction with their surrounding environment. It is still standard practise to use these serums although there is a drive to move away from them for therapy translation (112,113). This is because these materials tend to be animal derived materials with a risk of carrying pathogens.

Some cell types are sustained better in pre-condition media by another cell type population previously cultured. During culture, cells secrete factors to communicate and mediate their surroundings and these pre-conditioned media provide better cell culture conditions for some cell types compared to unconditioned media. In a study, human embryonic stem cell (hESCs) self-renewal and differentiation potential was assessed on protein conditioned and unconditioned biological substrate (matrigel). Mass spectrometry revealed 80 extracellular proteins in matrix conditioned by hESC (114).

The key lever on cell response is the relationship of the biomaterial with adsorbed proteins described in detail by (115) and (92). Proteins reach the surface before cells due to their smaller size. Once water molecules are moved away the interaction between the surface and protein starts and proteins cover the surface (surface conditioning) (92). Figure 1.7 shows the sequence and interactions of the components of the cell culture system.

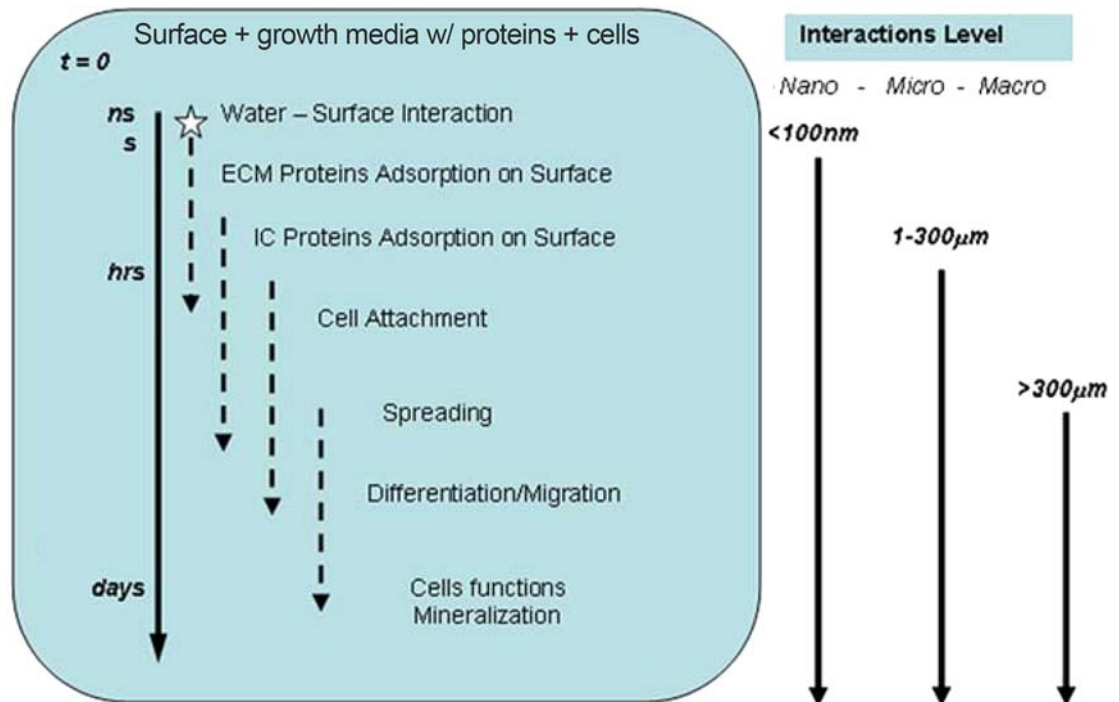


Figure 1.7: Interactions between surface-proteins-cells in time and level. Adapted with permission from (92). RightsLink license 4050271025585, Springer.

In the adsorption process, proteins undergo structural changes until they become energetically favourable. The protein adsorption process is different for each biomaterial design. Previously unavailable protein domains can be exposed (116) presenting peptide binding sequences for anchoring molecules on the cell membrane such as integrins (117). An example of an adhesion molecule for neural cell types is NCAM (118).



### 1.3.7 Presenting chemistry of biomaterials

Biomaterial surfaces can be modified with a variety of methods such as chemical gradients (41), self-assembled films, surface active bulk additives, chemical reactions and molecular grafting (90). The most widely used method to modify the surface chemistry is with self-assembled monolayers (SAMs) which is a process of coating the surface with molecules that form highly ordered structures on specific substrates. SAMs form chemically and physically stable covalently attached monolayers on surfaces such as gold and glass (Figure 1.8). In the literature, SAMs are used to study the effect of well-defined chemistry on biological processes such as protein adsorption and cell response. SAMs model biomaterial surfaces and relevant work employ thiols on gold surfaces (119–122) and alkylsilanes on silicon (71,123–127).

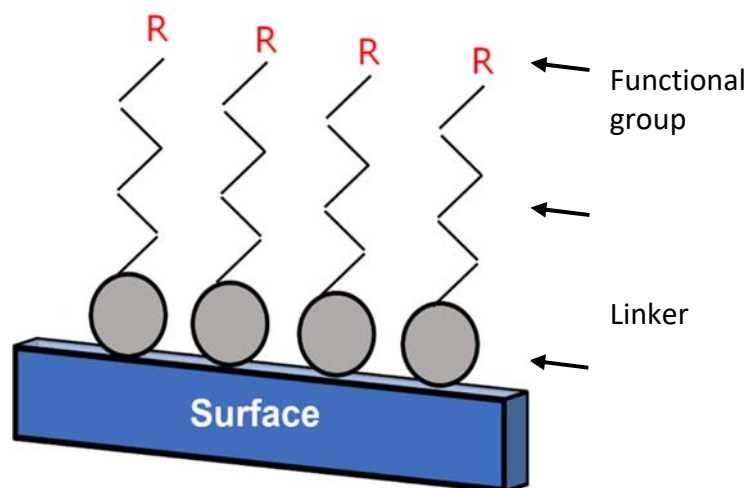


Figure 1.8: Self-assembled monolayer on a solid surface.

SAMs provide an efficient and effective method to change the presenting chemistry of the biomaterial surface. Changing the atom in the terminal group alters the properties of the surface and therefore the protein adsorption and cell interaction. Studies agree the initial cell adhesion to SAMs is greatly affected by the surface functional groups and displacement of adsorbed serum proteins with cell adhesive proteins playing an important role in cell adhesion (85,107,128). SAMs can be functionalised on surfaces providing the opportunity

to tailor the chemical properties of surfaces such as wettability/lipophilicity and acidity. These properties can be tuned with the SAM linker and the terminal group (129). SAMs are comprised of three parts, the head group, the alkyl chain and the terminal group (Figure 1.8). The head group anchors the rest of the molecule on the substrate (e.g. triethoxysilane). The alkyl chain provides stability of the monolayer due to hydrophobic and Van der Waals interactions that influence the SAMs ordering as well. Lastly, the terminal group introduces chemical functionality in the monolayer system (130,131).

SAMs allow the investigation of fundamental physical properties of interfacial chemistry, solvent molecule interaction and self-organisation and these, most likely, made them popular (132,133). A range of functional groups such as alkyl, thiols, carboxylic acids, phenols have been studied with each providing a better understanding over molecular surface interaction and the role of different molecules in the surface chemistry. Protein adsorption has been investigated with SAMs examining surface chemical characteristics importance on adsorption kinetics, adsorbed concentration and biological activity of the protein layer (71,134).

Understanding the effects of surface chemistry with adsorbed proteins and cells in culture will lead towards better neural stem cell control *in vitro* (90). It is also believed the investigation of the interactions between the components of the NSC culture system may hide clues as to how connections or mis-connections may arise in the central nervous system (CNS) (90,135). It is hoped to direct cells *in vitro* to develop into functional nervous tissue that can be used in therapeutic strategies.

## 1.4 COMPUTATIONAL APPROACHES

Sciences in molecular biology, regenerative medicine and neurogenesis have unravelled a plethora of biological facts such as genome sequences, stem cell therapies, and dividing neurons in the adult brain (due to neural stem cells) (136,137). A crucial aim of biology is to understand biological components interaction in a dynamic, parallel or concurrent fashion. The spectrum of biological components ranges from molecules, cells, tissues and organs to complete organisms. In complex biological systems, interactions of smaller components such as molecules and cells induce new, emergent properties that are observable at higher scale, on tissues and organs.

The components of biological systems undergoing specific interactions have been defined by evolution. These are fundamental processes behind physiological and pathological conditions: ranging from tissue formation, response to stem cells therapy and cancers. An understanding to the system-level should be the goal. Insights into the function of biological systems however cannot result purely from intuitive paradigms due to the intrinsic complexity of such systems and experimental limitations. A combination of experimental and computational approaches can tackle this problem (138–145).

### 1.4.1 Data science and overlapping sub-fields

Statistics is a branch of mathematics dealing with data. This includes collection, organisation, analysis, interpretation, and presentation of data. Statisticians anticipate what can go wrong with experiments and fallacies can be drawn from naïve data uses. There are techniques to solve an abundance of problems, but the approaches have an inherent conservatism attached. It finds what could go wrong through testing of hypotheses. Informatics and bioinformatics, after the biology is separated, deal with data

infrastructure and matching algorithms. The aim here is to create and manage data stores, databases and design efficient matching and query algorithms. Big data is a hot topic these days. It is essentially the infrastructure and the platform to perform modelling or generating reports. Simulation is an exploratory approach allowing the generation of possible outcomes for a given problem. This is useful when certain assumptions are made about the data but are not represented in the data itself. Simulations allow the generation of variance in the outcomes and the testing of model stability.

Data mining is the discovery of implicit, previously unknown and potentially useful information from data. The idea is to automate this process by creating computer programs that sift data and seek patterns or regularities. Machine learning (ML), provides the technical basis of data mining and is an optimistic field. It overlaps largely with scientific methods, math and statistics (Figure 1.9). The notion is to create predictive models aiming to be indistinguishable from “correct” models. Perhaps prediction here should be rephrased to optimisation problem (146). Predictive analytics is a set of goals and techniques with emphasis in constructing models and overlaps with data science. Data science represents the ownership and management of the entire modelling process. This includes discovering the need, collecting and managing data, building and deploying models into production.

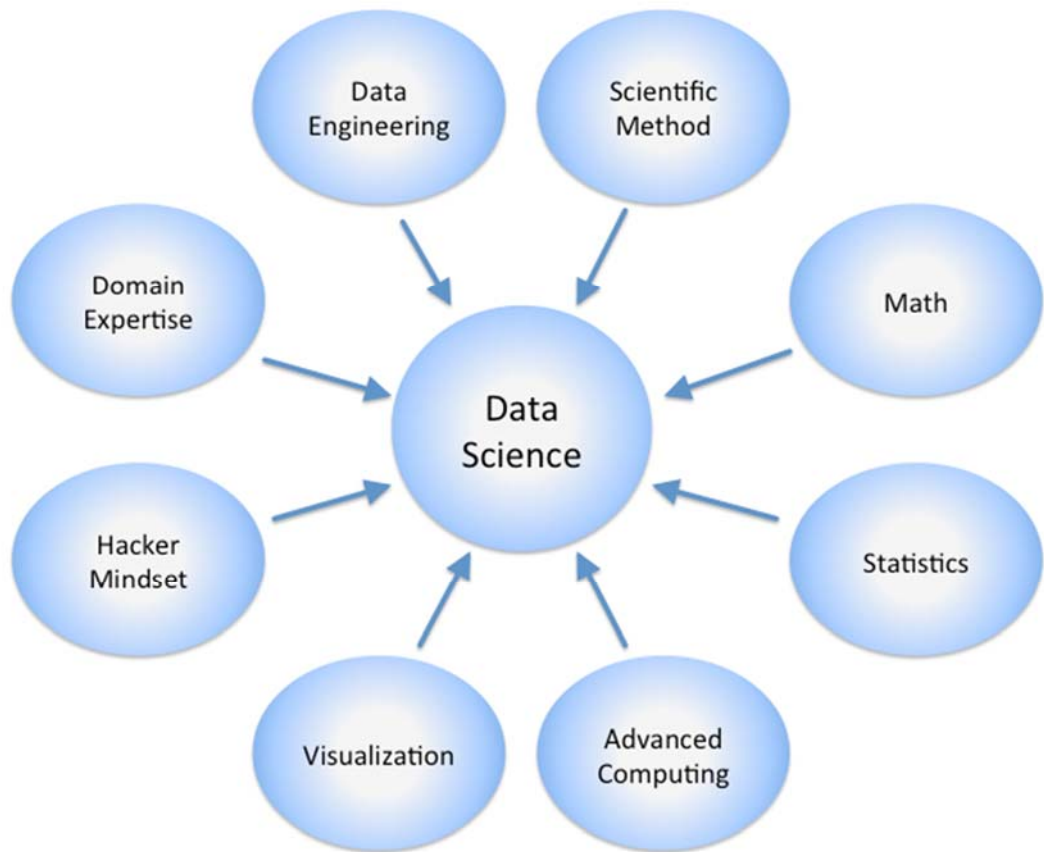


Figure 1.9: Data science and related sub-fields. Taken with permission from (147).

### 1.4.2 Relevant literature

Data mining has been used to analyse in multiscale registered trials for stem cell-based regenerative medicine (148). The authors used chord diagram and phylogenetic-like tree visualisations to assist in knowledge discovery of clinical trials registered at ClinicalTrials.gov. They screened 5,788 trials, 939 were included and 51% of these were related to mesenchymal stem cells (MSCs). More than half the MSC studies concerned allogeneic MSCs and received more support from industry than autologous MSC studies. The authors found the use of cultured cells have increased greatly since 2009. In trials, the use of cells derived from adipose tissue has also increased compared to bone marrow cells. The use of adipose-derived stromal cells was predominantly autologous, restricted to European countries and supported by industry compared with other MSCs.

Another data mining example is the work of a research group that showed the continuous depletion of neural stem cell pool, due to their division, might be responsible for age-related decreased neurogenesis in the adult hippocampus (149). The authors used various computational approaches to determine the age-related changes in the pools of stem and progenitor cells, Hayes & Nowakowski (150) approach to model the single- and double-label pulse dose experiments (related to cell staining for microscopy), and to determine the parameters of the cell cycle of stem and progenitor cell populations. They found upon exiting their quiescent state, adult hippocampal stem cells rapidly undergo a series of asymmetric divisions to produce dividing progeny destined to become neurons and mature astrocytes.

Adult neurogenesis has benefitted from computational neuroscience. This field is about modelling new neuron function and it is unravelling the sophisticated biological processes of adult neurogenesis *in vivo* using data mining, machine learning and simulations. *In vivo* relevant studies fall outside the scope of this project. For those interested, noteworthy and recent reviews are in (142,151,152). Protein adsorption on biomaterial has benefitted from computational methods in the recent years. Noteworthy literature can be found here (92,115,153–156).

Cellular automata (CA) is a 'top-down' discrete modelling approach used to simulate cell morphogenesis and tissue development (157). CA is a discrete modelling approach that captures system-level mechanisms of complex biological phenomena by defining a series of decision rules implemented in a parallel and dynamic manner (158). In a typical tissue growth model, a cell moves in one of  $n$  directions with a certain probability in each simulation cycle. A number of simulation cycles are performed iteratively. The advantages of such models include the relative simplicity of visualisation, implementation, and its

design extensibility and flexibility. Cellular Potts model is an extension of CA and is a simulation approach that incorporates mathematical descriptions of cell motility and connectivity (159,160). Cellular Potts models can specifically define cell structure in terms of shape and volume parameters (161–163), unlike traditional CA. In tissue growth simulation, these models foster cell movements in those directions that minimise a local energy function (163). The idea is that tissue geometry, area and localisation are regulated by favouring stronger bonds, i.e. their contact energies as well as larger cell boundaries (164). Thus, the Cellular Potts model is a powerful approach to incorporating quantitative cellular information into discrete or cell-based models.

In molecular biology, a group developed a computational tool to analyse and sequence the genome-wide data from the mechanism controlling gene expression (DNA methylation, (MeDIP-seq data) in human embryonic stem cells (hESCs) along the endodermal lineage (165). The group coined their tool MEDIPS and it processes the inherently complex MeDIP-seq data faster, more accurate, with increased sensitivity and with better correlation with sequenced results compared to existing methods. MEDIPS belongs closer to data science as a method since it performs a multitude of functions with the MeDIP-seq data. Its strength is that it significantly reduces the imbalance of sequenced data generation and analysis. The authors were able to investigate the effect of other mechanisms controlling gene expression in cells (differential methylation). In addition, MEDIPS allowed the analysis of the interplay between silencing genes (DNA methylation), histone modifications, and transcription factor binding and show that in contrast to ‘from scratch’ methylation, demethylation (activating genes) is mainly associated with regions of low CpG densities in regions of the DNA.

In relevant work, data mining was used to elucidate regulated gene expression to provide insight in disease and development. Yeo *et al.* (166) have decoded functional RNA elements (involved in protein synthesis) *in vivo* by studying the FOX2 binding protein to identify its targets in hESCs. FOX2 is a key regulator of gene editing (exon splicing) for cell survival, differentiation and development in the nervous system and other cell types. The mapped FOX2 targets revealed other splicing regulators and allowed the creation of a model that shed light into binding or skipping splicing events in a position-dependent manner. The authors did not provide model specifics but they did mention the model was created from consensus binding motifs for FOX2 depletion-induced gene editing. FOX2 was discovered to be a critical regulator of a splicing network and important to the survival of hESCs.

Wilson *et al.* (167) found hematopoietic stem cells (HSCs) in mice reversibly switch between dormancy and self-renewal contrary to popular belief they are turn over every few weeks. The authors used identified the cells using flow cytometry and label-retaining assays (BrdU and histone H2B-GFP). They then fit ordinary differential equation (ODE) model(s) on their experimental data on one (dormancy) and two-population (dormancy and self-renewal) versions. The two-population ODE model had a much better fit on their data and the data of a previous, less extensive study (168) in support of the two-population HSC hypothesis. The authors used stochastic markov simulations to find when HSCs divide in support of the dormancy and self-renewal hypothesis. The results revealed that HSCs divide every 145 days, or 5 times per lifetime agreeing with experimental findings.

Data mining is found in transcriptional studies as well. A group investigated the proteins involved in regulating genes (transcription factors, TF) and their specific interactions with targets necessary for programming the synthesis of gene-products such as proteins in embryonic stem cells (ESCs) (169). The authors acquired transcription factor and



transcription regulator data from the interaction of proteins and DNA within cells and DNA sequencing. These factors are known to play different roles in ESC biology as signalling pathways (LIF and BMP), self-renewal regulators, and key reprogramming factors. Computationally, the authors developed their own approach to find associations between TF occupancies and gene expression based on TF-binding data. They then performed k-means clustering (grouping data based on proximity) and defined five classes of genes that are associated with a similar set of transcription factors. Based on these associations between binding and expression, they have constructed a transcriptional regulatory network model that integrates the two key signalling pathways (LIF and BMP) with the intrinsic factors in ESCs. Collectively, the comprehensive computational and experimental mapping of TF-binding sites identified important features of the transcriptional regulatory networks that define ESC identity (Figure 1.10).

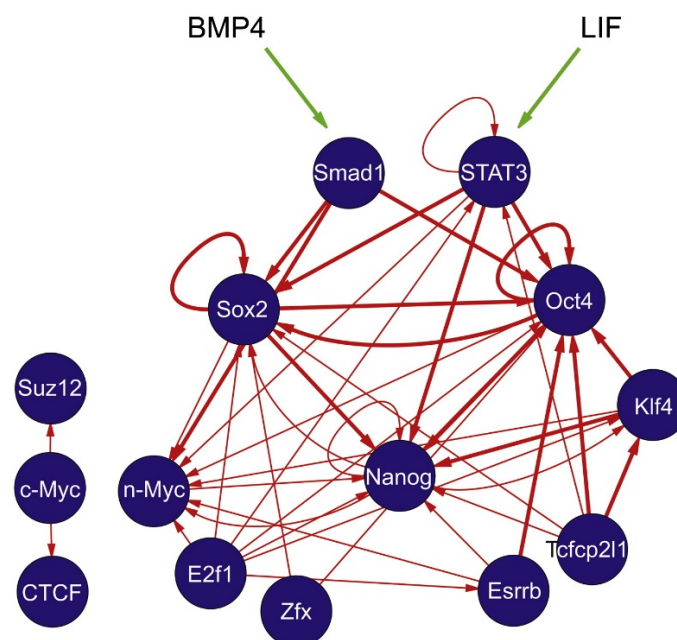


Figure 1.10: Transcriptional regulatory network inferred from real data during embryonic stem cell differentiation. Transcription factors are in blue bubbles (nodes). BMP4 and LIF are signalling pathways in cells. Thick arrow represent interactions inferred from binding data and both expression experiments whereas thin arrow represent interactions inferred from binding data and one expression experiment. RightsLink license 4070220942845, Elsevier.

Computational simulation work in bone tissue engineering with biodegradable scaffolds, the geometry of the porous scaffold microstructure is a key factor controlling mechanical function of bone-scaffold system in the regeneration process and after. A research group

claims to have found the optimal scaffold microstructure design using a three-dimensional computation simulation (voxel finite element method) of bone tissue regeneration consisting of scaffold degradation and new bone formation (170). The focus of their work was developing the computational simulation framework and a comparison with experimental results have not been conducted.

Computational methods have been used to understand cell signalling. This field is called systems biology where experimental and computational research is integrated. Cookson *et al.* (171), discovered the cell pathways activated by signalling molecules at the single cell level are heterogeneous. The authors' experimental findings agreed with computational results. Muller *et al.* (172) reconstructed an extended stem cell regulatory network from gene expression patterns of 150 samples of pluripotent, multipotent and differentiated human cells types (the 'stem cell matrix' database). Using a computational clustering technique, they found that pluripotent stem cells (ESCs and iPSCs) gene expression data clustered together. The authors also used an algorithm performing interaction and similarity module analysis (MATISSE) to identify a putative pluripotency network (called PluriNet). This algorithm searched for connected sub-networks involving pluripotency-related factors from a pre-compiled background network of human protein-protein and protein-DNA interactions including NANOG – a transcription factor involved with self-renewal of undifferentiated stem cells. Although PluriNet is undirected by the user(s) and many interactions have not been experimentally characterised in most cell types, it is still a useful method to 'project' experimentally derived datasets onto pre-compiled databases and interpret new findings from known biological processes (40). Both authors in this paragraph used their own data mining methods and discovered new knowledge from their existing data.

An example of machine learning (ML) work is the computational model of cell migration in three-dimensional matrices (173). The authors used a force-based dynamics approach. The model determines overall locomotion velocity vector for speed and direction for individual cells based on internally generated forces transmitted into external traction forces. The model also considers timescales where multiple attachment and detachment events are integrated. Model predictions agree well with experimental findings for both 2D substrata and 3D natural tissues and synthetic gels.

Others create their own models from the literature and empirical observations. N'Dri *et al.* (174) created a computational model of cell adhesion and movement using continuum-kinetics approach. The models considers molecular mechanics and macroscopic (cell-level) transport. The model is assessed using an adherent cell, rolling and deforming along the vessel wall under imposed shear flows. Experimental findings agree with the model's results and the authors discovered the intracellular viscosity and interfacial tensions directly affect the rolling of a cell. In addition, the presence of a nucleus increases the bond lifetime, and decreases the cell rolling velocity. Bigger cells roll faster and have decreased bond lifetime. In conclusion, the rheological properties of cells have significant effect on the adhesion process contrary to what has been hypothesised in the literature.

## 1.5 PROBLEM DEFINITION

The problem is that we cannot generate fit for transplantation tissues as we cannot guarantee they are free of undifferentiated cells. This makes therapy translation impossible as there is no guarantee stem cells will stop dividing and therefore give rise to a risk of teratomas (175). Our best *in vitro* cell environments rely on biologically derived materials. There is a drive to move away from such materials (112,113,176) and switch to synthetic

materials that are reproducible, with reduced cost and can be made pathogen free (90,176). Cell performance for this project is defined as cell cluster area, cell projection for neurons and glia, and cell type proportion. The benchmark cell performance is that of the *in vitro* biological environments with laminin. Cell performance observed on synthetic materials do not match that of the *in vitro* biological environment. Bioengineering strategies focus on experimental process methodologies in the laboratory. These are costly, time consuming and may also be limited by the need to use animal derived tissues (16).

An experimental method to assess stem cell performance on surfaces is to use a chemical gradient approach. Gradient surfaces presenting change in overlaid chemistry also present a small topographic feature. Although these allow experimental investigations of synergistic effects of multiple parameters, they are still limited to 2-3 variables per investigation. In addition, the presentation of a gradient may itself have an impact on the overall output observations. For example, cell migratory direction hindering differentiation potential (16).

A synthetic environment that produces cell performance closer to that of the *in vitro* biological environment has not been found yet (16). Current experimental methodologies in the lab for surface engineering rely on intuitive approximate solutions (16,176). Previous surface generations with improved cell performance serve as examples for the intuitive perturbation of the components of the cell's microenvironment (16,177). The illustration below shows the experimental process in the lab (Figure 1.11).

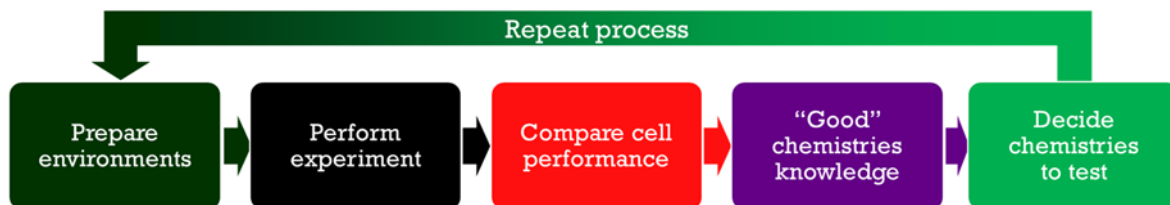


Figure 1.11: Traditional cell culture procedure in the lab.

In synthetic material design, the surface chemical properties to tune and by what degree is still unknown. We believe there is room for improvement for cell performance on synthetic environments and here we focus on finding better chemical designs that better fit the purpose. There are 13 million chemical designs to test, in theory. This means testing all of them to find a better chemistry is next to impossible given the time it takes to prepare a cell microenvironment and assess cell performance (16). It has taken a total of 6 months collectively to test 13 environments with traditional cell culture experiments in the lab. It is time to move computationally to solve this problem. Kohn, 2004 (145) mentions the adoption of computational methods in biomaterial design is the way forward for tailored materials to satisfy the requirements of biomedical applications.

## 1.6 AIMS AND OBJECTIVES

The aim is two-fold:

1. Find synthetic surface chemistries that provide better cell performance than our current synthetic standard amine and as close as possible to the *in vitro* biological environment (laminin protein coated)
2. Explain which chemical properties of the cell's environment affects cell performance and by what degree

The objectives are:

1. Find and describe the relationship(s) between chemical parameters and cell performance using computational techniques
2. Create a tool to perform cell culture experiments computationally with the use of the models discovered in the previous objective (Figure 1.12 & Figure 1.13)
3. Use the same tool as previous objective to discover the effect of one or a few chemical parameters on cell performance
4. Validate findings experimentally with cell culture experiments

Figure 1.12 below shows the end results of the conversion from real chemistries to numerical chemistries:

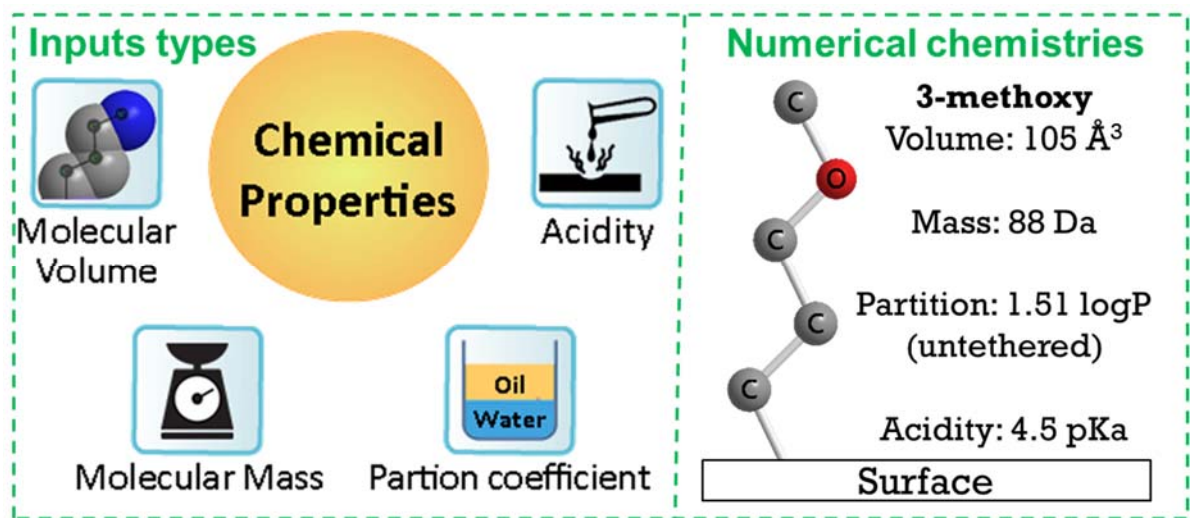


Figure 1.12: (Left) parameters defining the chemical properties of surface chemistries. (Right) A synthetic chemistry with its chemical values.

Figure 1.13 below shows a methodology schematic for computational cell culture experiments:

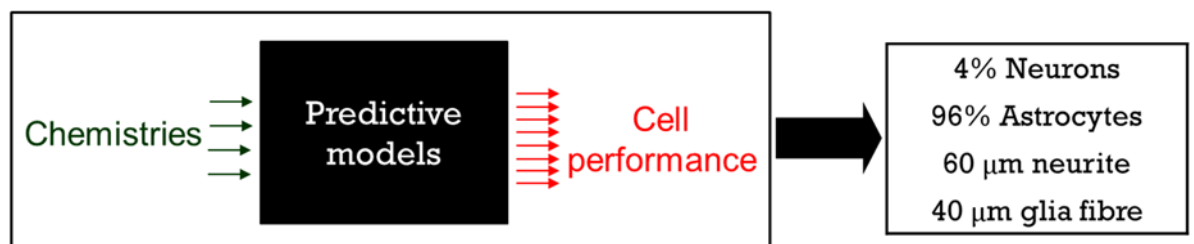


Figure 1.13: (Left) computational cell culture experiment logic. (Right) Example outcomes of cell performance.

## 2 MATERIALS AND METHODS

---

Finding surface chemistries for improved neural cell attachment and differentiation raises the question “which synthetic chemical design enhances nerve tissue engineering comparable to a biological benchmark?” Answering this effectively necessitates addressing a limitation found in the literature which is assessing numerous chemical properties simultaneously. For the experimental methodology, data collection is to perform cell culture experiments on a variety of surface chemistries and collect morphological cell response data for analysis such as correlations and machine learning. Below is the methodology to change the presenting chemistry of glass coverslips used as neural cell culture surfaces.

### 2.1 MODIFYING PRESENTING CHEMISTRY OF SURFACES

The 13 mm coverslips (Thermo) were left in 70% industrial methylated spirit (IMS) for at least 24 hours to remove dust and unwanted debris. After rinsing with isopropyl alcohol (IPA), the coverslips were air dried immediately prior to modifying their presenting chemistry. In previous work, nine synthetic chemistries were used. Some of these have also been used in recent work, thirteen in total, shown in Table 2.1 with the surface chemistries used in previous work (41) and current work indicated with a P or a C respectively in column “Use”.

For each synthetic chemistry, 30 glass coverslips were added in separate vials with 5 ml of solvent (toluene, ethanol from Fisher or tetrahydrofuran from Sigma-Aldrich) and 50  $\mu$ l of the silane solution. These were left for 24 to 48 hours for the coverslips to acquire the

functionality through a condensation reaction. Glass coverslips have hydroxyl groups at the surface, which is the bond site for self-assembly molecules (synthetic chemistries).

Modifying surfaces required a single-step process for the majority of chemistries but two, the carboxylic acid surfaces and the *in vitro* biological control surfaces with PDL and laminin. The carboxylic acid terminated surfaces required a two-step process. The carboxylic acid functionality is acquired by reacting amine (index 5 in Table 2.1) surfaces with succinic anhydride. APTES surfaces were rinsed with toluene then the coverslips were placed in a second vial with toluene and 0.005 moles of dissolved (sonicator) succinic anhydride for 48-hours to react the terminal amine to form the carboxyl terminal group above it. All modified surfaces with synthetic chemistries were:


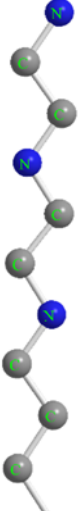

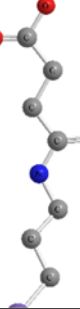
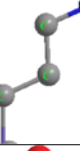

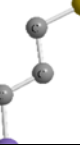
- Rinsed with the same solvent used previously then annealed for 1 hour in an oven at 150° C (178)
- Placed in well plates in LFH for cell culture or stored in a desiccator until needed

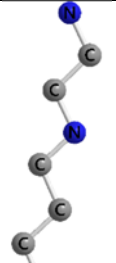

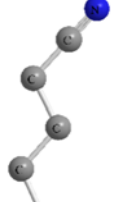

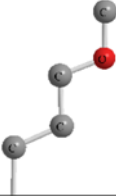
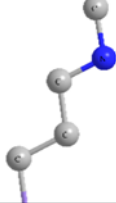
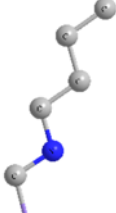
The *in vitro* biological surfaces were made ready 1 day before neurosphere micro-culture. These surfaces are made with poly-d-lysine (PDL, Sigma-Aldrich) and laminin from Engelbreth-Holm-Swarm murine sarcoma basement membrane (Sigma-Aldrich). Coverslips were placed in well plates and they were sterilised with 70% IMS then rinsed with IPA in a laminar flow hood (LFH) used for cell culture. The coverslips were dried in an oven at 150° C for 1 hour. PDL (1:10 dilution) was added on each sterilised coverslip and left for 1 hour in the LFH. PDL enhances electrostatic interaction between negatively charged ions on laminin. PDL increases the number of positively charged sites available for laminin adsorption (179). After rinsing the coverslips with sterilised distilled water 3 times, laminin (1:100 dilution) was added and left overnight in an incubator at 37.5° C. Prior to use with



cells, the laminin surfaces were washed 3 times with sterilised water and air dried for 30 minutes in the LFH.

Table 2.1: Synthetic chemistries used in previous (41) range from index 1-8 and recent work chemistries range from index 3-14.

Index	Use	IUPAC name	Structure
1	P	<a href="#">Phenyl triethoxysilane, Sigma</a>	
2	P	<a href="#">N1-(3-Trimethoxysilylpropyl)diethylenetriamine, Sigma</a>	
3	P and C	<a href="#">Methyltriethoxysilane, Sigma</a>	
4	P and C	<a href="#">4-oxo-4-((3-(triethoxysilyl)propyl)amino)butanoic acid</a>	
5	P and C	<a href="#">(3-Aminopropyl)triethoxysilane, Sigma</a>	
6	P and C	<a href="#">Triethyl hydrogen orthosilicate, Sigma</a>	
7	P and C	<a href="#">(3-Mercaptopropyl)triethoxysilane, Fluorochem</a>	

8	P and C	<a href="#">N-[3-(Trimethoxysilyl)propyl]ethylenediamine, Sigma</a>	
9	C	<a href="#">2-(Carbomethoxy)ethyltrichlorosilane, Fluorochem</a>	
10	C	<a href="#">3-Cyanopropyltrimethoxysilane, Fluorochem</a>	
11	C	<a href="#">N-(6-Aminohexyl)aminomethyltriethoxysilane, Fluorochem</a>	
12	C	<a href="#">3-(Methoxy)propyltrimethoxysilane, Fluorochem</a>	
13	C	<a href="#">N-Methyl-3-aminopropyltrimethoxysilane, Fluorochem</a>	
14	C	<a href="#">n-Butylaminopropyltrimethoxysilane, Fluorochem</a>	

## 2.2 SURFACE CHARACTERISATION

After modifying the presenting chemistry of surfaces, ensuring the chemistry has bonded on the underlying bulk material (glass) involves proving it is there using at least three techniques/instruments. A popular method is contact angle experiments where a drop of a solvent is released and the inner contact angle measured, and other, more advanced methods are Raman Spectroscopy and X-ray Photoelectron spectroscopy. In Raman spectroscopy, the sample is excited with a laser interacting with the molecular vibrations resulting in light (Raman) scattering. This causes an energy shift either up or down and gives information about the vibrational modes of the sample. Doing so provides structural fingerprint insights and molecules can be identified. X-ray photo electron spectroscopy measures the elemental composition by irradiating with an X-ray beam while measuring the kinetic energy and number of electrons escaping from the top 10 nm of the sample.

### 2.2.1 Contact angle measurements (CAMs)

All modified surfaces were measured for their water contact angles as a form of chemical verification of each chemistry. With the exception of laminin surfaces, modified surfaces of each batch of synthetic chemistries were dried in 50° C overnight. Solvents used include sterilised and filtered distilled water (dH<sub>2</sub>O) and decanol for hydrophilicity and lipophilicity measures respectively. OneAttension Theta Lite was put to focus and set to 160 frames per second to capture with a software trigger for 3 seconds once the syringe was retracted with high speed due to the designed mechanism of the instrument. Solvent drop volumes of 1 to 2 µl of dH<sub>2</sub>O and decanol respectively were measured using live pendant analysis before being dropped on modified surfaces.

Freshly made laminin surfaces and the dried synthetic chemistry surfaces were then used for decanol contact angle on the side that has not been used previously. Modified surfaces were placed on the imaging stage and images were acquired immediately and automatically by the software and trigger after the solvent drop was released at room temperature. Results were analysed with OneAttension software and the drop shape curve was fitted from subpixels using Young-Laplace equation detailed in (180,181). The baseline was corrected for each sample. The surface free energy of the solid was calculated with the following equation:

$$\gamma_{SG} = \gamma_{SL} + (\gamma_{LG} \cdot \cos \theta_C)$$

Equation 2.1: Young's equation to determine the surface free energy ( $\gamma_{SG}$ ) of solids.  $\gamma_{SL}$  is the interfacial tensions,  $\gamma_{LG}$  is the surface tension, L is liquid phase (solvent), S is solid phase, G is gas/vapour phase, and  $\theta_C$  is the contact angle (180).

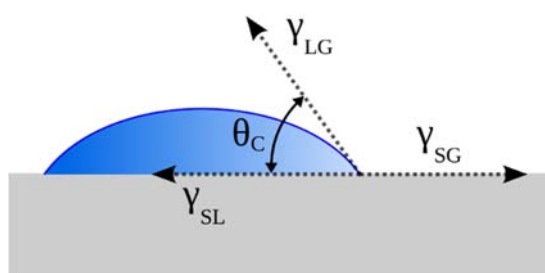


Figure 2.1: Illustration of contact angle measurement on solid surfaces (Young's equation). Blue blob is the solvent and the gray bar is the solid surface. Taken with permission from (182).

## 2.2.2 Surface Enhanced Raman Spectroscopy (SERS)

The instrument Thermo Scientific DXR Raman microscope was fitted with a laser, filter and full range grating all designed for 532 nm wavelength. After warming up the laser, alignment and calibration of the instrument was conducted as per the manufacturer's manual to enhance the Raman signal and reduce noise. Samples were handled with needle tip forceps and placed on a microscope slide. For the SERS technique, gold nano-particles (GNP) 30 nm in diameter (OD 1) stabilised in citrate buffer (Sigma-Aldrich) were pipetted in 2  $\mu$ l volume on two debris-free areas of each functionalised silicon wafer and were left

to air-dry. The slide with the samples was mounted on the stage and the laser was focused on each sample on debris-free areas.

For SERS, the focus was on areas near the edge of concentrated dried gold nanoparticle solution indicated by sky to light blue colour. The oscillating electric fields of light rays cause GNP electron charge that oscillate with greater amplitude than the frequency of visible light (183). Due to this effect and the size of the GNP (30 nm) light is absorbed in the blue-green portion of the spectrum (450 nm) (184). Samples were analysed at 4' exposure time and 75 sample exposures with 50  $\mu\text{m}$  both slit and pinhole apertures. Data was collected and analysed, normalised, smoothed and peaks were identified with OMNIC v8.2 software.

### 2.2.3 X-ray Photoelectron Spectroscopy (XPS)

The XPS analysis was performed with the Theta Probe instrument equipped with a monochromatic AlK $\alpha$  x-ray source (Thermo Scientific) as stated by national EPSRC XPS user's service (NEXUS) facility at Newcastle University. A high-energy pass (200 eV, step 1.0 eV) is performed as survey spectra and a low-energy pass (40 eV, step 0.1 eV) is performed for high-resolution spectra of the elements of interest (e.g. carbon, nitrogen, oxygen, Table 2.2). A flood gun was used for charge compensation to deal with electron loss from the sample. The XPS analysis of this project's samples was conducted at the UK National XPS facility at Newcastle University under the guidance of Prof P Cumpson. Data acquired were analysed with CasaXPS software v2.3. The table below shows X-ray energy (in electronvolts) to excite the elements of interest to this study:

Table 2.2: X-ray energy level exciting elements of interest.

Energy / eV	Element	Level
69	Br	3d
168	S	2p
284	C	1s
399	N	1s
532	O	1s

## 2.3 CELL CULTURE ON MODIFIED SURFACES

The raw biological material for nerve tissue engineering is cortical tissue was dissected from rat embryos aged 16 days. Cortical tissue was chosen for studies as cortical neural progenitors and neural stem cells can be sourced over a long period (185,186). Rat tissue is similar to human neurospheres (187) and by using it we avoid the trouble of producing high quality neurons from human stem cells. Rats gestate for 22 days (E0-E21), so E16 cortical tissue was selected for dissection as past experiments show that the first markers of mature neurons emerge around E15-E17 (186,188). E16 tissue strikes a balance between the number of cells acquired and cell lineage determination (185). The older the tissue, the more determined it becomes towards the cell lineage destined for mature cell types.

### 2.3.1 Neural tissue dissection

Tools required were sterilized using a glass bead sterilizer (Steri 250, Simon Keller AG) for approximately 15 seconds at 250°C. These include large and small scissors (Fine Science Tools), bracken forceps (Roboz Surgical Instrument Co) and Dumont forceps (Fine Science Tools). After cooling down, the tools were placed in a sterile container. Trypsin and deoxyribonuclease (DNase) solutions were thawed at room temperature. Trypsin with EDTA solution was made up of 0.1% trypsin and 0.05% DNase I (both from Worthington Biochemical Corp., Reading, UK) in Dulbecco's Modified Eagle's Medium (DMEM, Sigma-Aldrich). DNase solution was made up of 0.05% DNase also in DMEM. Both solutions were

kept in 1.5 ml Eppendorf tubes at  $-20^{\circ}$  C. 50 ml tubes (both Greiner Bio-One) were filled with DMEM medium and placed on ice.

Sprague-Dawley rats bred in-house at Keele University were sacrificed by approved Schedule 1 methods, following guidelines from the UK Animals, Scientific procedures Act, 1986 and authorization from Keele University's local ethics committee. The embryos were 16 days old (E16) with E0 defined as date of observing vaginal plug. After Schedule 1, the peritoneum was opened using scissors. First, horizontally through the skin then vertically to expose the uterine horns containing the rat embryos. Sterilized forceps were used to lift each uterine horn whilst a pair of small scissors was used to trim off tissue attached to the abdomen. Uterine horns were transferred into a 50 ml tube and were transported to the dissection hood. Uterine horns were then transferred into a sterile 100 mm petri dish (Greiner Bio-One). Scissors and forceps were used to remove one embryo at a time and to transfer embryos into a petri dish with DMEM medium.

The following dissection steps were performed with a dissection microscope (Leica DMIL Inverted Phase Contrast Microscope). Rat brains were then extracted after decapitating embryos (recognised Schedule 1 methods for embryos older than 11 days old). Scissors and forceps were re-sterilized using the bead sterilizer for 15 s. The embryo heads were split in 3 groups placed on their side in the petri dish, a vertical cut was made (Figure 2.2A) using a pair of small scissors (Fine Science Tools) and the brains were extracted (Figure 2.2B). Once ready, each of the brain groups was transferred into a petri dish containing DMEM medium (189,190). A longitudinal cut was made along the medial dorsal cortex, close to the midline using fine Vannas dissection scissors (Fine Science Tools, Figure 2.2C). It was opened up to reveal and remove the ganglionic eminence (heart-shaped structure shown in (Figure 2.2). The surrounding tissue remaining is the cortex and this was collected. Tissue

pieces were transferred into designated small Eppendorfs containing DMEM for each brain group, kept on ice.

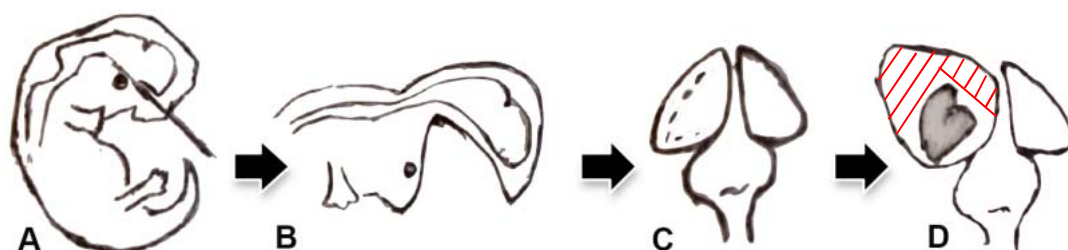


Figure 2.2: Dissection of E16 brain tissue. A cut is made above the eyes to expose and extract the rat brain (A, B). Another cut is made (C) to expose and remove the ganglionic eminence (D) and the cortex tissue (area shown with hashed lines) is collected. Adapted with permission from (191).

The dissected cortical pieces were digested to form single cells with 0.1% trypsin + 0.05% DNase in DMEM for 30 minutes at 37 ° C. A pellet formed and the trypsin solution is aspirated away followed by 3 washes of 200  $\mu$ l of DMEM with 0.05% DNase (Worthington Biomedical Corp) to digest extracellular nucleic acid released by lysed cells making the solution less viscous. Once the DNase is aspirated and the cells washed with DMEM, they were centrifuged for 3 minutes at 55 g (RCF) and aspirated once again to remove any residual enzyme. The cell pellet is suspended in media and mechanically dissociated to break the tissue to single cells.

### 2.3.2 Neurosphere expansion

The next step after neural tissue dissection is to expand single cells to neurospheres to maximise the number of neural stem cells and progenitors (51). Carrying from the final step of tissue dissection, the supernatant of the tube containing a pellet of single cells was aspirated. The pellet was then re-suspended in 1 ml of neural progenitor culture media (NPC, Table 8.1 in appendices section 8.1) to quench proteolytic activity of any residual trypsin following centrifugation. Cell counts were performed with a haemocytometer and T25 flasks (Greenier Bio-One) were seeded with 1 million cells/ml. Once seeded with cells, 5 ml of NPC media was added. NPC media contains  $\beta$ FGF (stem cell mitogen, Gibco aa 10-



155) promoting neurosphere formation due to cell proliferation (45). Resulting flasks were incubated at 5% CO<sub>2</sub> at 37 ° C. After 48 hours, an additional 2 ml of NPC was added to account for neurosphere growth. At this point and every 48 hours, 2 ml NPC media was replaced. For medium exchange, the T25 flask was placed upright to rest for 5 minutes to sediment the neurospheres and prevent their accidental removal.

### 2.3.3 Neurosphere passage

From a review on neurosphere cultures (192) it was discussed that neural progenitors isolated from the developing brain have potential for use in replacement therapies but suffer from limited availability and ethical concerns. Neurosphere-expanded cells are not easily committed to a neuronal fate and that expression of one gene normally involved in neuronal commitment is not sufficient to promote neuronal differentiation in a complex environment (192). This means additional methods are required to maximise the number of cells within a neurosphere. Passaging for neurospheres is essentially breaking them in smaller parts as this obviously controls their size and more importantly the cell types within the neurosphere. The reasons behind the passage include:

- to decrease the chances of necrotic cells at the centre of neurospheres from insufficient supply of nutrients from media
- to increase uniformity in neurosphere size
- for easier micro-culture technique and analysis
- to be able to make size comparisons with sphere spreading at designated time points in culture

For the last point, sphere spreading holds information of cell migration away from the sphere. In traditional cell culture, passaging is the process where cells are detached and transferred to fresh media. In the context of neurospheres, passaging is the process of

splitting up the spheres into single cells and transferring them to fresh media to reset their size. Remaining suspended single cells form new neurospheres if they are neural stem cells and progenitors.

Neurospheres were passaged after 7 days in culture. The neurospheres with NPC media were taken from the T25 and centrifuged at 55 g (RCF) for 5 minutes to create a pellet. The NPC was aspirated and the neurosphere pellet was re-suspended in 0.5 ml of fresh NPC. This was transferred to a 1 ml Eppendorf tube and neurospheres were dissociated mechanically to single cells with the pipetting technique. The single cell solution of 0.5 million cells/ml was transferred to a fresh T25 with 5 ml of NPC media to restart the neurosphere formation process. The T25 was incubated (37 ° C, 5% CO<sub>2</sub>) for 2 days.

#### 2.3.4 Neurospheres micro-culture

Micro-culture is a method to miniaturise cell culture experiments therefore increasing the scale of the study. It addresses the issue of having limited biological material such as cells and proteins when testing experimental conditions. Using this method, neurospheres must adhere to the modified surfaces and not the well plate before the well is filled with media.

The P1.2 (passaged once plus 2 days re-expansion) neurospheres were taken from the T25 and centrifuged at 55 g (RCF) for 5 minutes. The supernatant NPC was aspirated and the neurosphere pellet was re-suspended in 2 ml differentiation culture media containing fetal calf serum to promote cell differentiation. Neurosphere counts were performed using a haemocytometer. A stock solution of neurospheres and differentiation media was made to provide 30 µl micro-cultures containing 200 neurospheres. During the process of seeding cells on modified surfaces, the micro-culture solution was continuously rocked manually to

prevent the neurospheres from settling. 30  $\mu$ l was pipetted on the centre of each dried modified surface placed in individual wells in a 24 well plates. For each type of modified surfaces, 3 were used for each time point (6 in total), per experiment. The seeded surfaces were incubated at 5% CO<sub>2</sub> at 37 ° C for two hours. An additional 0.5 ml of differentiation media was added to every well and the next morning, the wells were topped up to 1 ml. From that point and for every 48 hours, 0.5 ml of differentiation media was replaced for each well. Micro culturing was performed on all surfaces as shown in Figure 2.3 below:

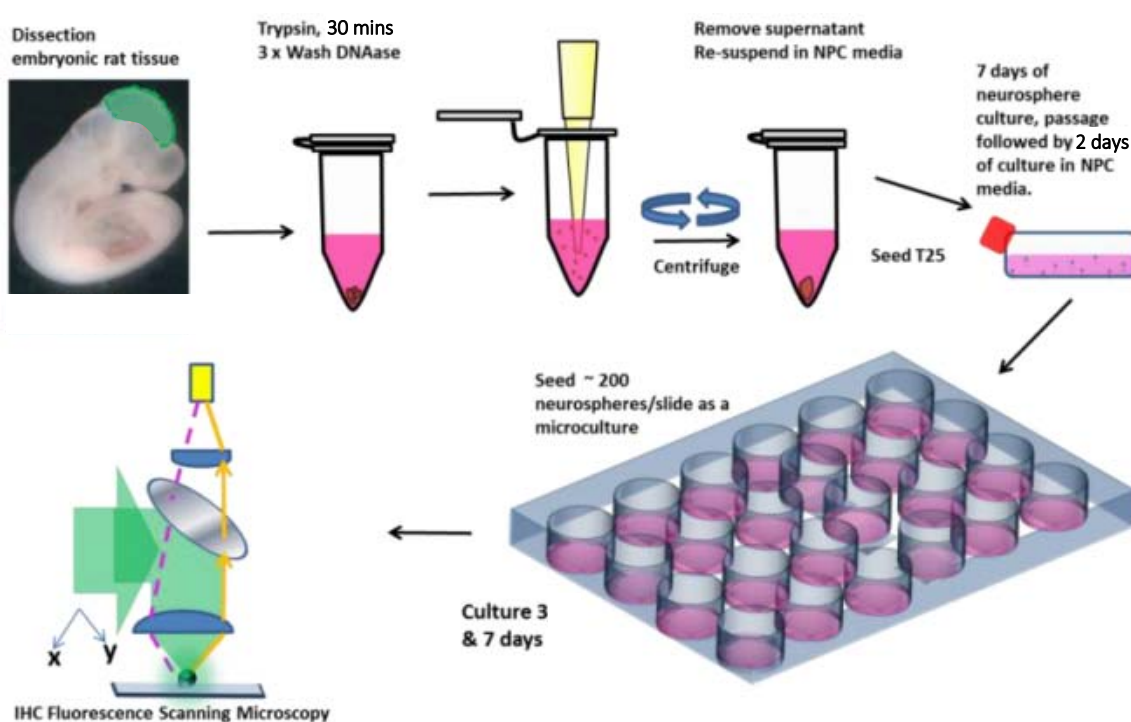


Figure 2.3: Workflow for cell culture on modified surfaces. Starts from top left. From cortical tissue dissection, dissociating tissue to single cells, resuspending cells in neural progenitor culture media in T25 flasks and left to grow for 7 days. Followed by a passage then, 2 days later seeding to modified surfaces. Samples were cultured for up to 7 days and samples were fixed and mounted on microscope slides for image capturing and analysis. Adapted with permission from (41).

## 2.4 FIXATION AND IMMUNOCYTOCHEMISTRY

Fixation is a critical to preserve biological tissues from decay (autolysis) and prepare them for immunocytochemistry (ICC). Fixation halts ongoing biochemical reactions and increases the mechanical stability of the treated samples. ICC is common technique used to identify

cell types using fluorescent light on samples. Anatomical visualisation of specific localised proteins (antigens) in cells is possible using a primary antibody that binds to them. The principle is simple. A primary antibody (e.g. murine  $\beta$ -tubulin) attaches to an epitope on a cell structure. Next, a secondary antibody (e.g. goat anti-mouse) is added and this contains a fluorophore. The fluorophore can be excited with fluorescent light from a UV light source and antigen positive cells will fluoresce.

At day 3 and 7, cells were fixed for ICC. In the first step, media was aspirated from the wells and cells were fixed with of 4% paraformaldehyde (PFA) solution for 20 minutes at 4°C to better preserve cell morphology (193). After 3 washes with tris buffer solution (TBS, 12 g trizma base from MERCK, 9 g NaCl, 1 L dH<sub>2</sub>O), the wells were inspected for the presence of cells under a standard upright lab microscope. With ICC, the antibodies can bind undesirably on plastic wells, other cells (false-positives) or debris. This non-specific binding issue is dealt with by 'blocking' binding sites on the cell's environment and cells with a serum. Samples were blocked for 1 hour at 4 ° C with a solution containing goat serum (1:20), Triton X (1:500) to digest lipids in cell membrane and allow antibody penetration, and TBS. After 3 washes with TBS, the primary antibody solution containing  $\beta$ -III-tubulin (neuronal marker, 1:500 dilution, Cambridge biosciences) and GFAP (clonal glial marker for astrocytes, 1:1000 dilution, DAKO) was added to bind with mature neural phenotypes. The composition of ICC solutions can be found in Table 8.2, in appendices section 8.2. The samples were incubated with primary antibody solution overnight at 4 ° C. Following 3 washes with TBS, a secondary antibody solution was added containing the FITC and TRITC fluorophore tagged antibodies (Cheshire Sciences). The samples were left in the dark for two hours followed by 3 washes with micro-filtered dH<sub>2</sub>O. After that time, the samples were mounted facing down on microscope slides with DAPI mounting media (Vector Labs). DAPI is a fluorescent dye that binds to nuclear material within the nucleus. Finally, the

samples were sealed with clear nail varnish around the edges to protect them from drying and fading, and to stabilise them on the microscope slide.

## 2.5 MICROSCOPY

Measuring morphological cell performance is performed from cell culture images captured using life sciences microscopes. These are instruments designed for measuring distances and cell counts where the area of view is defined with a pixel-to-distance calibration. In this sense, with the appropriate calibration, it is possible to take measurements and annotations digitally and convert these to real units such as  $\mu\text{m}$ , mm, and counts.

### 2.5.1 Epi-fluorescence

Optical fluorescence microscopes work by exciting previously deposited (ICC) antibodies or dyes with light (fluoresce) and these would emit light at different wavelengths (phosphoresce). High intensity light is split into fluorescent light wavelengths with a fluorescence filter. Fluorescent light makes its first pass through the dichroic filter (beam splitter) allowing light of a specific wavelength through. The fluorescent light shines on the sample exciting fluorescent material to phosphoresce. The emitted light goes through the objective lens and magnifies the sample in view, and then through the dichroic mirror. The emission is detected by an ICCD camera (intensified charged coupled device) to provide an image (Figure 2.4).

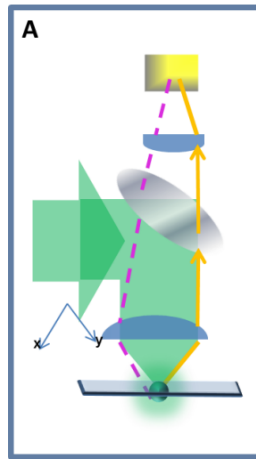


Figure 2.4: Epi-fluorescent microscopy schematic. Adapted with permission from (41).

## 2.5.2 Morphological cell performance measurements

Phosphorescent images of coverslips with neurons and glia were captured to assess attachment, migration, differentiation, and cell process elongation. The samples were scanned with an automated scanning XY stage epi-fluorescence Nikon Ti microscope (Nikon Instruments) with 5% overlap percentage. Scans were taken using a 100x objective lens with a monochrome Hamamatsu ORCA CCD camera (Hamamatsu Photonics). In addition, 200x and 400x images were taken for cell counts and to examine cell projections. Three filters listed in Table 2.3 (below) were used with as short exposure time as possible. All scans were acquired with 1x gain.

Table 2.3: Excitation and emission wavelengths of fluorescence microscopy used for this project.

Filter	Excitation/emission wavelength	Colour
DAPI	358/461 nm	Blue
FITC	488/518 nm	Green
TRITC	541/572 nm	Red

Image analysis to quantify cell performance was performed with NIS elements v3.2 (x64) software. Cell responses quantified include:

1. Cell cluster area. Upon adhesion, neurospheres spread on the culture surface and form a cluster. This cluster area was measured in  $\mu\text{m}^2$ .

2. Cell proportions for neurons, glia, and unknown type cells. From all counted cells (150-200 per coverslip), the proportion of cell types was determined.
3. Type I Astrocyte area. *In vitro*, spreading of type I astrocytes indicates they are under stress. *In vivo*, this is called reactive astrogliosis and it could arise due to injury to the nervous system (194). Type I astrocyte spreading area was measured in  $\mu\text{m}^2$ .
4. Cell projection lengths. These are neurites for neurons and astrocyte fibres for glia (measured in  $\mu\text{m}$ ).

Data was imported into Microsoft Excel for validation, manipulation and export.

## 2.6 STATISTICAL ANALYSES

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organisation of data. It is necessary to use statistical analyses as the sample size is rarely the population size. An oversimplification of statistical tests is they tend to measure the risk or how “wrong” one can be with made assumptions.

### 2.6.1 Variance tests

Variance tests are used to investigate whether the variance in data is homogeneous, satisfying a parametric assumption before choosing further parametric tests. Parametric statistics have a fixed length on parameters whereas non-parametric do not. Levene’s test checks the null hypothesis that the variances in different groups are equal (i.e. the difference between the variances is zero). This test does a one-way analysis of variance (ANOVA) conducted on the deviation scores; that is, the absolute difference between each score and the mean of the group from which it came (195). Equal intervals on the variable represent equal differences in the property being measured (e.g. the difference between 6

and 8 is equivalent to the difference between 13 and 15) (195). Levene's test uses the mean and has better statistical power for symmetric, moderate-tailed, distributions (196). The absolute deviations were used with this test to assign equal weights of data spread (195). The squared deviation would have emphasised the outliers (197). The Brown–Forsythe test on the other hand uses the median instead and is recommended as the choice that provides good robustness against many types of non-normal data while retaining good statistical power (195).

## 2.6.2 Data distribution

Another parametric assumption to use parametric tests is how normally distributed the data is. A normal distribution is has most of the data points cluster around the mean with the number tapering off symmetrically either side of the mean to a few extreme values in each of the two tails. While the assumption of normally distributed data is not the only one that must be satisfied in order to use parametric tests, the arithmetic of such tests is based on the parameters describing a symmetrical, bell-shaped curve (Gaussian) (196). Normality tests can determine whether sample data has been drawn from a distribution that is approximately normally distributed (195). Shapiro–Wilk and Anderson-Darling tests for normality are popular in the literature (198–200). The authors found Shapiro-Wilk provides a superior omnibus indicator of non-normality judged over the various short/long-tailed, asymmetric and symmetric alternatives and over diverse sample sizes used (198). Anderson-Darling test detects non-normality around the tails of a distribution (199,200) whereas Shapiro-Wilk is better around the centre of a distribution (198). Chen-Shapiro test extends the power of Shapiro-Wilk without losing power and also supports limited sample size (198,199).



If we add together many random variables with all having the same probability distribution the sum, as new random variable, will have a distribution that is approximately normal (Central Limit Theorem) (195). This theoretical basis explains the reason why so many variables in nature appear to have a probability distribution that approximates a bell-shaped curve (Gaussian). Random biological processes can often be viewed as being affected by a great number of random processes with individually small effects (201). The sum of all these random components creates a random variable that converges on a normal distribution regardless of the underlying distribution of processes causing the small effects (202). A Q-Q plot serves as a supplementary method checking for normality visually. It plots the data set in equal portions (quantiles) (195). Razali & Wah (2011) mention that Q-Q plots is an effective and common tool for visual inspection of data distribution (199) (Figure 2.5).

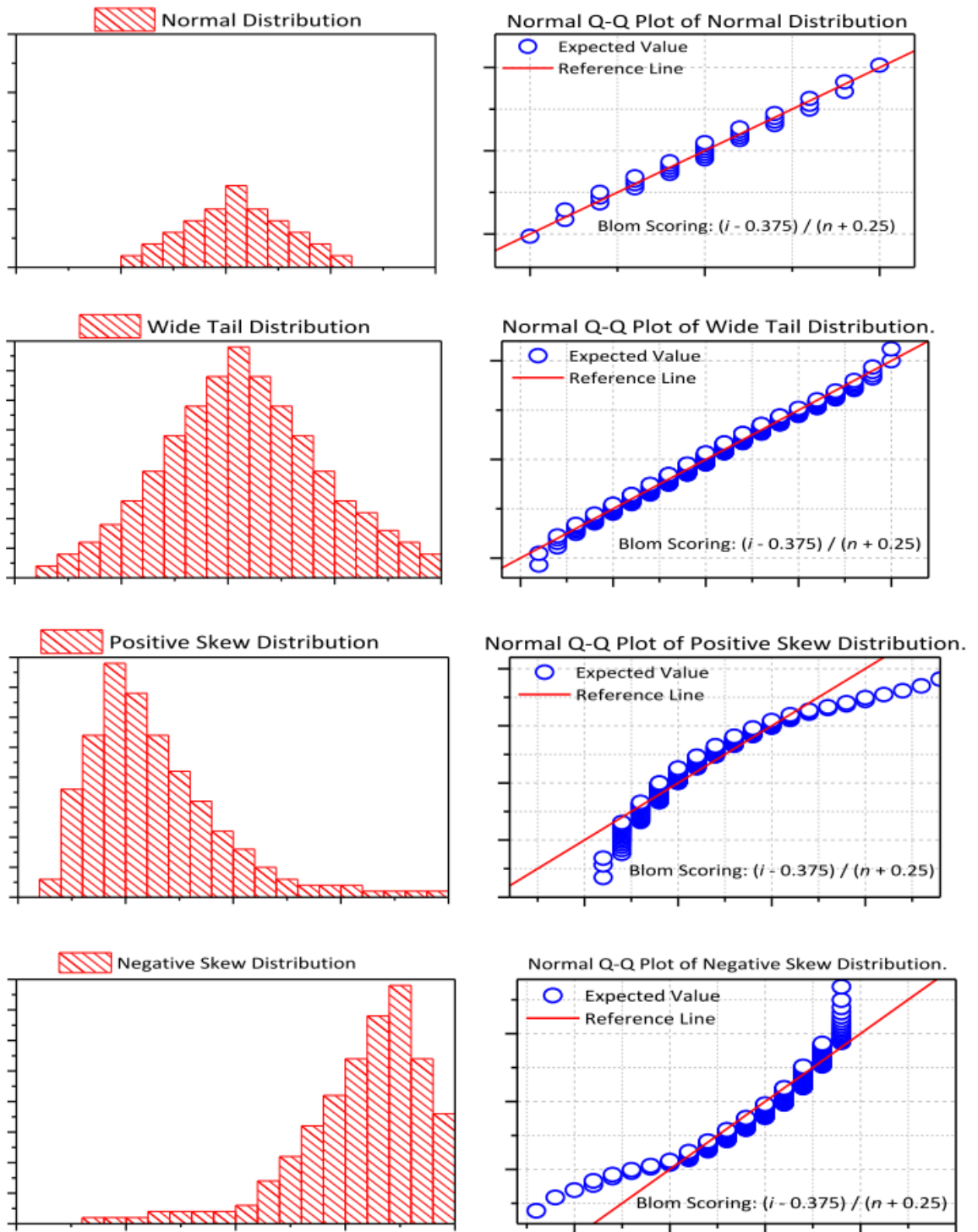


Figure 2.5: Data distributions and corresponding QQ-plots. Taken with permission from (41).

### 2.6.3 Correlation

Discovering better synthetic environments for use in nerve tissue engineering can be accelerated computationally by capturing the relationship between the surface chemistry and morphological cell responses. This will allow the exploration of environment

candidates from their chemical inputs (numerical). The selection of the best candidate (environment) from a set of alternatives is called mathematical optimisation.

The simplest form looking for relationships between two variables is correlation tests. Dependence is any relationship between two variables or sets of data. Correlations can be measured with the use of different indices (coefficients) (195). The coefficient value can range from -1 to +1 and this value tells us the strength and direction of correlations. The direction is indicated by the sign of the coefficient. A positive coefficient means both variables tend to increase together. If one variable tends to increase as the other decreases, the coefficient is then negative (195,203). It can inform which surface properties vary with cell performance to help understand how to better design biomaterials. The two most popular techniques are the parametric Pearson's coefficient ( $r$ ), and the non-parametric Spearman's rho coefficient ( $\rho$ ) (196,203).

Pearson's  $r$  is a correlation measure of a linear dependence between two variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable (Equation 2.2). This type of correlation is a parametric test where parametric assumptions are satisfied with the data in question.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)} \sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

Equation 2.2: Pearson's correlation equation (195).  $X$  and  $Y$  are the independent and dependent variables respectively.

Previous data for the first experimental chapter had missing data. This was dealt with pairwise deletion to minimise data loss compared to using listwise deletion (complete row). Pairwise deletion maximises all data available in analyses. Missing value pairs of either

dependent ( $Y_i$ ) and independent ( $X_i$ ) variables are deleted right before a test (e.g. correlation) and this can lead to absurdities in other tests (195). For practical reasons in bivariate correlation, the same number of values for both  $X$  and  $Y$  were used (195).

### 2.6.3.1 Effect size of correlations

A high valued test statistic does not necessarily mean the effect it measures is meaningful or important. To address this criticism we can measure the effect size we are testing in a standardised way. The effect size value for the correlation is the coefficient itself. An effect size is simply an objective and (usually) standardised measure of the magnitude of observed effect (195). Cohen (1988, 1992) suggested widely used effect sizes (195):

- $r = .10$  (small effect): This effect explains 1% of the total variance.
- $r = .30$  (medium effect): Accounts for 9% of the total variance.
- $r = .50$  (large effect): Accounts for 25% of the variance.

The effect size is intrinsically linked to three other statistical properties:

- 1) the sample size of population ( $n$ ) (195,196,204)
- 2) probability level (alpha value,  $\alpha$ ) and
- 3) statistical power ( $\beta$ ) (195,196).

If we know three of these properties then we can calculate the remaining one including the coefficient.

### 2.6.3.2 Correlation significance

Bivariate correlations are a measure of strength of a relationship between two variables. Any relationship should be assessed for its significance as well. This significance is expressed in probability levels ( $p$ ) and it tells how unlikely a given correlation coefficient will occur given no relationship in the population. This is also known as hypothesis testing and it tells us the confidence we are not accepting false positives (type I error). There is

also the danger of accepting false negatives (type II error) (195,196). Below is a table with the correctness and errors of hypothesis testing:

Table 2.4: Hypothesis testing correctness and errors.

		Truth	
		Null hypothesis True	Null hypothesis False
Decision	Reject Null Hypothesis	Type I Error (false positive)	Correct Decision
	Fail to reject Null Hypothesis	Correct Decision	Type II Error (false negative)

### 2.6.3.3 Type I errors and $\alpha$ -value

The probability ( $p$ ) value is usually decided beforehand and is the threshold for which the null hypothesis will be rejected if the value falls below it (false positive). This value is denoted as  $\alpha$  (alpha) value and is called significance level. Values for probability depends on the application of the outcome of the test, e.g. in drug research the  $p$ -value is 0.01 for 99% confidence and for other applications a  $p$ -value of 0.05 for 95% confidence level is more common and the minimum value as per Fisher's criterion (195,196). For example, we could take a correlation coefficient value of 0.5 with a  $p$ -value of less than 0.05 and be 95% confident the correlation coefficient differs from 0.

### 2.6.3.4 Type II errors and $\beta$ -value

Cohen (1992) suggested an acceptable  $\beta$  (beta) probability value is 0.2 (or 20%) (false negative). The corresponding level of power is  $1 - \beta$  (195,196,204). This gives 80% chance of not accepting a false negative. This means that if we took 100 from a population in which an effect exists on all, we would not be able to detect an effect in 20 of those samples (195,196).

There is a trade-off between the two errors; to make a Type I error (false positive) there has to be no effect in the population, whereas to make a Type II error (false negative) the opposite is true; there has to be an effect we overlooked. So, as the probability of making a Type I error decreases, the probability of making a type II error increases (195,204). The easiest way to minimise the occurrence of both errors is to increase the population sample size (204).

The table below shows the estimated power of Pearson's correlation coefficient with given sample size and  $\alpha$ -value (0.05 or 95% confidence):

Table 2.5: Estimates of power ( $1 - \beta$ ) of Pearson's correlation coefficient given effect size ( $r$ ), sample size ( $n$ ), and  $\alpha$  ( $p$ -value).

$\alpha = 0.05$ Two Tailed					
Effect Size: $r$					
$n$	0.10	0.30	0.50	0.70	0.95
10	0.03	0.11	0.29	0.63	0.99
11	0.03	0.12	0.33	0.69	0.99
12	0.04	0.14	0.37	0.74	0.99
13	0.04	0.15	0.40	0.78	0.99
14	0.04	0.16	0.44	0.82	0.99
15	0.04	0.17	0.47	0.85	0.99
16	0.04	0.19	0.50	0.88	0.99
17	0.05	0.20	0.53	0.90	0.99
18	0.05	0.21	0.56	0.92	0.99
19	0.05	0.22	0.59	0.93	0.99
20	0.05	0.24	0.61	0.94	0.99
21	0.05	0.25	0.64	0.95	0.99
22	0.05	0.26	0.66	0.96	0.99
23	0.06	0.27	0.69	0.97	0.99
24	0.06	0.28	0.71	0.97	0.99
25	0.06	0.30	0.73	0.98	0.99
26	0.06	0.31	0.75	0.98	0.99
27	0.06	0.32	0.76	0.98	0.99
28	0.06	0.33	0.78	0.99	0.99
29	0.06	0.34	0.80	0.99	0.99
30	0.07	0.35	0.81	0.99	0.99
31	0.07	0.37	0.83	0.99	0.99
32	0.07	0.38	0.84	0.99	0.99
33	0.07	0.39	0.85	0.99	0.99
34	0.07	0.40	0.86	0.99	0.99

For statistical power analysis G\*Power v3.1 was used for bivariate normal correlation in post hoc type of analysis. This type computes achieved power given  $\alpha$  probability ( $p = 0.05$ ), correlation sample size ( $n = 7$ ), and correlation effect size ( $r/\rho$ ). The power of each correlation coefficient is mentioned under each correlation graph under *Power* ( $1 - \beta$ ).

#### 2.6.3.5 Standard error of correlation

This standard error (SE) is a measure of dispersion from the correlation coefficient (195). It is calculated as:

$$SE = \sqrt{\frac{1 - r^2}{n - 2}}$$

Equation 2.3: Correlation coefficient standard error. This is the measure of dispersion from the correlation coefficient.

The correlation coefficient ( $r$ ) observed within a sample of  $XY$  values can be taken as an estimate ( $R$ ) of the correlation that exists within the general population of bivariate values from which the sample is randomly drawn.

## 2.7 DATA MINING AND MACHINE LEARNING

Machine learning is performed using learning algorithms to learn predictive/prescriptive models among other types. Correlation finds relationships between two variables, however learning algorithms can find relationships for more than two e.g. multiple regression, logistic regression. The data preparation step is crucial for machine learning and usually takes the longest when building a data driven solution. This is because it will influence the model fit and future-predictive ability (generalisation performance).

### 2.7.1 Data collection and aggregation

The predictors of the model are chemical descriptors of the surface chemistry. These chemical descriptors were decided from previous work (41,90,205), other relevant research groups (77,176) and the literature (206–210). They were collected using a variety of software and methods. For synthetic chemistries, the top 5-6 atoms of the backbone and branched side chains were considered.

- Partition coefficients ( $\log P$ ) for synthetic chemistries were calculated with ACD/ChemSketch 2016 for each backbone atom and side chains attached to it. This gave 5 levels of  $\log P$  values. ACD/ChemSketch 2016 is a variant of AlogP calculation



method extensively compared among many others in (106,209). For protein logP calculation, Ghose & Crippen's method (211,212) was used. Laminin constituents were downloaded from UniProt.org ([link](#)) and converted to PDB format using OpenBabel v2.4.1. VEGA ZZ v3.1 (213) ([publications](#)) was used with NAMD energy minimisation for proteins (214–217). This [guide](#) written by the author of the software was followed.

- Acidity measures (pKa) were obtained from the literature (218–221) and for proteins, the popular ProPKA v3 (206–208,222) was used.
- Molecular mass is calculated as the sum of atomic weights of each constituent element multiplied by the number of atoms of that element in the molecular formula. Laminin's mass was obtained from the literature and this agrees with the manufacturer's specification (223).
- Molecular volume for synthetic chemistries was found in the literature (224–226) or calculated with ChemDraw 2015. For proteins, ProteinVolume v1.3 was used (227).

Obtained values for each chemical input were added in columns. Each row was a different chemistry. Cell data were aggregated from previous work (41) and from cells culture on modified surfaces. These data include cell migration, and morphology. Multiple cell measurements were available for each chemistry so the chemical inputs were duplicated. The cell data were designated with the time point in binary (dummy variables) and chemical data corresponding to them. These were saved in a flat csv file.

## 2.7.2 Dataset selection, cleaning, and pre-processing

Although not a chemical parameter, the cell culture duration (time point) parameters were treated as temporal indicators. Cells remodel their environment in time (59) so an

indication of the time in culture helps handling the data. It also allows to include data from both time points. After collecting raw cell data, the central tendency was decided based on the sample distributions. There are 18 instances of data, one for each coverslip per chemistry. 9 of them are for day 3 samples and other 9 for day 7.

Surface contact angle measurements were excluded as input features. These would be impossible to acquire for theoretical (numerical) chemistries as not all of these can be made or are stable to be used. In addition, the 'ideal' contact angle would always show on top on the ranked numerical chemistries regardless of the chemical design. Chemical inputs were projected from raw to their  $\log_{10}$  and root (data not shown). These methods retain as much information as possible which is necessary for laminin's chemical values.

## 2.8 MODEL SELECTION

Computational models provide the ability to predict future cell outcomes without performing the actual cell culture experiment. This is necessary as there is a plethora of surface chemical designs to test with cells and there are resource limitations, especially time. The predictive models can be interpreted to give insights on which variables are used to make a prediction.

Predictive models are discovered with Waikato Environment for Knowledge Analysis (WEKA). The popular WEKA workbench developed at Waikato in New Zealand, is a collection of state-of-the-art machine learning algorithms (classifiers) and data pre-processing tools implemented in Java (228–230). It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating

learning schemes statistically and visualising the input data as well as the result of learning. Due to its ease of use and abundance of algorithms provided, WEKA has accumulated 13k citations leading to mid-2017.

There is an abundance of machine learning algorithms and each has specific parameters tuning a model's predictive performance on future outcomes. Finding the optimal set of both, requires problem domain knowledge, know-your-data, theoretical basis of algorithms, and value ranges of hyper-parameter boundaries tied with theory and discovered empirically.

Standard practices are to split the complete dataset into two parts. 70% for training and validation to find classifiers and optimise their parameters and the remaining 30% for testing on unseen data to calculate the error rate of the final, optimised method (231). Further splitting the training and validation set is required and when data is limited, the threshold for doing this is a dilemma. There is a trade-off between using more training data for potentially better knowledge representation and using more data for better testing of the model or for hyper-parameter discovery.  $k$ -fold cross-validation maximises the use of data for both training and testing by binning data to  $k$  bins of equal size (e.g.  $n_{tot} = 200$  then  $k = 10$  bins each consisting of  $n_k = 20$ ).  $k$  separate learning experiments are run and in each, one  $k$  subset is selected for validation and the remaining  $k - 1$  subsets are used for training. This is repeated  $k$  times averaging the validation results for all  $k$  experiments. Although more computationally expensive this maximised the use of data for model selection (231).

MultiSearch is a WEKA package that allows the testing of multiple learning algorithm parameters in order to find the best values. It chooses these values for each parameter

using DefaultSearch. This performs a 2-fold cross validation across the initial space to determine the point with the lowest mean absolute error (error metric to optimise). This is the centre point and now 10-fold CV evaluates adjacent parameter values. If better parameters are found, they are set as the new centre and the search continues until no further improvement can be found (king of hill-climbing). For numerical parameters, the MathParameter was set to start from the minimum value working its way up to the max with an increment value (STEP). For non-numerical values such as true, false among others, were tested with ListParameter. For groups of parameters such as those of Support Vector Regression with Kernels, ParameterGroup was used. MultiSearch reports the best classifier parameter setup by calling “multiSearch.getBestClassifier()”.

A dataset consisting of 10 chemistries and their corresponding cell responses was prepared as per section 2.7.1. Each cell output needs its own predictive model therefore, 8 models were sought. 10-fold cross-validation was selected as each chemistry has 9 instances for each two time points, with a total number of samples  $n = 180$  for training and validating classifiers.

### 2.8.1 Model performance

The model performance is a collection of prediction metrics used to assess predictive ability a.k.a. generalisation performance. Mean absolute error (MAE) is the average of the absolute differences between  $n$  predictions ( $p$ ) and actual values ( $a$ ) and this measure is widely used in machine learning (231). For this project, the MAE on its own is insufficient to assess model performance. The cell outcomes in experimental results have a central tendency and a spread. For the former it is usually average, median or trimmed average and the latter is standard deviation. Inherent biological variation goes together with

experimental methodologies. It is standard practice to include at least 3 replicates of a test condition and perform at least 3 experiments for outcome robustness. Inspired from the MAE and biological variation, we believe a better metric is a measure that uses the difference between actual and prediction proportional to the average standard deviation of all observed samples. This model performance ratio (MPR) is calculated as:

$$\frac{|\tilde{y} - \hat{y}_i|}{\bar{\sigma}}$$

Equation 2.4: Model performance ratio. The absolute difference between real median and prediction is standardised by the spread of data, the average standard deviations.  $\tilde{y}$  is the median of real data,  $\hat{y}_i$  is the prediction estimate, and  $\bar{\sigma}$  is the average standard deviation of all observed samples.

## 2.8.2 Attribute evaluation and selection

Variable selection for machine learning is case-specific meaning for each computational problem a different set of data may provide predictions with lower MAE for example. The data used directly affect the learning of the model and experimentation to find the best set is good practice.

Correlation feature subset evaluation (CfsSubsetEval) (231,232) is a method that evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. If the  $-L$  switch is not set, the acceptance of a feature will depend on its ability to predict the class if they have not already been predicted by other features. Its function is as follows:

$$M_S = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}$$

Equation 2.5: Attribute evaluation method: correlation feature subset evaluation (232).  $M_S$  is the heuristic merit of a feature subset  $S$  containing  $k$  features.  $\bar{r}_{cf}$  is the mean feature-class correlation ( $f \in S$ ).  $\bar{r}_{ff}$  is the average feature-feature inter-correlation.

*CfsSubsetEval* evaluates and gives merit scores to attribute subsets found by other search algorithms. One method is *BestFirst* search (231,233) that uses greedy hill-climbing technique where it incrementally changes a single element in the attribute subset until a better one is found. Once found, the process repeats until no further improvements can be found. *BestFirst* terminates when the performance starts to decline but keeps a list of all attribute subsets evaluated along with their performance measure. This technique is called backtracking and it allows the algorithm to revisit an earlier subset configuration. The  $-N$  parameter is the number of features to keep.

The other one is *GreedyStepwise* (231,234) shares the greedy trait searching through the space of attribute subsets. Like *BestFirst*, it can progress forward from the empty set or backward from the full set. Unlike *BestFirst*, it cannot backtrack but it does terminate as soon as adding or deleting the best remaining attribute decreases the evaluation metric. In the ranking mode ( $-R$ ), the search is forced to the far side of the search space to go through all subsets. Attributes continue to be added even if the addition reduces the merit of the current best subset. At each stage, the best attribute is added. At the point where additions begin to reduce the overall goodness, the attribute that degrades the subset the least is added. The ranking is determined by the order in which attributes are added. The  $-N$  parameter works only in ranking mode and allows the best  $N$  attributes from the ranked list to be retained.

### 2.8.3 Cell cluster area

Modelling cell cluster area was achieved with *RandomTree* (235). Decision trees owe their name to their tree-like structure. The paths from root to leaf represent classification rules. Classification of examples start at the top node – the root – and the value of the attribute

it corresponds to this node is logically tested. The example moves down the branch to another node that corresponds to a particular value of an attribute and this repeats until the example reaches the end node – the leaf – and instead of a logic test, a value of the target attribute is given. All examples arriving at the same leaf, the same target value will be predicted. Generally, the upper the attributes are on the tree the stronger influence on the target variable.

In RandomTrees, trees are learnt from top to bottom with an algorithm known as divide-and-conquer (DAC). In DAC, the problem is recursively broken down into sub-problems until these become simple enough to be solved directly. Features are selected at random for each node from  $K$  number of input variables. Among these, the attribute with most information gain is selected for the root and subsequently the same occurs for each node. After the root is decided, the examples are split into disjoint sets and the corresponding nodes and branches are added to the tree. The simplest splitting criteria for attributes is in the test form of:  $t \leftarrow (A < v)$  where  $v$  is one possible threshold value of attribute  $A$ . The corresponding set  $S_t$  contains all training examples for which  $A$  has values above or below  $v$ . After the dataset is split accordingly to the selected attribute, the procedure moves further down recursively for the remaining dataset. The stopping criteria is when the remaining examples have the same outcome or no further splitting is possible. For the latter, this is due all possible splits have been exhausted or because all remaining splits will have the same outcome for all examples. For all other sets, an interior node is added and associated with the best splitting attribute for the corresponding set as described. Hence, the dataset is successively partitioned into non-overlapping smaller datasets until each set contains only examples of the same outcome (pure node). Ultimately, a pure node can always be found via successive partitions unless the dataset contains examples with identical feature values but different outcome values (contradictions).

All attributes were made available ( $K = 10$ ) each time a node was selected. Decision tree models can fit on any training set that does not contain contradictions. This makes them more prone to overfitting and come in the form of overly complex trees. At the same time, the tree complexity has a crucial effect on its accuracy. Preventing overfitting is usually accomplished by limiting the node depth of the tree and setting a minimum for the number of instances per leaf ( $M = 19$ ). The higher this parameter is set, the more general the tree will be since having many leaves with a low number of instances yields a too granular tree structure. For cell cluster area, the  $depth = 6$ . In the regression case, the outcome probability (mean) is estimated based on a holdout set (backfitting). This holdout set  $N$  was set to 3 parts. One part is held for backfitting and the remaining parts for growing the tree.

#### 2.8.4 Neuron proportion

Modelling neuron cell proportion was achieved with LWL and RandomForest. Locally Weighted learning (LWL) (236,237) is used to select data then pass them to the classifier to construct a model. LWL is a lazy method (instance-based) with memory-based learning. Processing data is deferred and they are stored in memory until needed. The need here is training classifiers and relevant data are found and used to build them. Relevance is measured with Euclidean distance function with nearby points having high relevance. Attribute normalization is turned on by default to deal with different units and scales for distance calculation. Nearest neighbour local models, simply choose the closest point and use its output value. Weighted learning assigns weights using an instance-based method and builds a classifier from the weighted instances. The classifier can be selected and the



number of neighbours used, which determines the kernel bandwidth, and the kernel shape to use for weighting.

The subsets of data used to train each locally weights classifier are determined by a nearest neighbours algorithm. A user-specified parameter  $k$  controls the number of instances used. This is implemented by using a weighting function setting its width to the distance of the  $k$ th nearest neighbour. Let  $D$  be the Euclidean distance to the  $q$ th nearest neighbour  $x_q$   $d = \sqrt{(x - q)D(x - q)}$ . This metric is an important parameter that describes the size and shape of the receptive field. All attributes have been normalised before the distance is computed.  $f$  is a weighting function with  $f(y) = 0$  for all  $y \geq 1$ . Weight  $w_i$  is set for each instance  $x_q$  to:

$$w_i = \exp\left(-\frac{1}{2}(x_i - x_q)^T D(x_i - x_q)\right)$$

Equation 2.6: Weighting function of each instance  $x_i$  before distance computation.  $x_i$  are the training points. Function  $D$  is the distance metric describing the size and shape of the receptive field (diagonal matrix)

Instance  $x_k$  receives weight of 0 so do all instances further away from the test instance, and an identical instance to the test one received weight of 1. LWL was set to a linear nearest neighbour search with Euclidean distance and all neighbours were included in the weighting process.

RandomForest (235) is an ensemble learning technique. They operate by constructing numerous decision trees (explained in section 2.8.3) during training and output the average prediction of individual trees. Random decision trees correct for the individual decision tree's habit of overfitting to their training set. Bootstrap aggregating (bagging) (238) is employed where  $N$  learners are presented with a randomly sampled subset of training points (instances) so that learners will produce different models and their outcome is

averaged. Doing so reduces variance as the conditional probability distribution is averaged  $N$  times. Bagging works best with unstable learners, those that produce differing generalisation patterns with small changes to training data. Therefore, bagging does not work well with linear models. In addition, RandomSubspace method is used (239) to select a uniform number of random samples of features  $n$  to train classifiers from the full set  $N$ . In a situation where discriminative information is spread across the features, will result to reduced correlation between estimators. As a rule of thumb  $n = N/2$ .

Each tree is constructed from a bootstrap sample from the original dataset. Resulting trees are not subject to pruning allowing them to partially overfit their own sample of data. To further diversify the classifiers, at each branch in the tree, the decision of which feature to split on is restricted to a random subset of  $n$  size from the full training feature set. The random subset is chosen anew for each branching point. Breiman (235) suggests  $n$  to be  $\text{int}(\log_2 N_p + 1)$ , where  $N_p$  is the size of the full feature set.

From relevant instances provided by LWL, further selection of instances and 3 features were randomly chosen out of all to construct a classifier. RandomForest iterations was set to ( $-I = 9$ ) for the equivalent number of decision trees. The minimum number of instances that reach a leaf was set to  $-M = 1$  (generalisation term) and the  $-depth = 0$  for unlimited length of trees.

### 2.8.5 Type I astrocyte proportion

Modelling type I astrocyte cell proportion involved RandomSubSpace and IBk. RandomSubSpace method is used (239) to select a uniform number of random samples of features  $n$  to train classifiers from the full set  $N$ . In a situation where discriminative

information is spread across the features, will result to reduced correlation between estimators. As a rule of thumb  $n = N/2$ .

Instance Based k-nearest neighbours (IBk) (240) is a non-parametric method that can be used for regression. The input consists of  $k$  closest training examples in the feature space. The output of IBk is the average of the values of its  $k$  nearest neighbours. Like locally weighted learning (LWL), IBk is a lazy method where the function is only approximated locally and all computation is deferred until prediction. Weights are assigned to the contributions of the neighbours so that the nearer ones contribute more to the average than the distant ones. A common weighting scheme, gives each neighbour a weight of  $1/d$  where  $d$  is the Euclidean distance to the neighbour with continuous variables. For regression, the neighbours are taken from a set of objects of which their property value is known. IBk is sensitive to the local structure of data. The value of each  $k$  nearest point is multiplied by a weight proportional to the inverse of the distance from that point to the test point. The most intuitive of IBk is the 1-nearest neighbour classifier that assigns point  $x$  to the class of its closest neighbour in the feature space presented as  $C_n^{1nn}(x) = Y_{(1)}$ . The nearest neighbour classifier guarantees error rate no worse than twice the minimum achievable error rate given the distribution of the data (Bayes error rate).

RandomSubspace chose 8 features (out of 10) in random then passed them to IBk to construct the classifier. This was repeated 18 times ( $I = 18$ ) and the result of all models was averaged. IBk selected 10-nearest neighbour ( $K = 10$ ) of the target class and all were averaged to provide the outcome. The weighting scheme selected gives each neighbour a weight of  $1/d$  ( $I$ ). IBk minimised the mean squared error (MSE) ( $E$  switch) of residuals. MSE applies more weight for predictions further away from the mean of the  $K$  neighbours.

## 2.8.6 Type II astrocyte proportion

Modelling type II astrocyte cell proportion was achieved with Support vector regression (SVR). SVR is an optimisation algorithm. SVR (241–244) produced models depend only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. In  $\varepsilon$ -SVR, the goal is to find a function  $f(x)$  that has at most  $\varepsilon$  deviation from the actually obtained targets  $y_i$  for all training data and at the same time is as flat as possible. Smaller  $\varepsilon$  values means the closer the function needs to be to  $y_i$ . Points outside the margin are the vectors supporting the actual regression model. Points outside are deemed not important. This characteristic is referred to as sparsity of the solution as only a small set of relevant objects present in the input data are considered to obtain the regression model.

Consider a dataset  $\{(x_1, y_1), \dots, (x_n, y_n)\}$  with  $x \in R^d$  ( $d$ -dimensional input space) and  $y \in R$ . SVR tries to find the function  $f(x)$ , which relates the measured input object (e.g. chemical data) to the desired output property of this object (e.g. cell response). The formula for this is:

$$f(x) = \langle w, x \rangle + b \quad (w, x \in R^d)$$

Equation 2.7: Support vector regression formula.  $w$  and  $b$  are the slope and offset respectively of the regression function.

Both parameters are estimated by minimising the following cost function:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n L_{\varepsilon}(f(x_i), y_i)$$

With:

$$L_{\varepsilon}(f(x_i), y_i) = \begin{cases} 0 & \text{if } |y_i - f(x_i)| \leq \varepsilon \\ |y_i - f(x_i)| - \varepsilon & \text{otherwise} \end{cases}$$

Equation 2.8: Estimating  $w$  and  $b$  parameters for support vector regression.

Where  $\frac{1}{2} \|w\|^2$  is the term characterising the model complexity. This is the flatness of  $f(x)$  and  $L_{\varepsilon}(f(x_i), y_i)$ . The latter being the  $\varepsilon$ -insensitive loss function introduced by Vapnik (243) which does not penalise errors less than  $\varepsilon \geq 0$  (Figure 2.6).  $C$  is the regularisation constant that determines the trade-off between the model complexity  $f(x)$  and the amount up to which deviations larger than  $\varepsilon$  are accepted. Tuning both  $\varepsilon$  and  $C$  should achieve a well-performing model. Literature explaining in greater detail is found in (242,244,245).

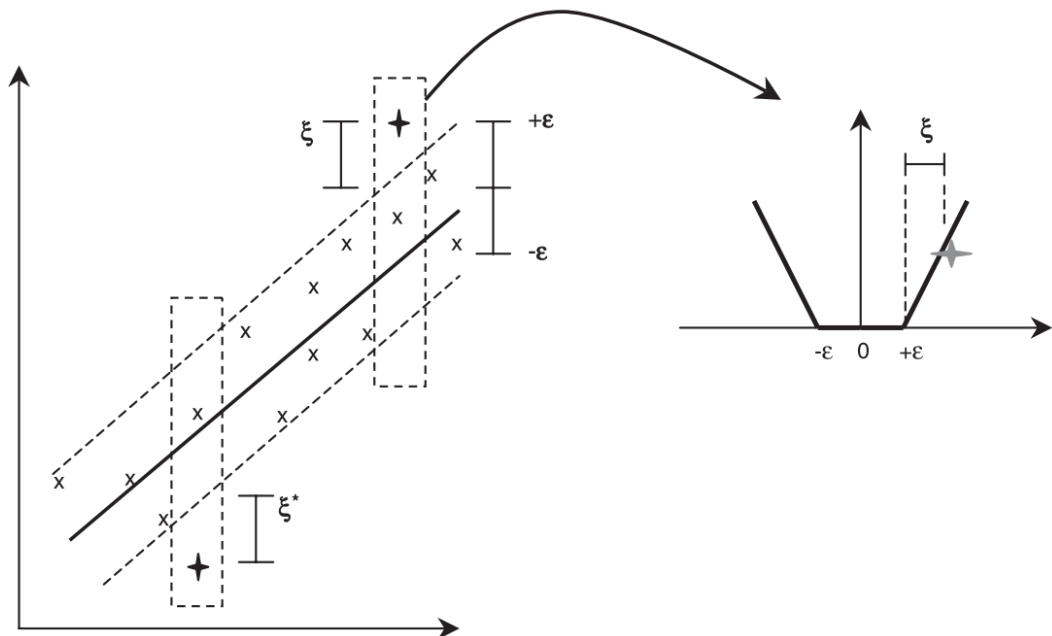


Figure 2.6: Support vector regression example. Left: a line with radius  $\varepsilon$  is fitted to the data. The trade-off between model smoothness (model complexity) and data points lying outside the model is determined by  $C$ .  $\xi$  are the accepted deviations beyond  $\varepsilon$  (243). Data points outside  $\varepsilon$  are called support vectors and denoted as bold + symbols. These support the actual regression model. Right: Vapnik's  $\varepsilon$ -insensitive loss function is shown. The slope is determined by the regularisation constant  $C$  and the support vector position is denoted with a bold grey + symbol. RightsLink license 4081470324618, Elsevier.

Sequential Minimal Optimisation regression (SMOreg) is an algorithm proposed by (246) for solving the mathematical optimisation problem (quadratic programming) arising during training Support Vector Machines (243,245). It works with  $\varepsilon$  insensitive loss function. SMOreg runs iterations to solve optimisation problems like the one described in Equation 2.7. It splits the problem into series of tiny sub-problems (2 Lagrange multipliers) that are solved analytically (approximately) without the explicitly invoking a quadratic optimiser. With SMOreg and the increase of dimensionality, the time required to train SVR models increases linearly.

### 2.8.6.1 Kernel

One of the main reasons support vector machines (SVM) is popular is its ability to model complex linear relationships by using a suitable kernel function. The kernel functions transform the input space into a high dimensional feature space where non-linear relationships can be represented in a linear form. Popular kernels include polynomial, Gaussian or radial basis function (RBF) (Table 2.6), and sigmoid.

Table 2.6: Kernels for use in kernelised models such as Support Vector Regression. Kernel functions transform the input space from low dimensional to high dimensional feature space where non-linear relationships can be described linearly.

Kernel name	Function
Linear (dot product)	$G(x_1, x_2) = x_1' x_2$
Gaussian (RBF)	$G(x_1, x_2) = \exp(-\ x_1 - x_2\ ^2)$
Polynomial	$G(x_1, x_2) = (1 + x_1' x_2)^p$ where $p$ is in the set $\{2, 3, \dots\}$

Peason VII universal kernel (Puk) (247) is the another type of kernel function that can be used in support vector machines. The choice of a kernel function depends on the nature of data, i.e. the kind of relationship that needs to be. The nature of data is usually unknown and the best mapping function must be determined experimentally by apply various kernel functions and the one yielding the highest generalisation performance is selected. Puk

kernel has excellent flexibility because it can be adapted for many kinds of data by adjusting the kernel parameters. It can turn to linear, polynomial, Gaussian, and Sigmoid kernel.

$$f(x) = \frac{H}{[1 + (2(x - x_0)\sqrt{2(\frac{1}{\omega}) - \frac{1}{\sigma}})^2]^\omega}$$

Equation 2.9: Pearson VII universal kernel function (247). This kernel can be adapted to other kinds of kernels such as polynomial, Gaussian, Sigmoid among others by adjusting the parameters  $\omega$  and  $\sigma$ .

Where  $H$  is the peak height at the centre of  $x_0$  of the peak, and  $x$  represents the independent variable. The parameters  $\omega$  and  $\sigma$  control the half (Pearson) width and the tailing factor of the peak.

SMOreg's regularisation constant was set to  $C = 0.52$ . This determines the trade-off between the model complexity and the amount up to which deviations larger than  $\varepsilon$  are accepted. The insensitive loss function was set to  $\varepsilon = 0.001$ . Each attribute was standardised to have zero mean and unit variance with  $x' = \frac{x - \bar{x}}{\sigma}$  (where  $\sigma$  is the standard deviation). Puk kernel's omega parameter was set to  $O = 0.22$  and the sigma was set to  $S = 2.98$ .

### 2.8.7 Proportion of unknown type cells

SMOreg (246) was used to model unknown type cell proportion with the universal Puk kernel both described in the previous section.

SMOreg's parameter that sets trade-off between the model complexity and the deviations larger than  $\varepsilon$  are accepted was set to  $C = 1.12$ . The insensitive loss function was set to  $\varepsilon = 0.001$ . Each attribute was normalised to fall between 0 and 1 have zero mean

and unit variance with  $z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$ . Puk kernel's omega parameter was set to  $\omega = 0.91$  and the sigma was set to  $\sigma = 0.19$ .

### 2.8.8 Neurite length

Modelling neurite length was achieved with RandomCommittee and RandomTree. RandomCommittee (248) builds an ensemble of randomisable base classifiers. Each base classifier is built using a different random number seed on the same training data. The final prediction is the average of all outcomes from each base classifier. RandomTree (235) is explained in section 2.8.3. A decision tree is built where the paths from the root to the leaf represent classification rules. The nodes represent a logic test on a particular value of an attribute.

RandomCommittee was set to iterate 32 times ( $\text{--- } I = 32$ ) for the equivalent number of RandomTrees. RandomTree was configured to use all features available ( $\text{--- } K = 10$ ). The minimum number of instances reaching a leaf (weight) was set to ( $\text{--- } M = 7$ ) and the maximum depth of the trees to unlimited ( $\text{--- } \textit{depth} = 0$ ). The outcome probability (mean) is estimated on a holdout set (backfitting) set to 5 parts ( $\text{--- } N = 5$ ). One part for backfitting and the remaining for growing the tree. Some classifiers may be unable to provide an outcome. This is referred to unclassified instances. This was allowed with the ( $\text{--- } U$ ) switch.

### 2.8.9 Type I astrocyte area

Modelling Type I astrocyte area was achieved with M5Rules. M5Rules (249–251) is a model tree technique that deals with continuous class problems. They have a typical decision tree structure (e.g. RandomTree) but use linear functions at the leaves (outcome). M5 builds a tree by splitting the data based on the values of predictive attributes. The algorithm



chooses attributes that minimise intra-subset variation in the class values of instances that go down each branch. The algorithm starts with a tree learner applied to full training dataset. Next, the best leaf according to a heuristic is turned to a rule and the tree is discarded along with all instances covered by the rule. This process occurs recursively to all remaining data and stops when all data are covered by at least one rule. This is known as the separate-and-conquer (SAC) strategy for learning rules. The trees build at each stage is a partial one and this leads to computational efficiency without affecting size and accuracy of resulting rules.

For the initial tree, the splitting criterion is based on the standard deviation of the class values that reach a node as an error measure for that node. The expected reduction in error is calculated by testing each attribute at that node. The attribute that maximises the expected error reduction is selected. The standard deviation reduction is calculated by:

$$SD_r = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

Equation 2.10: Building initial M5Rules trees. Standard deviation reduction formula to select attributes that minimise expected error.

Where  $T$  is the set of examples that reach the node and  $T_1, T_2, \dots$  are the sets that result from splitting the node according to the attribute chosen. Splitting the tree halts when the class values of all instances that reach the node vary very slightly or when very few instances remain.

The 'best' leaf selection heuristic is coverage and the percent root mean squared error. The former is the number of instances the rule applies for and the equation for the latter is:

$$\% RMSE = \frac{\sqrt{\sum_{i=1}^{N_r} (A_i - p_i)^2 / N_r}}{\sqrt{\sum_{i=1}^N (A_i - \bar{A})^2 / N}}$$

Equation 2.11:  $A_i$  is the real values for example  $i$ ,  $p_i$  is the prediction by the linear model at the leaf level,  $N_r$  is the number of examples covered by leaf,  $\bar{A}$  is the mean of real values, and  $N$  is the total number of examples.

The percent mean squared error favours accuracy at the expense of coverage. Other measures may trade-off accuracy against coverage. In the literature, there is no consensus on the best measure for rule value as no method proposed so far resolves this problem satisfactorily. An extensive theoretical study is in (252).

The expected error of each node is calculated by averaging the absolute difference between the predicted value and actual class value of each instance reaching the node. These optimistic errors on training data are compensated by multiplying with a factor. This factor takes into account the number of parameters in the model representing the class value at the node and the number of training examples that reach it. Constructing trees can lead to sharp discontinuities between adjacent linear models at the leaves. These differences are compensated with a procedure called smoothing. The procedure computes a prediction using the leaf model then passes that value back to the root. On its way there, the value is smoothed at each node by combining it with the value predicted by the linear model for that node that was produced at the time the tree was built. Past experiments have shown that smoothing substantially increases the accuracy of predictions (250,253).

$$p' = \frac{np + kq}{n + k}$$

Equation 2.12: Model tree smoothing procedure to reduce sharp discontinuities inevitably occurring between adjacent linear models at the leaves.  $p'$  is the prediction passed up to the node higher,  $p$  is the prediction passed to this node from below.  $q$  is the value predicted by the model at this node,  $n$  is the number of training instances that reach the node below, and  $k$  is a smoothing constant.

The  $-N$  switch disables pruning (simplifying) the trees generated and the parameter determining the minimum number of instances to create a leaf node was set to  $-M = 2$ .

### 2.8.10 Astrocyte fibre length

Modelling astrocyte fibre length was achieved with AdditiveRegression and Decision Stump. AdditiveRegression (254) is a stochastic gradient boosting method that enhances the performance of 'base' classifiers. It is a method to increase model complexity and improve its fit by combining models learnt from base learners. AdditiveRegression starts with a simple predictor such as the mean. Subsequent models from each iteration builds the model stage-wise on a subsample of data, drawn at random (without replacement) to reduce computation time and add randomness. Randomness reduces the chances of overfitting. The residuals left from the previous iteration are modelled again. Overall prediction is given by the sum of the collection.

Gradient boosting is usually employed in conjunction with the base learners such as decision trees. Each base learners quality of fit is improved using Friedman's modified gradient boosting method (254,255). Consider a function estimate problem with  $x$  inputs  $\{x_1, \dots, x_n\}$  and  $y$  outputs  $\{y_1, \dots, y_n\}$ . Gradient boosting at the  $m$ th step would fit a decision tree  $h_m(x)$  to the pseudo-residuals. These are the gradient of the loss function being minimised. Let  $J_m$  be the number of the base learner's leaves. The tree partitions the input space into  $J_m$  disjoint regions  $R_{1m}, \dots, R_{J_m m}$  and predict a constant value in each region. The output of  $h_m(x)$  for input  $x$  can be written as the sum:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} I(x \in R_{jm})$$

Equation 2.13: Decision tree fit to the residuals of the gradient of the loss function being minimised.  $b_{jm}$  is the predicted value in  $R_{jm}$  region.  $I$  is a function defined on a set  $x$  that indicates membership of an element in a subset  $R_{jm}$ .

The  $I$  function returning with a value for a set of  $x$  of 1 indicates being a member of  $R_{jm}$  and 0 if not. The optimal value  $\gamma_m$  is chosen separately for each of the tree's regions using line search. This method is called TreeBoost (254) and is a basic iterative approach to find a local minimum to optimise the loss function,  $L$ . Regularisation is the term used for training too closely to the dataset and this leads to degradation of the model's generalisation ability (overfitting). Shrinkage is a method for regularisation that modifies the update rule as:

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m h_m(x), \quad 0 < \nu \leq 1,$$

$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

Equation 2.14: AdditiveRegression gradient boosting with shrinkage as the regularisation method.  $J_m$  is the number of leaves for the base learner.  $\gamma_{jm}$  is a value chosen with line search that minimises the value of the loss function  $L$ .

Where the  $\nu$  parameter is called the learning rate. This parameter was found that small values ( $\nu < 0.1$ ) gives fantastic improvement in the model's generalisation ability over gradient boosting without shrinking ( $\nu = 1$ ) (256). It also comes with a price – an increased computational time both training and querying as lower learning rate performs more iterations and combines more models.

A base learner has been used called DecisionStump (257,258) (DS). Recently it was used as a tree for classifying cancer gene expression data (259). DS is also known as 1-rules because it consists of a one-level decision tree. It is a decision tree with one internal node (the root) where this is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. For continuous

features, usually, a threshold feature value is selected, and the stump contains two leaves. One for values that fall below and another for values that fall above. DecisionStump performs surprisingly well on some commonly used benchmark datasets (UCI repository, (257) demonstrating that learners with high bias and low variance may perform well as they are less prone to overfitting. In machine learning, decision stumps are often used as components ('base learners') for ensemble techniques such as boosting and bagging.

Each iteration in AdditiveRegression fits a model to residuals left by the classifier from the previous iteration. This parameter —  $I$  was set to 2. DecisionStump does not have any parameters to configure.

## 2.9 COMPUTATIONAL CELL CULTURE EXPERIMENTS

A program coined 'Get-Chem' was created in php v7 (x64) that automates the process of performing cell culture experiments computationally. The program can be used for optimisation problems. Chemistries can be screened in minutes to determine cell performance. This tool will be used to discover chemistries better than our synthetic standard (amine). In addition, manipulating the inputs in conjunction with the use of the models can shed light to the chemical input effect on cell performance. Get-Chem takes in chemical variables and their possible values then recursively combines them to create numerical chemistries (test cases). Predictive models are then called to produce estimates of cell performance in pseudo-MIMO (multiple inputs, multiple outputs) fashion. Finally, results are stored and sorted in an SQL database. The flowchart of Get-Chem is as follows:

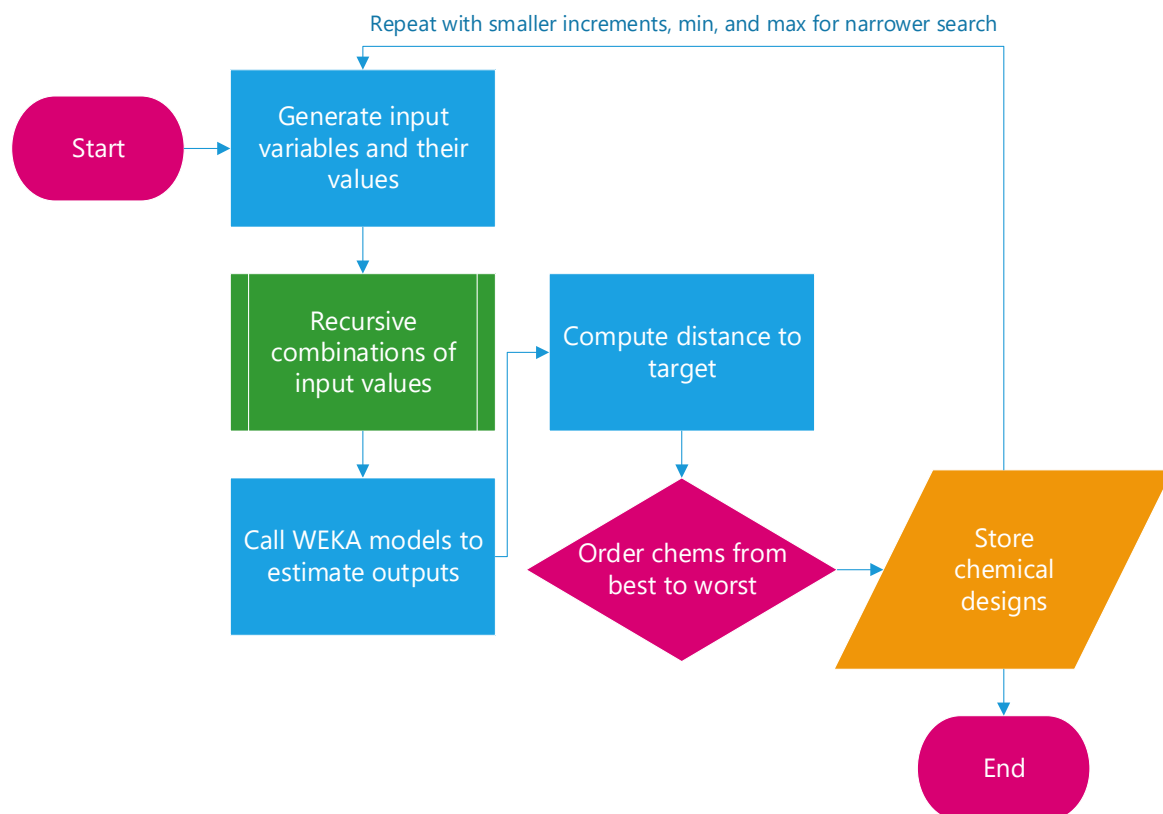


Figure 2.7: Get-chem flowchart. This program automates the process of running computational cell culture experiments. The user decides the test cases (chemistries) and cell performance estimates are computed for each. Results are stored in an SQL database.

### 2.9.1 Generating test cases

The first step was to generate test cases (numerical chemistries). The idea is to list all inputs as variables and the possible values they can take. Input variables were decided beforehand and the min/max and increments between these are editable in the configuration file of the software. The min and max values were discovered by extreme theoretical designs (e.g. very electronegative) drawn in ACD/ChemSketch 2016. The increment values determine the search width relative to the min/max set previously. Large increments indicate a broad search and lower values indicate the opposite, narrow search. Then, recursive combinations of these was performed shown below:

```
private function recursive_combinations($arrays, $i = 0) {
...
    if ($i == count($arrays) - 1) {
        return $arrays[$i];
    }

    // get combinations from subsequent arrays
    $tmp = $this->recursive_combinations($arrays, $i + 1);

    $result = array();

    // concat each array from tmp with each element from $arrays[$i]
    foreach ($arrays[$i] as $v) {
        foreach ($tmp as $t) {
            $result[] = is_array($t) ?
                array_merge(array($v), $t) :
                array($v, $t);
        }
    }

    return $result;
}

public function generate_combinations() {
...
    $this->reset_timer('Generating combinations from all input values');

    $combinationsInput = [];
    foreach($this->input_values as $valueTypeArray) {
        $combinationsInput[] = $valueTypeArray;
    }

    $this->combinations_array = $this->recursive_combinations($combinationsInput);
    $this->combinations_count = count($this->combinations_array);
...
    return true;
}
```

Code snippet 1: Generating test cases as input vectors. The software takes in input variables with possible values to them. Each variable is treated as an array (table) and each arrays are recursively combined to generate test cases.

### 2.9.2 Ranking method

The ranking system is a distance metric for all predicted cell outputs of a chemistry compared to the real cell outputs of our target, laminin. The Bray and Curtis statistic is used to quantify compositional dissimilarity between two objects, based on counts for each (260,261). This metric is normally used in ecology and biology. The dissimilarity index is calculated by taking the summed differences between the variables and standardising

them with the summed variables of the objects (Equation 2.15). The ranking metric used discriminates against smaller values therefore large values in cell outputs have more influence in *in vitro* cell performance.

$$d^{BCD}(i, j) = \frac{\sum_{k=0}^{n-1} |y_{i,k} - y_{j,k}|}{\sum_{k=0}^{n-1} |y_{i,k} + y_{j,k}|}$$

Equation 2.15: Bray and Curtis dissimilarity index (261) between objects  $i$  and  $j$ .  $k$  is the index of a variable and  $n$  is the total number of variable  $y$ .

### 2.9.2.1 Weighting cell performance indicators

Weights for cell outputs are necessary, as some of the cell performance indicators are more important than others for the purposes of this study. The most important cell output is monolayer formation *in vitro* as this is an indication of cells maximising interaction with their environment (59). Cell proportion of the cell types investigated is second in order as cells perform different functions. A tissue with deviation from the natural cell proportion may have undesirable levels of biological function. In addition, undifferentiated cells cannot be transplanted into patients (175). Neurons are the functional component of the nervous system. These electrically excitable cells process and transmit information through electrical and chemical signals occurring via synapses. Once matured to neuron axons, these connect to other cells such as neurons, muscles and glands for information transmission (262). From macroglial cells, astrocytes were chosen as they are the most abundant type in the central nervous system (263) and have numerous projections that link neurons to their blood supply. Due to astrocyte abundance, astrocyte fibres have been set with lower importance. *In vitro*, type I astrocyte spreading indicates they are under stress. *In vivo*, this is called reactive astrogliosis and it could arise due to injury to the nervous system (194). Type I astrocyte spreading area has medium importance. Cell types not investigated include oligodendrocytes, ependymal cells and radial glia (136). The table below shows the importance and weights applied to cell outputs:



Table 2.7: Cell output weighting for importance in finding a better synthetic environment to culture neural stem cells. This weighting is applied after the dissimilarity index is calculated. Higher values have more impact on the rank of cell performance associated with numerical chemistries.

<b>Cell variable</b>	<b>Weight</b>	<b>Importance</b>
Cell cluster area	0.1	Highest
Neuron proportion	40	High
Proportion of unknown type cells	50	
Neurite length	0.5	Medium-High
Type II astrocyte proportion	10	
Type I astrocyte proportion	0.2	Low
Type I astrocyte area	0.1	
Astrocyte fibre length	0.1	

### 2.9.3 Storing results

Results were saved in a MySQL database (v5.6) and the user can export all or some results for further analysis such as sensitivity. The order of these, ascending or descending, is determined by their ranking metric value. The code that performs this action is as below:

```

private function db_export($type, $benchmark = false) {
    $order = [
        'closest' => 'ASC',
        'furthest' => 'DESC',
    ];
    echo "\n\033[33mNumber of $type results to be exported: \033[0m";

    while($input = fgets(STDIN)){
        if (intval($input)) {
            $selectFields = "Name, CONCAT(".$this->config['db_table'].", id) as ID, ".implode(', ',
array_merge($this->config['input_variables'], $this->config['output_variables'])).", ".$this-
>comparison['code'];
            if($benchmark) {
                $sql = "(SELECT $selectFields FROM " . $this->config['db_table'] . " WHERE Name = "
ORDER BY ".$this->comparison['code']." ".$order[$type]." LIMIT " . intval($input) . ")";
                $sql .= " UNION ALL ";
                $sql .= "(SELECT $selectFields FROM " . $this->config['db_table'] . " WHERE Name <> ")";
                $sql .= " ORDER BY IF(Name <> ", 0, 1) ASC, ".$this->comparison['code']." ".
$order[$type] . " .";
            } else {
                $sql = "SELECT $selectFields FROM ".$this->config['db_table']." ORDER BY ".$this-
>comparison['code']." ".$order[$type]." LIMIT ".intval($input).";";
            }
            break;
        } else {
            $fileName = $this->config['exports_path'].$this->db_table.'-'.$type.intval($input).'.'.$this-
>get_total_time().'.csv';
            $this->reset_timer("Fetching ".intval($input)." $type results from DB");
            $data = "";
            $result = $this->db->query($sql);
            $this->reset_timer("Exporting data to ".$fileName);
            $finfo = $result->fetch_fields();

            foreach ($finfo as $field) {
                $data .= $field->name.';';    }
            $data = rtrim($data, ',');
            $fp = fopen($fileName, "w");
            fwrite($fp, $data);

            while ($row = $result->fetch_row()) {
                $data = "\n";
                foreach($row as $col) {
                    $data .= $col.';';
                }
                $data = rtrim($data, ',')."\n";
                fwrite($fp, $data);
            }
            fclose($fp);
        }
    }
}

```

Code snippet 2: Computational cell culture experiment results. This code shows how data are saved in a database then exported by the user choosing the number and the order of results, based on the ranking metric.

#### 2.9.4 Numerical chemistry conversion

The results contained theoretical (numerical) chemistries and not all of them could be synthesised or are stable. Numerical chemistries needed to be converted from theoretical ones to ones that can be created in practice. Designs were re-drawn in ACD/ChemSketch 2016. Using the pKa and logP values for each, the head group and possible side chains of the molecule was first drawn. From that, one atom at a time was added to the backbone and possible side branches if necessary. The choice of atoms at each level was directed from the chemical values of results for each variable. Finally, the molecular mass and volume was calculated for the re-constructed chemistries and they were shortlisted only if they matched with those of the theoretical chemistries.

#### 2.9.5 Reassessing converted chemistries

The re-constructed chemistries have slightly different values than the theoretical ones. Reassessment was necessary as another step in the process to validate findings. A separate test was conducted from the chemistries that made it in the shortlist. The re-constructed chemistries were fed into the same predictive models used previously and the cell outputs with their distance to laminin's was calculated. The next step was to look for the chemistries as an off-the-shelf product preferably in the form of self-assembly molecules. The similarity search was conducted in [e-molecules](#) and [ChemSpider](#) with different labile groups and without. At this point, some chemistries could not be found and inquiries were sent to laboratories to synthesise them.

## 3 DISCOVERING RELATIONSHIPS COMPUTATIONALLY

---

### 3.1 INTRODUCTION

Finding synthetic cell culture environments where cells perform similarly to the *in vitro* biological control is possible by testing a large number of chemistries. This entails cell culture experiments and they come with limitations such as:

- High costs e.g. materials, reagents, cells
- Time required e.g. 6 months for 13 environments
- Personnel to obtain results faster
- Large number of experiments due to large number of possible environments to test
- Animals are still required as stem cell source is rat foetal neural stem cells

Another methodology mitigating these limitations is to move to model cell culture experiments on a machine. We coined this methodology “computationally informed surface engineering”.

The study of data may shed light in the direction of the computational tools to proceed and describe the relationships (if any) between chemical and cell parameters. These relationships in the form of computational models will allow testing of millions of environments with cells in minutes. We hope to find better candidate synthetic environments that will allow us to develop cell therapies *in vitro* as well as to be able to understand the effect of the chemistry of the synthetic environment on cell performance.

This chapter has two aims:

1. Exploring previous data (41) and testing parametric assumptions
2. Searching for relationships with correlation tests

The objectives are to:

1. Investigate surface chemistry and cell performance data from synthetic environments where nervous tissue was developed from neural stem cells
2. Compare cell performance with that of the biological control environment
3. Find relationships between chemical surfaces and cell parameters
4. Discover insights relating to the effect of chemical properties of the environment on cell responses

### 3.1.1 Previous work data

The data for this chapter were acquired from previous work (41). The overlap with this project is to improve control of neural cell responses through chemically defined microenvironments. The aim is to assess the response of neural stem cells and progenitors expanded as spheroids of proliferating cells (neurospheres) in a range of surface chemistries (functionalities). Wright *et al.* (2014) used rat ventral mesencephalon derived cells from 12-day old embryos (E12) of rats (41). We discovered that terminal surface chemistry directs fractional populations of neurons and astrocytes (264). The authors used self-assembly molecules to modify the presenting chemistry of solid surfaces. The functional groups include amine (NH<sub>2</sub>), hydroxyl (OH), carboxyl (COOH), methyl (CH<sub>3</sub>), phenyl (Ph) and thiol (SH). A list of these chemistries is in Table 2.1. Neurospheres spread and cells attached and populated surfaces differently in each environment. This will be discussed in detail in the Results section (3.2).

The chemistry of culture surfaces was verified with surface chemistry characterisation techniques such as contact angle measurements, infrared, Raman, and X-ray photoelectron spectroscopy. Data describing the properties of surface chemistries used in experiments

have been retrieved from the literature or from peer reviewed computational models. Such data define unique chemical designs. The adopted chemical parameters investigated in biomaterial sciences are in the table below. These are explained in 2.7.1.

Table 3.1: Chemical parameter and value origin.

<b>Chemical parameter</b>	<b>Values acquired from</b>	<b>Protein specific methods</b>
Partition coefficients (logP)	ACD/ChemSketch 2016 (106,209)	Ghose & Crippen's method (211,212)
Acidity measures (pKa dissociation constant)	(218–221)	ProPKA v3 (206–208,222)
Molecular mass	Calculated	(223)
Molecular volume	(224–226), ChemDraw 2015	ProteinVolume v1.3 (227)

Cells from the neurospheres differentiated to neurons and astrocytes, migrated and elongation of cell processes was either promoted or retarded. All cell response comparisons were against cell performance of the biological environment (glass coated with biological material, PDL and laminin). The author demonstrated that the presentation of chemical cues provide a path towards improving the robustness of *in vitro* neural culture environments controlling multiple cell responses attributed to surface-cell interactions.

Cell performance was characterised from images of cells cultured in the environment of interest. The cells were stained with cell-type-specific dyes to characterise the types, morphology and processes and then images are captured through microscopy. Morphological cell performance attributes were selected from neuro-regeneration literature (41,265,266). These attributes for monolayer cultures included cell cluster size, cell density of neurons and astrocytes, and projection length of axons and astrocyte fibres.

The data from previous work (41) were investigated to provide insights as to the direction and computational tools to choose for establishing modelling relationships between

chemistry and cell performance. For this task, parametric tests are preferred as they provide greater statistical power and can handle heterogeneous variance compared to nonparametric tests (267). Parametric assumptions are tested on cell data for variance and distribution. Correlation tests with significance follow between cell and chemical data in the search for relationships, their strength, and direction. Plots of cell data against chemical data accompany the relationships discovered. All findings are discussed then the chapter ends with conclusions.

## 3.2 RESULTS

The cell culture surfaces prepared with synthetic chemistries consist methyl (-CH<sub>3</sub>), carboxyl (-CO<sub>2</sub>H), amine (-NH<sub>2</sub>), hydroxyl (-OH), phenyl (-Ph), thiol (-Sh) and the *in vitro* biological standard made of laminin on top of poly-d-lysine (P/LAM) (41). The chemical parameters of these environments that were retrieved from the literature consist of the partition coefficient (logP), acidity measure (pKa), molecular mass and molecular volume. Cell performance consists of scores of cell cluster area, cell densities and ratios of neurons and astroglia, and cell projection length of neuron axons and astrocyte fibres. The source of the cells is E12 ventral mesencephalon and the cell performance was measured on days 3, 5 and 7 in culture on modified surfaces.

### 3.2.1 Variance tests

To use parametric tests certain assumptions regarding the data used need to be tested. One assumption is homogeneity of variance between different groups. This means the variance between cell scores from each environment needs to be almost equal (195) to meet this assumption. Parametric tests can perform well even with heterogeneous

variance (267), however the variance may be useful to the computational methodology proposed here. The variance of same time-point cell data (e.g. day 3) from different environments was compared. Levene’s test performs a one-way ANOVA on the differences between each score and the mean of the group whereas, the Brown-Forsyth test uses the median (195). The tests were set to a significance level of 95%. This means the chances of accepting a false positive are less than 5% ( $p \leq 0.05$ ). Here, a false positive is identical variance between samples where in actuality, that is not true. Below the variance test results where  $p \geq 0.05$  mean that the homogeneity of variance assumption holds and  $p \leq 0.05$  means the assumption is violated:

Table 3.2: Variance tests ( $F - values$ ) on previous data (41).  $F - values$  are reported with 2 degrees of freedom parameters in brackets ( $df1, df2$ ). Levene’s test was performed on absolute deviations and uses the mean whereas, the Brown-Forsyth test uses the median. Probability values that are  $p \leq 0.5$  reject the null hypothesis meaning the assumption of homogeneity of variance is violated.

	<b>Cell Cluster Area</b>	<b>Neuron Density</b>	<b>Astrocyte Density</b>
<b>Levene’s</b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.00	0.00	0.00
Day 7	0.00	0.00	0.00
<b>Brown-Forsyth</b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.01
Day 5	0.00	0.00	0.00
Day 7	0.00	0.00	0.00

	<b>Neuron/Astrocyte Ra.</b>	<b>Neuron Axon Length</b>	<b>Astrocyte Fibre Length</b>
<b>Levene’s</b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.27	0.00	0.00
Day 7	0.15	0.00	0.00
<b>Brown-Forsyth</b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.32	0.00	0.00
Day 7	0.20	0.00	0.00



As seen in Table 3.2, only Neuron-Astrocyte Ratio on day 5 and day 7 have homogeneous variance ( $p > 0.05$ ). This ratio is  $\frac{\text{Neuron Density}}{\text{Astrocyte Density}}$ . The variances for the remaining cell variables are heterogeneous since the computed probability for each is  $p \leq 0.05$ . In time point day 3 in the early stages of cell development a few examples of homogeneous variance were expected since the neurosphere seeding conditions are identical. It seems the chemical properties such as wettability, lipophilicity affect biological interactions as soon as the seeding of the cell solution takes place. On day 5 and 7, heterogeneous variance was expected as cell performance can vary vastly in different environments (41,265).

### 3.2.2 Distribution

Another parametric assumption is a normal distribution. A normal distribution has most of the data points fall in the middle of the range (cluster around the mean) with the number tapering off symmetrically either side of the mean to a few extreme values in each of the two tails. While normality is not the only assumption for using parametric tests, the arithmetic of such tests is based on the parameters describing a symmetrical, bell-shaped curve (Gaussian) (196). Normality tests can determine whether sample data has been drawn from a distribution that is approximately normally distributed (195).

Popular methods for visual distribution inspection are quantile-quantile (q-q) (199) and box plots. Q-Q graphs are actual data plotted against a normally distributed version of given data. The original data are arranged in ascending order in percentiles (quantiles) and the normally distributed data points are obtained from the z-scores of the original data. Z-score is a measure of how many standard deviations below or above the population mean a raw data point is. Box-plots are non-parametric and make no assumptions regarding underlying data distribution. Popular statistical methods for normal distribution are Shapiro–Wilk and

Anderson-Darling (198–200). The former is more sensitive around the centre of the distribution whereas the latter at the tails. The figures below show q-q plots:

### 3.2.2.1 Quantile-quantile (QQ) plots

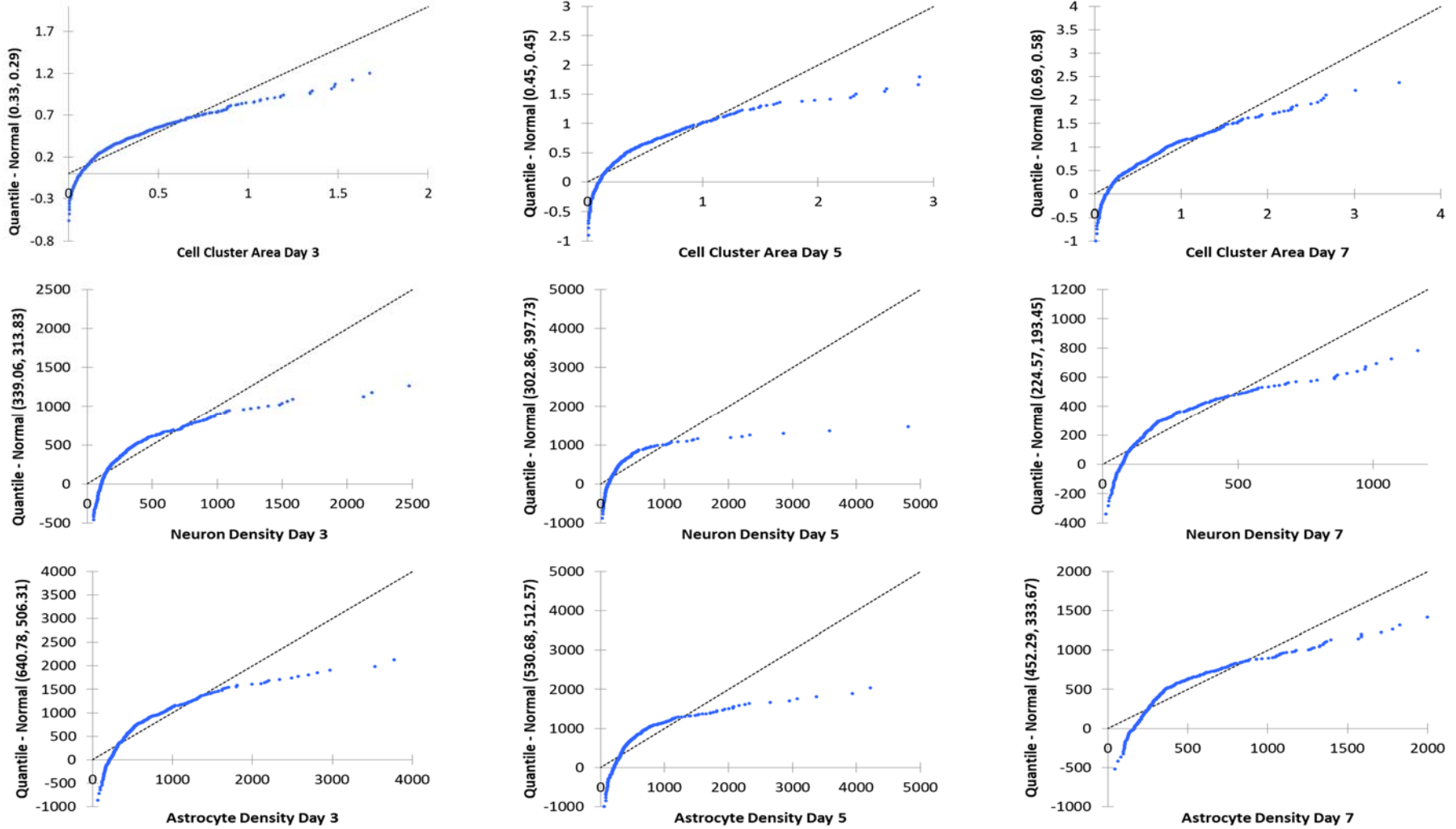


Figure 3.1: QQ probability plots of Cell cluster area and Cell density variables. This is a graphical method comparing two probability distributions by plotting their quantiles against each other.

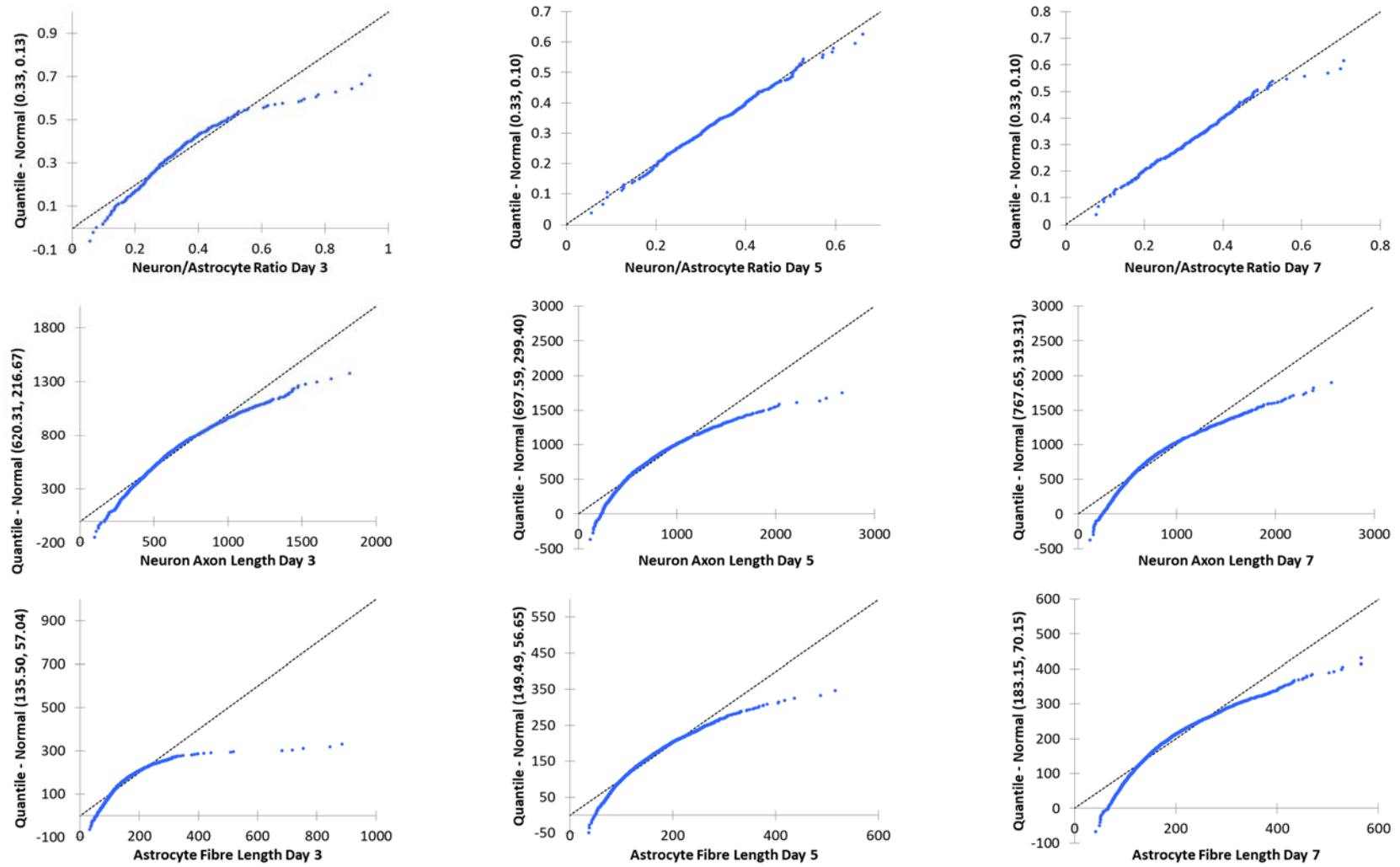


Figure 3.2: Quantile-Quantile probability plots of neuron/astrocyte ratio and cell process variables.

Except for neuron/astrocyte ratios, remaining data have a distribution closer to gamma distribution. The image below (left) shows QQ plots where a gamma distribution is compared to a normal distribution which is very similar to what is seen in the graphs above.

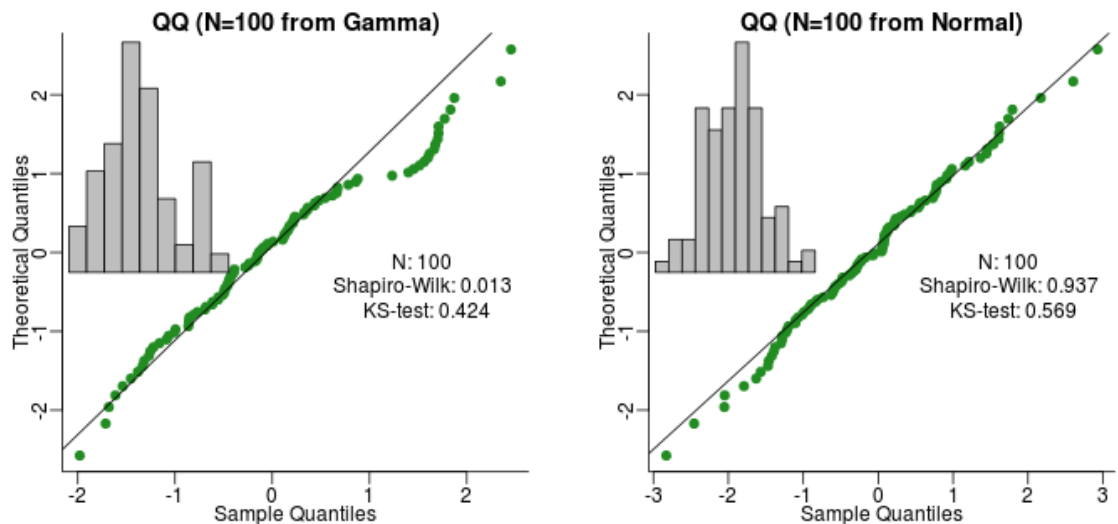


Figure 3.3: Left: normal QQ plot fitted on gamma distribution ( $k = 0.5, \theta = 1$ ). Right: normal QQ plot fitted on normal distribution.

This is not uncommon in life sciences (268) and most likely the sampling method is not at fault despite the presence of possible outliers. Most of the morphological data quantified from cell images are from synthetic environments. In most of these, cells do not behave as they do in the biological control (laminin) where a distribution closer to normal is expected (41). In synthetic environments, smaller values appear more frequently than larger values giving right skewed distributions. In addition, this was observed and discussed in detail in previous work (41). An example of an approximately normal distribution is shown in neuron/astrocyte ratio day 5 and 7. Below are then box-plots showing the data spread and distribution:

### 3.2.2.2 Box-plots

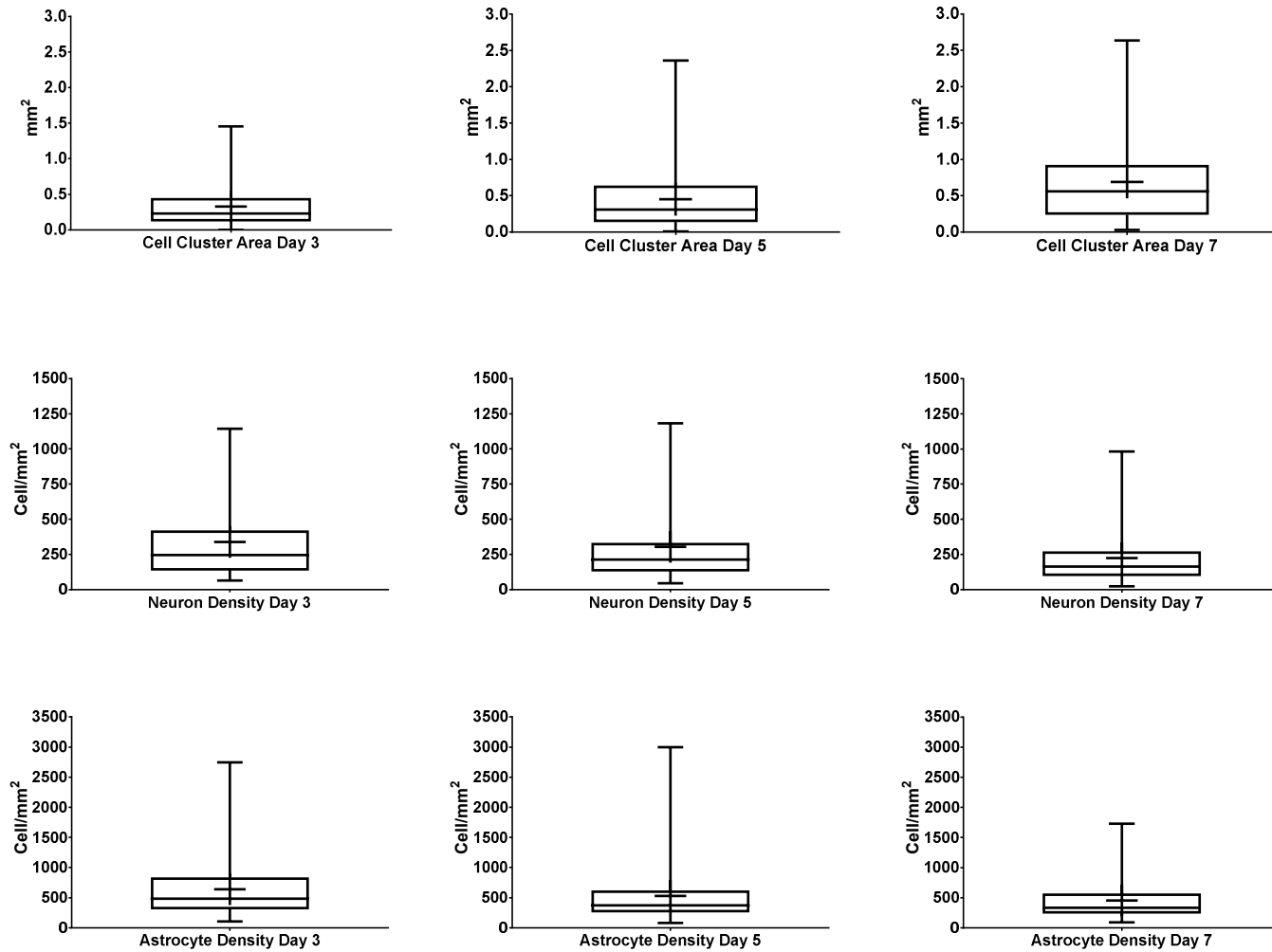


Figure 3.4: Box-plots of cell cluster area and cell density variables. These show cell variables and their quartiles. Top and bottom whiskers denote the upper and lower quartiles. Box spacing indicate data spread and skewness. + indicates the average and the black line inside the box is the median (2nd quartile). Displaying 1-99 percentile of data.

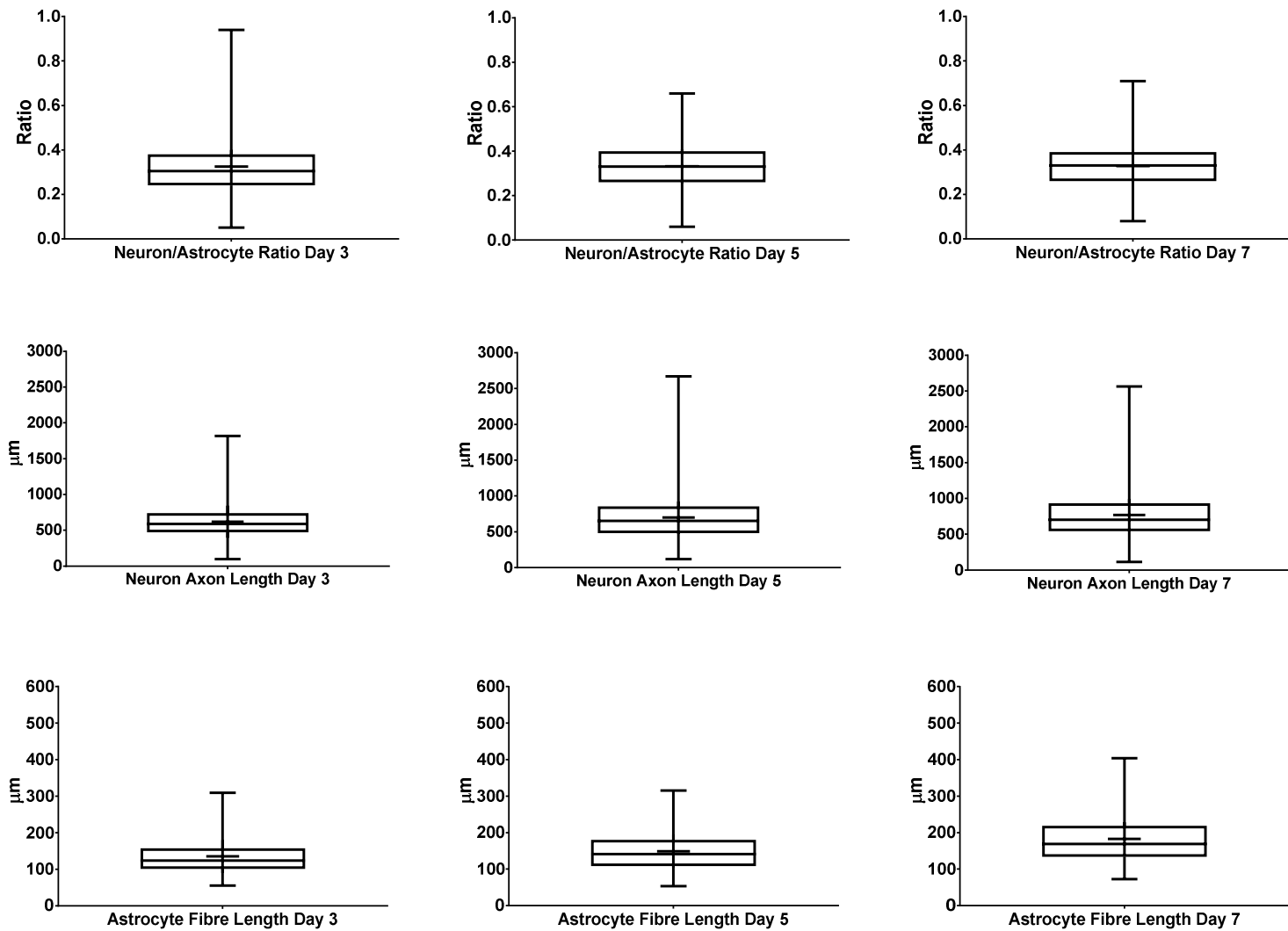


Figure 3.5: Box-plots of neuron/astrocyte ratio and cell process variables. These show cell variables and their quartiles. Top and bottom whiskers denote the upper and lower quartiles. Box spacing indicate data spread and skewness. + indicates the average and the black line inside the box is the median (2<sup>nd</sup> quartile). Displaying 1-99 percentile of data.

The box-plots above show similarities with data distribution findings. Most are gamma distributions except for neuron/astrocyte ratios. In a normal distribution, the box is expected around the centre of the whiskers and the mean indicated with a +, should be overlapping or very close to the median indicated with a line in the box. Cell Cluster Area is the area where cells migrate away from the sphere they came in. This measure increases with time as expected from developing and migrating cells.

Neurons, the functional component of the nervous system, tend to appear early in the culture but later diminish in density. This could mean the neuron population decreases over time, that they migrate or, that they are carried away by the rapidly proliferating astrocyte “carpet”. The latter is more likely as a similar effect is observed with the supporting cells called astrocytes - astrocyte density decreases with time as well. In both “vanishing” neuron and astrocyte densities, variation decreases in day 7 ascertaining the effect of cell migration. Alone, cell density per  $\text{mm}^2$  cannot answer whether cell populations die or migrate further apart. For this reason, the neuron/astrocyte ratio (NAR) has to be used in conjunction to answer the question for the non-proliferating neuron population. Similar NAR in different time points means neurons migrated. Reduction in NAR in time points means neurons may have died. From the results above, the average NAR value does not change with time meaning neurons density decreases mostly due to cell migration.

In Figure 3.5, neuron axons explore the environment searching for other neurons to form neural networks (circuits). Their length increases with time, as does the variation. On time point day 3, neurons have not had enough time to differentiate and send out processes but later on day 5 and 7 they elongate. *In vitro*, astrocyte fibres are mainly used to envelop synapses made by neurons, so they too are exploring for other cells. As with neuron axons,



astrocyte fibre length increases with time and the variation increases on day 7. For both cell types, projection elongation is expected to reach distant migrating cells.

### 3.2.2.3 Normality tests

Testing for normal distributions inform statistical whether a distribution is normal or not through the test statistic and probability values ( $p$ ) of false positive risk ( $\alpha$ ). Shapiro-Wilk test detects non-normality around the centre of the distribution whereas, the Anderson-Darling is better suited for the tails (198). Below are the normality results for each cell parameter with the test  $p$  value:

Table 3.3: Normality tests. Approximate normal distribution are those with  $p \geq 0.05$ .

	<b>Cell Cluster Area</b>	<b>Neuron Density</b>	<b>Astrocyte Density</b>
<b><i>Shapiro–Wilk (W)</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.00	0.00	0.00
Day 7	0.00	0.00	0.00
<b><i>Anderson—Darling (A<sup>2</sup>)</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.00	0.00	0.00
Day 7	0.00	0.00	0.00
	<b>Neuron/Astrocyte Ra.</b>	<b>Neuron Axon Len.</b>	<b>Astrocyte Fibre Len.</b>
<b><i>Shapiro–Wilk (W)</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.53	0.00	0.00
Day 7	0.00	0.00	0.00
<b><i>Anderson—Darling (A<sup>2</sup>)</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>	<b><i>p-value</i></b>
Day 3	0.00	0.00	0.00
Day 5	0.5	0.00	0.00
Day 7	0.28	0.00	0.00

Both normality tests (Shapiro-Wilk, Anderson-Darling) are statistical inference drawing conclusions from sample data by emphasising the frequency or proportion of data. This inference framework is well established and is the basis for hypothesis testing and confidence intervals. From the table above (Table 3.3), only neuron/astrocyte ratio day 5 and day 7 are normally distributed therefore their averages will serve well as their central tendency. From the q-q, box-plots and normality tests (Figure 3.4-Table 3.3), the remaining variables have closer to gamma distribution. For these, the median will be used instead as the central tendency of each. The central tendencies of cell measurements are required for each chemistry to perform correlation testing of two variables (bivariate). The sample size needs be equal to the number of environments.

### 3.2.3 Correlation and visual relationships

Correlation shows the statistical relationship between variables. These relationships assume dependence and linearity. Pearson's correlation has an advantage over using untransformed data to find correlation between variables but is also sensitive to outliers (203). To alleviate this problem, the median is used as the central tendency (196,269) for all cell variables except neuron/astrocyte ratio day 5 and 7. Since these two have approximately normal distributions, the average is used instead.

Correlation tests are performed on chemical and cell data where measurements were taken on 3 time points in culture (day 3, 5 and 7). Correlation significance and frequency follow next. After that follows a correlation heat map for all chemical vs cell variable combinations. Next, graphs of actual cell and chemical data with moderate-high correlation are shown. The order of these graphs is shown in the correlation heat map in Figure 3.9. The final section of results consists of correlation graphs between cell variables and

partition coefficient (logP) constituents. The purpose of this section is to show the relationship of each molecule constituent with cell variables. Each logP variable represents the logP value of previous (if any) and current constituents in the molecule. Starting from logP5, this represents the logP value of terminal group (2 constituents). LogP4 represents the terminal group and the constituent that follows, and this carries until logP1 representing the logP value of up to 6 constituents of a molecule. Below are graphs of correlation significance relative to the sample size and test chosen and in Figure 3.7 are the critical correlation values accepted:

### 3.2.3.1 Correlation significance

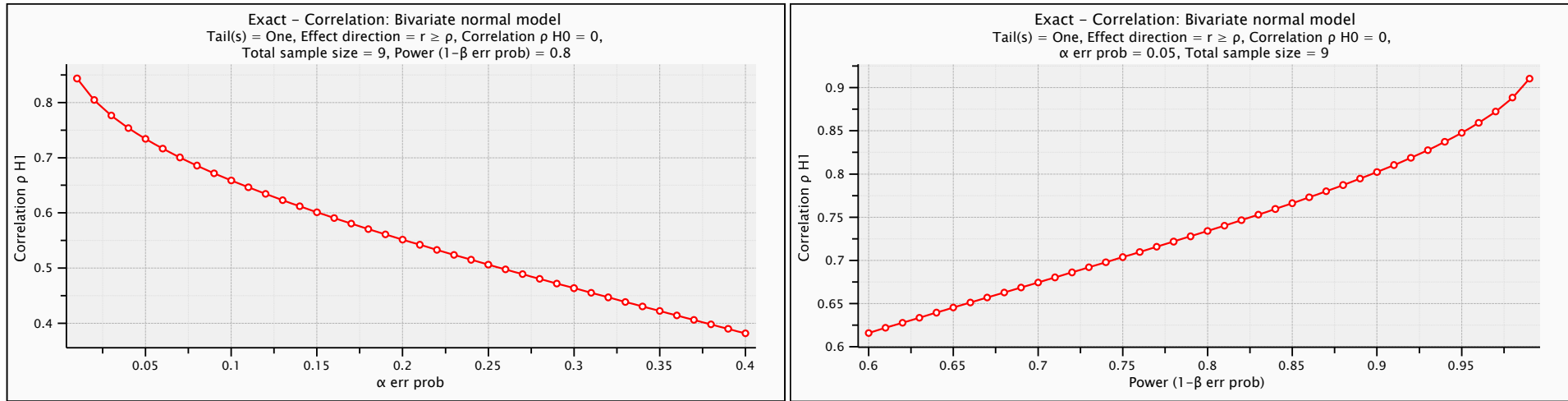


Figure 3.6: Correlation significance with sample size  $n = 9$ . Left: y axis is the correlation coefficient ( $H1$ ) and x axis is the  $\alpha$  probability accepting false positives. Power was set at  $1 - \beta = 20\%$  chance accepting a false negative. Right: y axis is the correlation coefficient ( $H1$ ) and x axis is the  $\beta$  probability accepting false negatives and  $\alpha$  was set at 5% chance accepting a false positive.

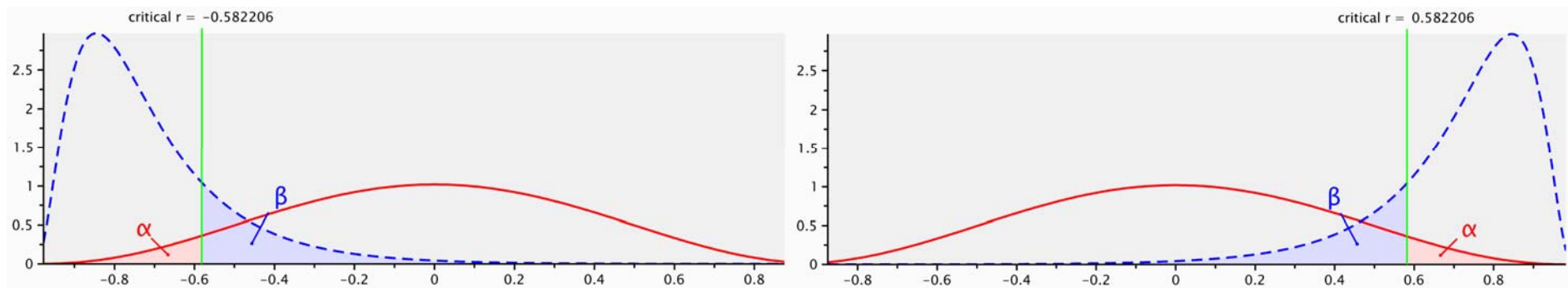


Figure 3.7: Critical correlation coefficient accepted as significant with sample size  $n = 9$ . y axis is the probability density for  $\alpha$  and  $\beta$  distributions and x axis is the correlation coefficient. Left graph shows the critical correlation coefficient for negative correlations and the right one for positive correlations. Correlations  $\geq -0.58$  or  $\leq 0.58$  are accepted as significant.

With a sample size of  $n = 9$ , the graphs (Figure 3.6) show the correlation coefficient ( $r = 0.74$ ) setting the risk of accepting false positives to 5% (left graph) and false negatives (right graph) to 20%. Graphs in Figure 3.7 establish thresholds accepting correlations as significant if they are  $\leq -0.58$  or  $\leq 0.58$ . Outside of these thresholds, the chance accepting a false positive (type I error) and false negatives (type II error) increase. Correlations outside the threshold need further evidence to support them. Below are graphs with significant correlation frequency and below that are the correlation graphs:

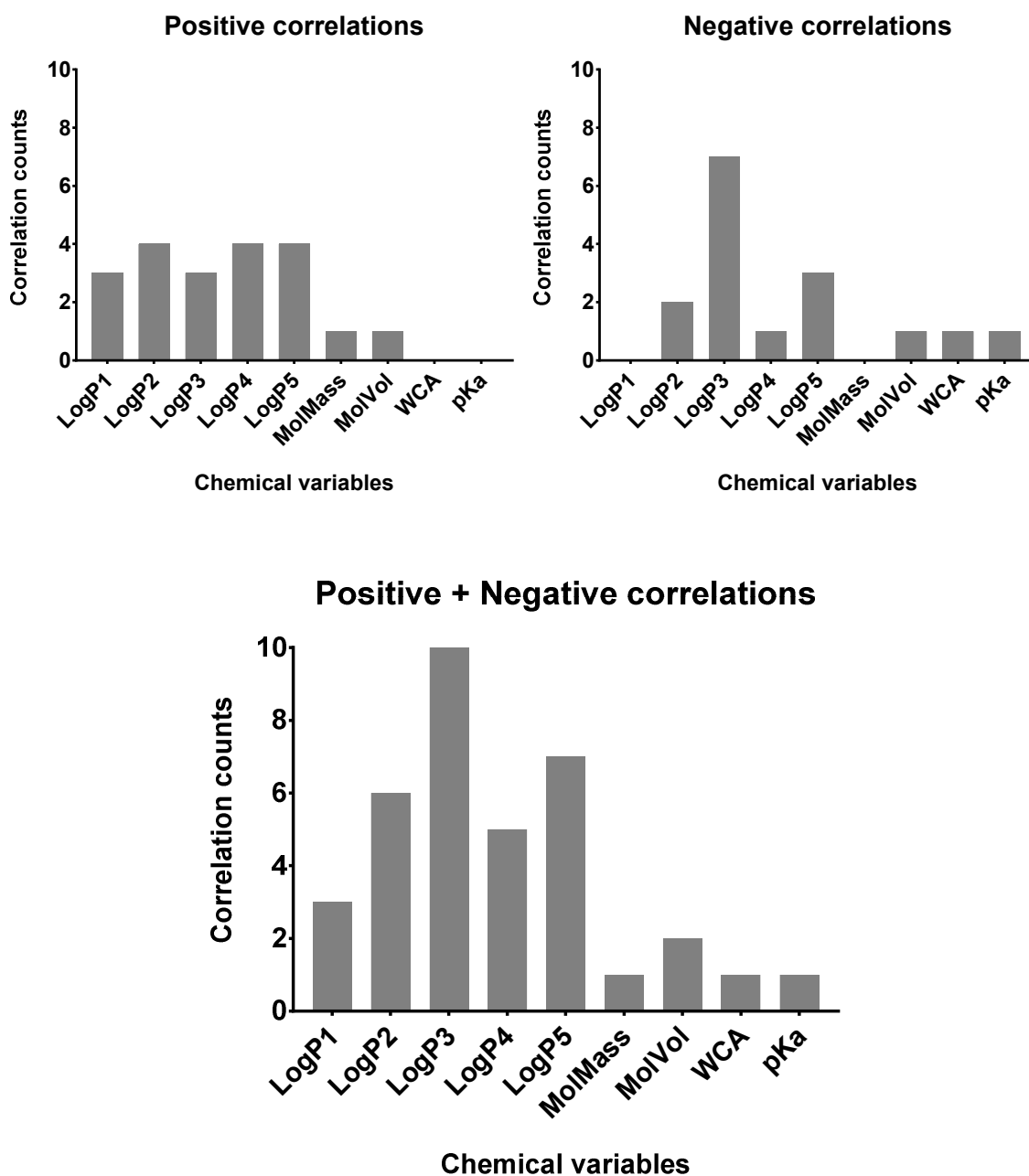


Figure 3.8: Significant correlations counts.  $x$  axis represents the chemical variables and the  $y$  axis represents correlation counts. Top left: positive correlation counts, top right: negative correlation counts, and bottom: total correlation counts.

Positive (+) correlation mean as one variable's value goes up so does the other variable's value. In negative (−) correlation, as one variable's value goes up the other one's decreases. From the top left graph in Figure 3.8, the logP (lipophilicity) group has 3-4 significant positive (+) correlations. From the top right graph, logP3 has most negative (−) correlations followed by logP5. The bottom graph shows the frequency of significant correlations. More correlations were expected from the popular hydrophilicity measure (water contact angle, WCA) and terminal acidity measured (acid dissociation constant, pKa). Nevertheless, the surface lipophilicity measures (logP group) are the most interesting from the above results. Below is a heatmap of all correlation where the darker a cell is the stronger the correlation:

### 3.2.3.2 Correlation heat map

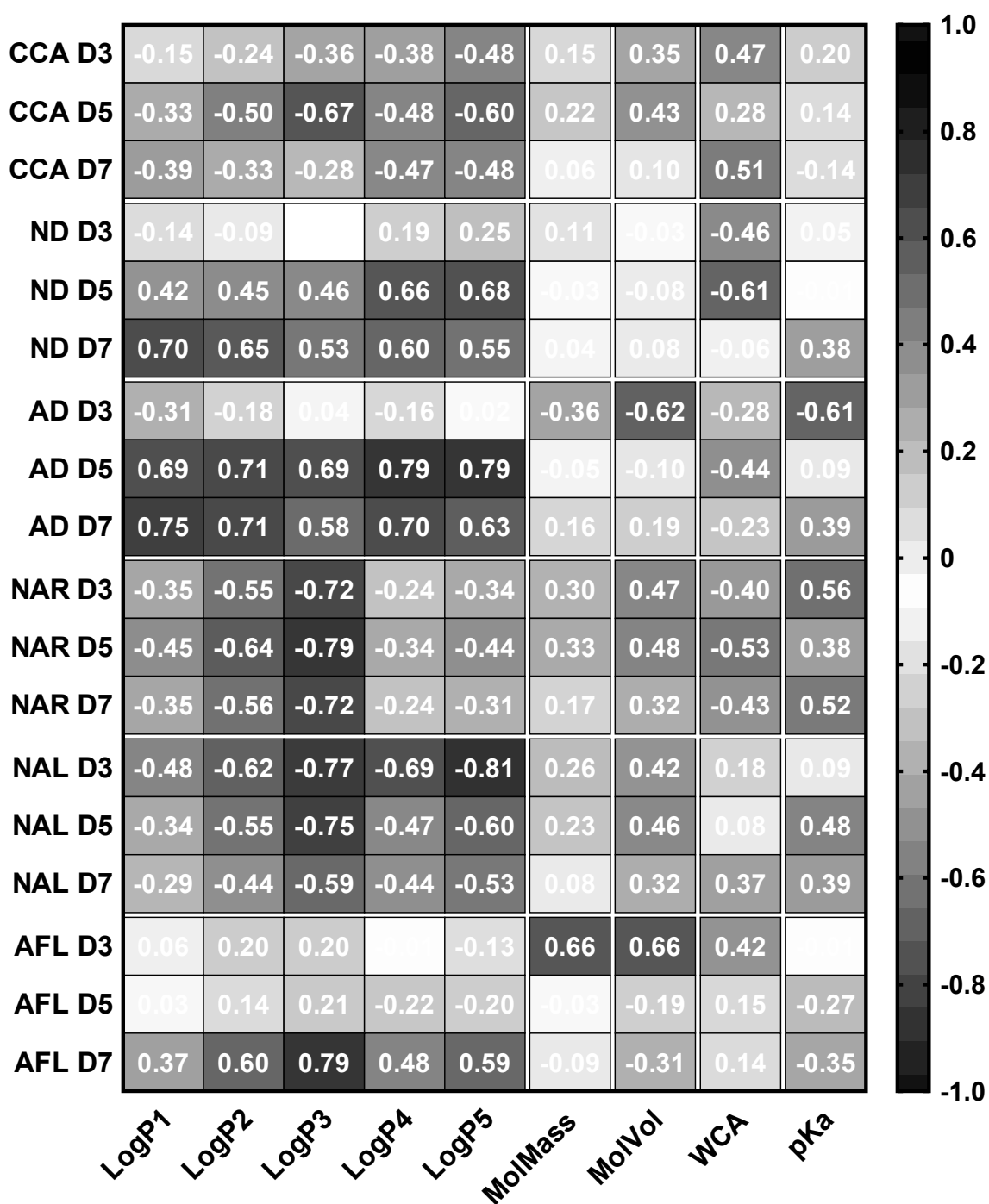


Figure 3.9: Correlation heat map between cell and chemical variables ordered by cell variable type and time point. Absolute correlation values were used. The darker the square the stronger the correlation. Abbreviations: CCA = Cell Cluster Area, ND = Neuron Density, AD = Astrocyte Density, NAR = Neuron-Astrocyte Ratio, NAL = Neuron Axon Length, AFL = Astrocyte Fibre Length, D3 = Day 3, D5 = Day 5, and D7 = Day 7.

The correlation map above in Figure 3.9 shows the absolute correlations and their strength. Correlation directionality will be reported and discussed in the foremost section. This correlation heat map is to study relationships and patterns for each chemical parameter (horizontally), or for each cell parameter per time point (vertically). Significant correlations

here are  $r \geq 0.58$  discovered in section 3.2.3.1108. The “interestingness” of the logP group is shown clearly with cell densities and processes lengths. Other correlations are apparent as well from molecular mass, volume, water contact angle, and acid dissociation constant. Next, we investigate each significant correlation with data plots and cell performance ranks.

### 3.2.3.3 Cell cluster area (CCA)

Neurospheres attached within 1-2 hours on modified surfaces and attachment takes longer on more hydrophobic surfaces. Upon attaching, the neurosphere breaks and cells interacting with the surface differentiate due to cell adhesion molecules (integrin) (48). Glia migrate away from the sphere initially, providing a “carpet” for neurons to migrate (270) as well as providing them with peptides or small proteins for maintenance (neurotrophic factors). Neurons perform independent short-range migrations out of the spheres in a process called chain migration (271).

Fluorescence microscopy and chemical markers were used to identify cell types. Cell images were captured in three time-points day 3, 5 and 7. Day 3 informs of biological/material interface and day 7 informs on biological remodelling of the environment. Exceeding 7 days in culture is challenging with differentiated neurons. The conversion of neurospheres to cell clusters is relevant to neural stem cell differentiation. Larger areas of cell clusters means more stem cells and progenitors differentiate to mature nervous cells (271). Below are graphs with raw data of cell cluster areas against chemical parameters for all time points. Data selected have significant correlations ( $r \leq -0.58, r \geq 0.58$ ) in at least one time point. Below is a figure with significant correlations between cell cluster area vs chemical parameters followed by a cell performance rank table:



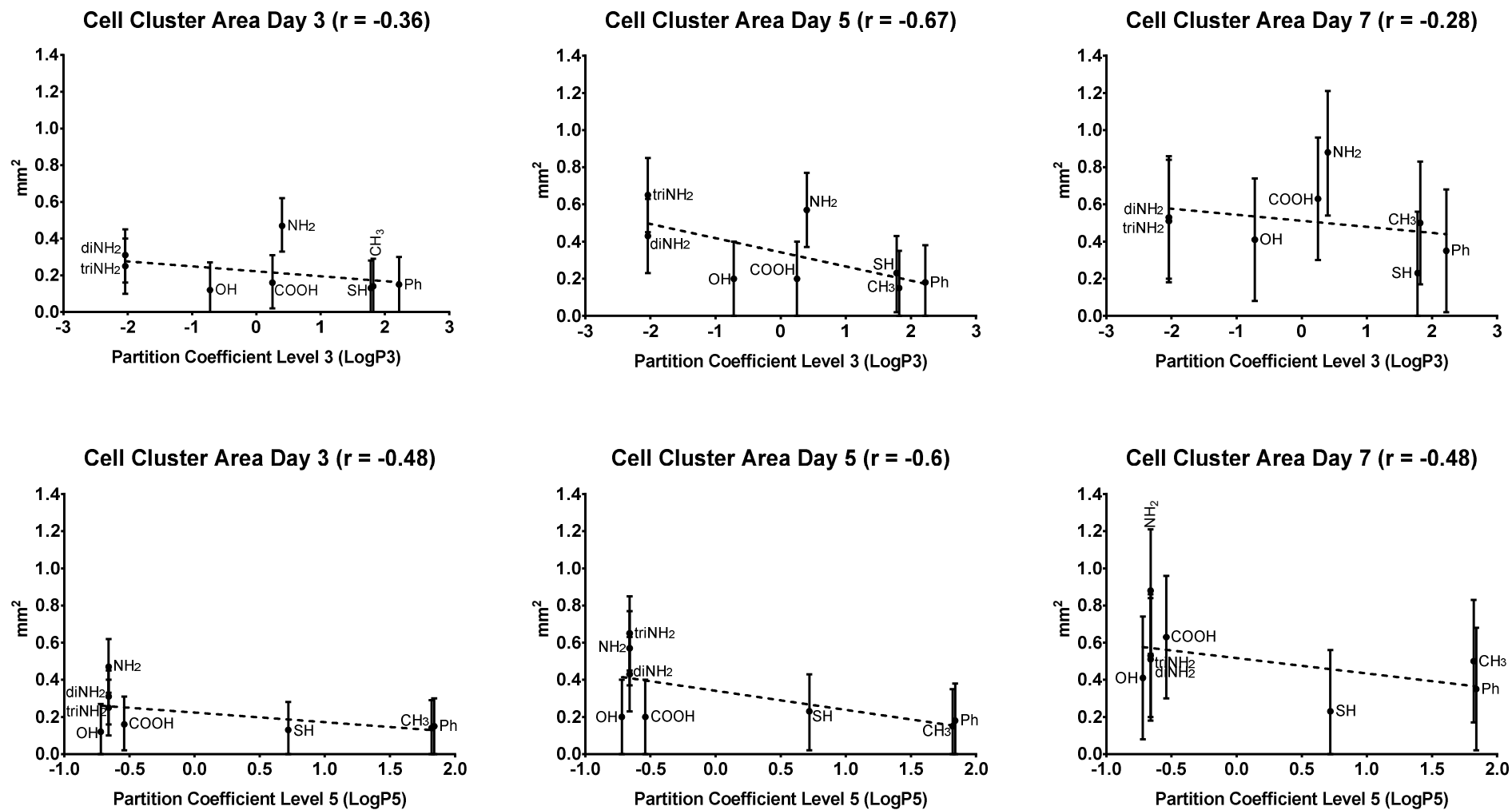


Figure 3.10: Data plots of cell cluster area vs logP3 and logP5. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

The table below shows the cell cluster area ranks for all cell culture environments for comparison:

Table 3.4: Cell cluster area rank in culture environments. Ranks are calculated with Bray & Curtis dissimilarity on related cell variables from all 3 time points. The lower a chemistry's rank is to 0 the closer its cell performance is to that of laminin's. Median absolute deviation is  $\pm 0.23 \text{ mm}^2$ .

Environment	Median value ( $\text{mm}^2$ )	Rank
P/LAM	0.74	0.00
NH <sub>2</sub>	0.64	0.17
triNH <sub>2</sub>	0.47	0.23
diNH <sub>2</sub>	0.42	0.28
COOH	0.33	0.38
CH <sub>3</sub>	0.27	0.47
OH	0.24	0.51
Ph	0.23	0.53
SH	0.2	0.58

Referring to The table below shows the cell cluster area ranks for all cell culture environments for comparison:

Table 3.4, laminin (P/LAM) with and amine (NH<sub>2</sub>) with surfaces give the largest cell cluster areas overall. Both environments had a marked difference in area expansion 50%. A surprise here is the carboxylic acid surfaces (COOH) had triple expansion on day 7, closing in on laminin. Thiol (SH), phenol (Ph), hydroxyl (OH) and methyl (CH<sub>3</sub>) provide the smallest cluster areas overall. Methyl (CH<sub>3</sub>) cell cluster areas changed very little in time. Median absolute deviation of cell cluster area measurements is  $\pm 0.23 \text{ mm}^2$ .

The data show cell cluster area (CCA) to correlate (–) moderate-strong with logP3, logP4, and logP5 on both day 5 and 7 (Figure 3.10). In addition, there is a (+) strong correlation with water contact angle (WCA) but this is not significant as it falls below the critical  $r$  value ( $< 0.58$ ). (Figure 3.9). Additional evidence in support is required for this relationship.

### 3.2.3.1 Neuron density (ND)

Successful cellular therapies to regenerate nervous tissue depend partly on the amount of neurons delivered. Neuronal network is the functional component of the nervous system. Cells around the cell cluster but not the dense centre were quantified in stratified random sampling. Increasing the density of transplant relevant populations is a key element in scaling up the therapy. Cell culture environments with synthetic chemistry provide greater degree of control compared to alternatives such as special culture media, and hypoxia as an environmental culture condition among others.

Day 3 neural density informs on neural differentiation. At this stage, high density means cells reside inside the neurosphere. Low neural density is a strong indicator of differentiation. Day 5 and 7 time-points inform on biological remodelling of the environment and cell proliferation due to the longer duration in culture (101). In a situation where neural density is similar but the cell cluster area is larger means neural cells are dividing. In tissue slices and xenografts, higher cell density means smaller extracellular volume and amount suggesting cells use the resources in the vicinity quicker (272,273). Low cell density promotes internal cell signalling for changes within individual cells (autocrine signalling); high cell density promotes cell-cell communication inducing changes in nearby cells (paracrine signalling) (274). Cell densities are comparable between different environments by standardising the cell counts with their cell cluster area. Below are graphs with raw data of neuron density against chemical parameters and right below those are the cell performance ranks:

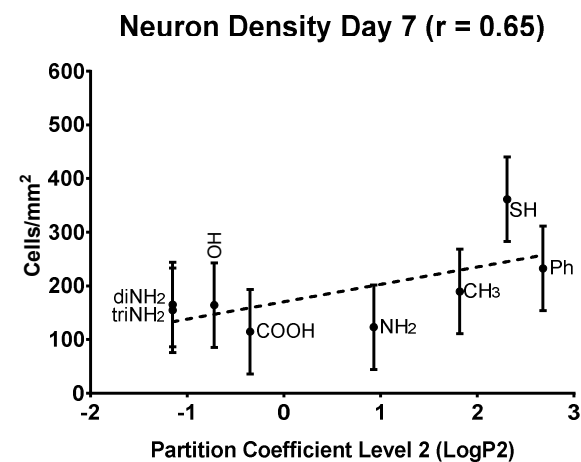
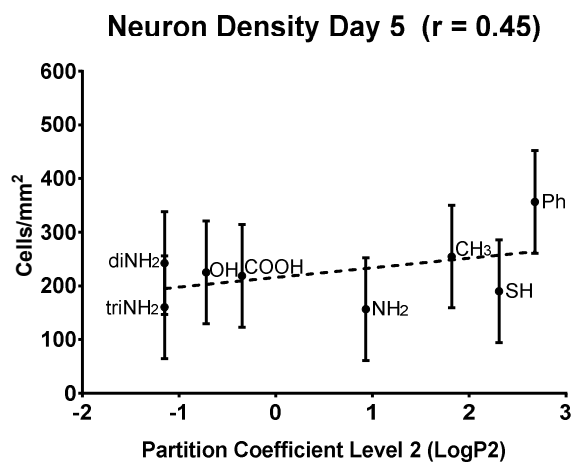
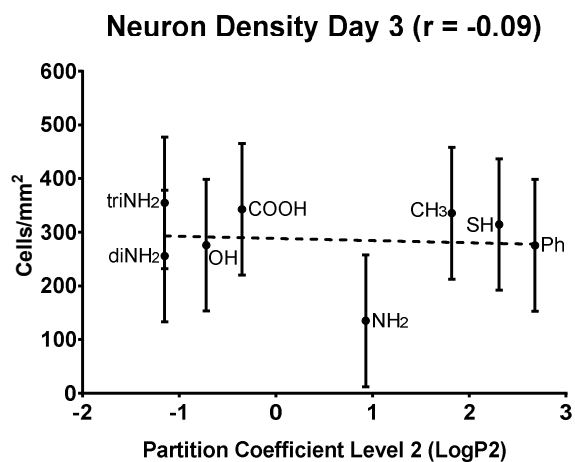
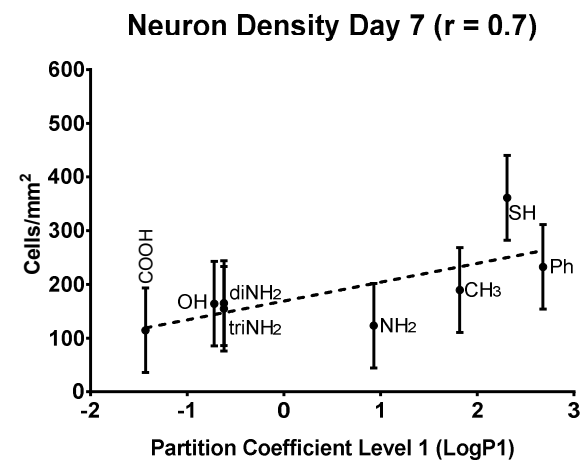
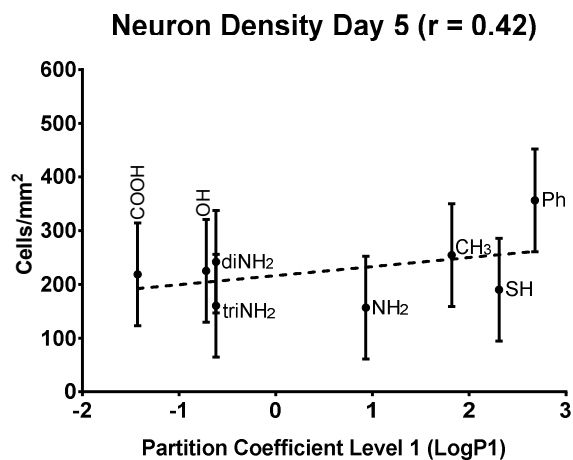
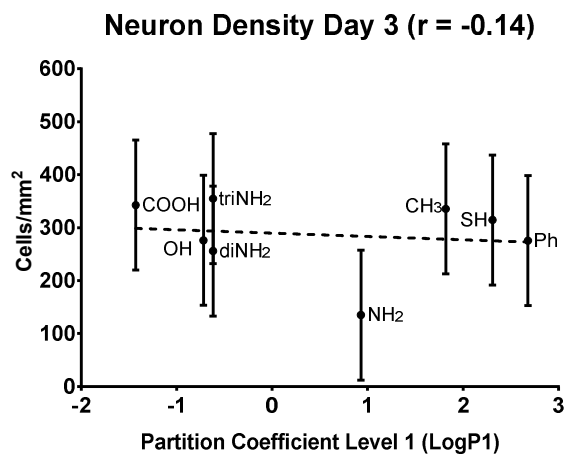


Figure 3.11: Data plots of neuron density vs logP1 and logP2. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

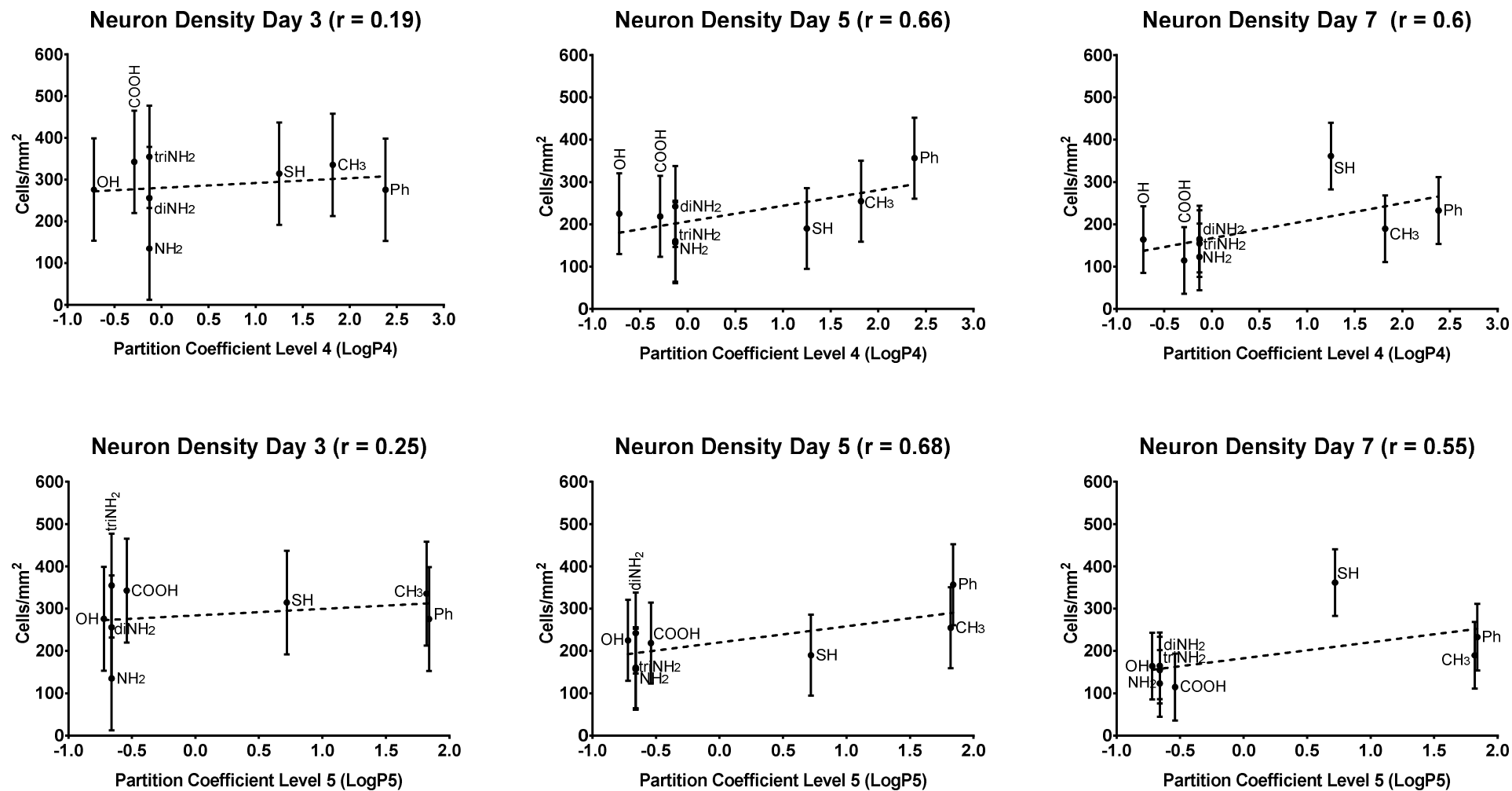


Figure 3.12: Data plots of neuron density vs logP4 and logP5. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

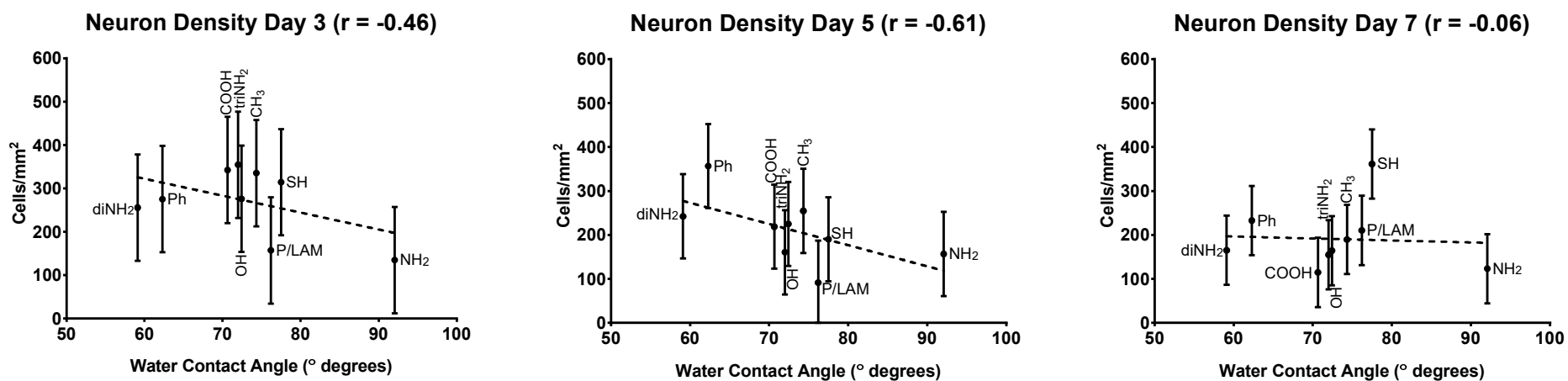


Figure 3.13: Data plots of neuron density vs water contact angle. Column 1 (left) represents cell data from day 3; column 2, data form day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

Table 3.5: Neuron density rank in culture environments. Ranks are calculated with Bray & Curtis dissimilarity on related cell variables from all 3 time points. The lower a chemistry's rank is to 0 the closer its cell performance is that of laminin's. Median absolute deviation is  $\pm 99$  cells/mm<sup>2</sup>.

Environment	Median value (cells/mm <sup>2</sup> )	Rank
P/LAM	152.87	0.00
NH <sub>2</sub>	138.25	0.20
diNH <sub>2</sub>	221.01	0.26
OH	221.75	0.27
triNH <sub>2</sub>	223.28	0.29
COOH	225.30	0.29
CH <sub>3</sub>	259.87	0.29
Ph	288.25	0.31
SH	288.65	0.31

Referring to Table 3.5, laminin environment is our biological standard and matching its performance is the priority. The order of results in the table is the mathematical distance to laminin's neuron density. Amine (NH<sub>2</sub>) is the best performer followed by laminin (P/LAM) in providing lowest neuron density. Lowest performers providing high neuron density are thiol (SH) and phenol (Ph) followed by carboxylic acid (COOH). Median absolute deviation of neuron density measurements is  $\pm 99$  cells/mm<sup>2</sup>.

From the graphs above in Figure 3.11 and Figure 3.12, neuron density correlates (+) moderate to strong with logP1, 2, 4, and logP5 on day 5 and (+) strong on day 7. In addition, ND has a (–) moderate to strong correlation with WCA on day 3 and 5 ( ). Thiol is changing the relationship with WCA on day 7 and a (–) correlation is expected here as well.

#### 3.2.3.2 Astrocyte density (AD)

Astrocytes are robust glial cells that play several roles in the central nervous system. They manage chemical signals (neurotransmitters) exchanged by neurons, strengthen neuron connections (synapses) called long-term potentiation (275). They also regulate ion concentration (e.g. potassium) in the extracellular space where excess amounts depolarise neurons that could result in epileptic activity (276). Other purposes of astrocytes include antioxidant defences, anti-inflammatory response, and energy metabolism (275,277).

There is a body of evidence in the literature of the importance of astrocytes in neuro-regeneration and neuro-repair. During development, ependymal cells and astrocytes form glial tubes used by migrating neuron pre-cursors (neuroblasts). In these tubes, astrocytes provide support for migrating cells as well as insulation from chemical and electrical signals released from surrounding cells. For tissue replacement therapies, astrocyte and neural stem cells (NSCs) exhibit a suppressive effect on an allogeneic immune response due to

cell–cell interaction (278). This means nervous tissue transplants containing astrocytes are more likely to be accepted by the patient’s immune system.

*In vitro*, astrocytes regulate the ionic or chemical milieu of neurons to aid neuron signalling (279). They were also found to direct neurite alignment to a greater extent compared to structured surface cues, highlighting their importance for biochemical signalling and cellular architecture (279). Lastly, conditioned media with biomolecules produced by astrocytes increase NSCs’ proliferation, differentiation, and participate in the modulating the cells (280). These findings mean astrocytes have important roles as early as the development stage and even the repair stage of the nervous system.

As previously, day 3 astrocyte density informs on differentiation. High density means cells reside inside the neurosphere whereas low astrocyte density is a strong indicator of differentiation and migration (271). Day 5 and 7 time-points are good indicators of proliferation (101). For the purposes of this project, astrocyte density may provide insights for the effect of chemistry on neural cells. Below are data plots of astrocyte density vs chemical variables followed by the cell performance ranks (Table 3.6):



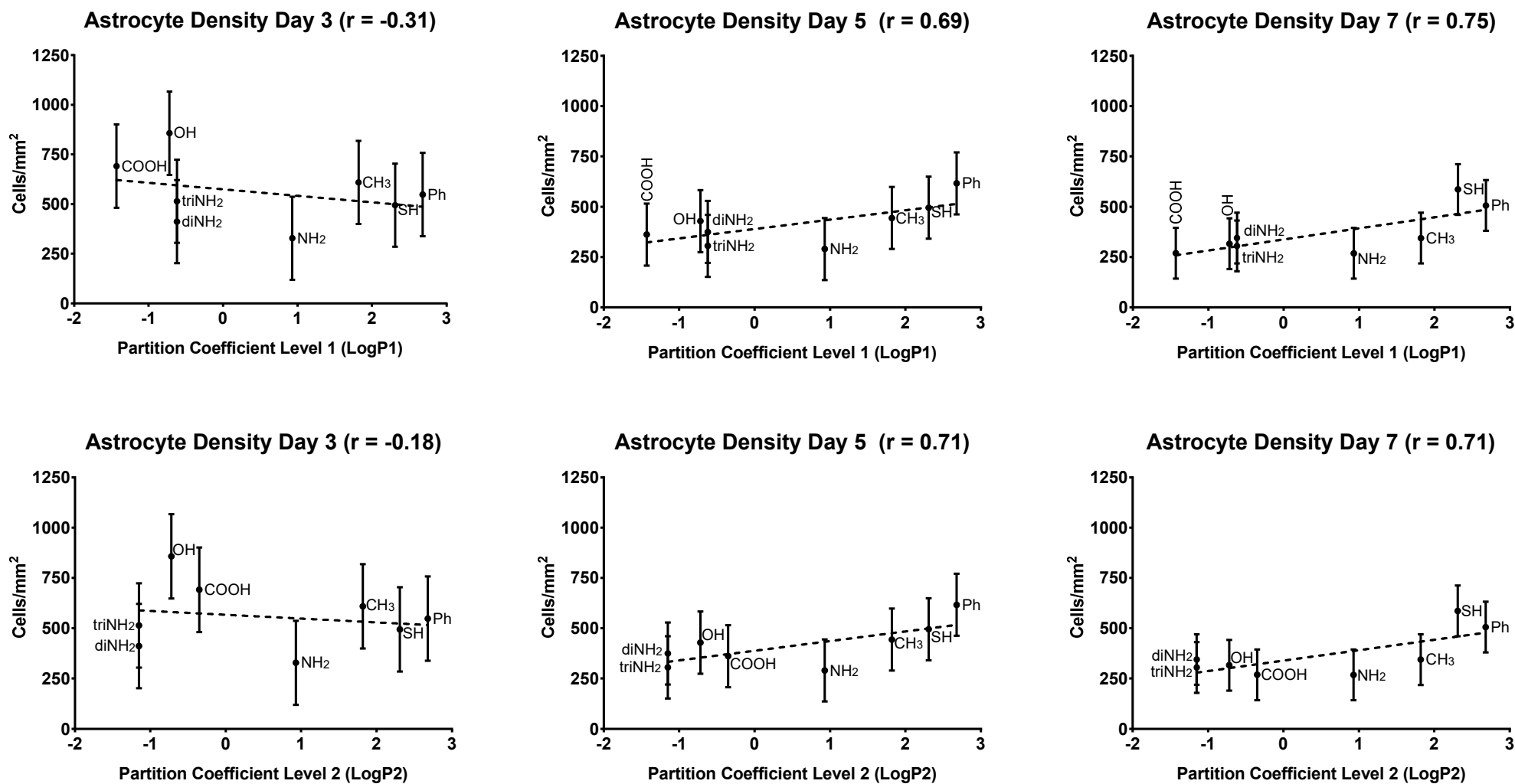


Figure 3.14: Data plots of astrocyte density vs logP1 and logP2. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

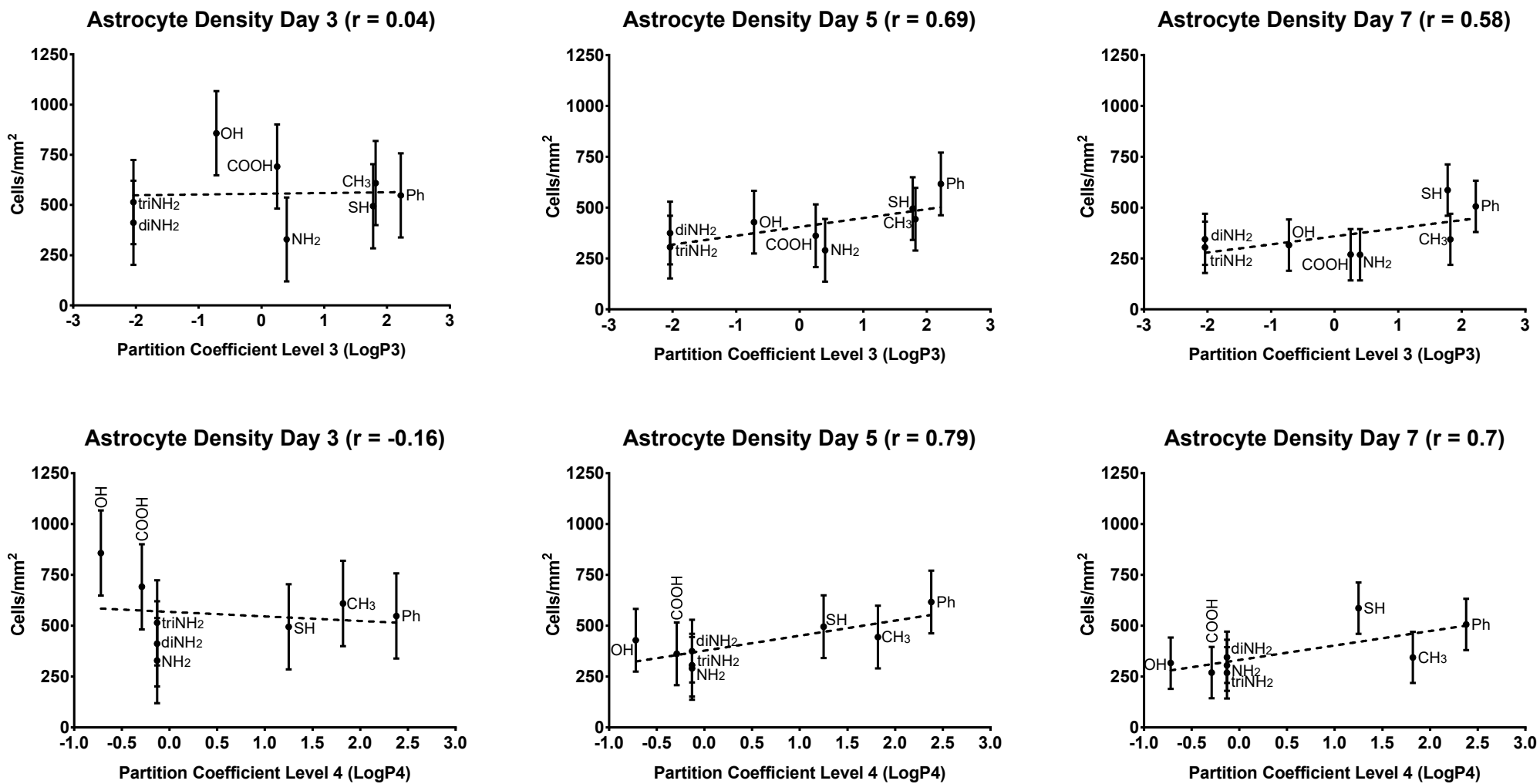


Figure 3.15: Data plots of astrocyte density vs logP3 and logP4. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

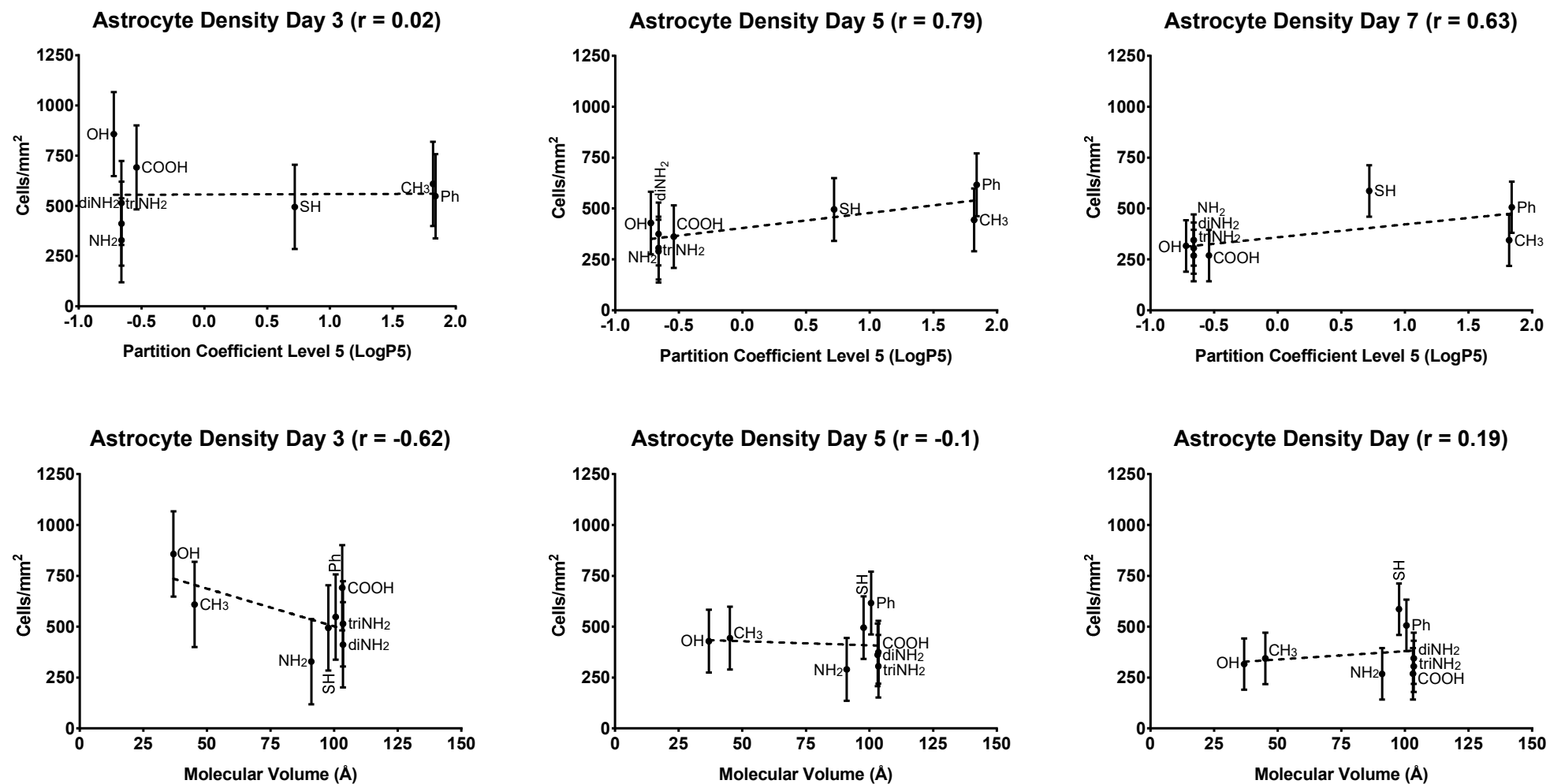


Figure 3.16: Data plots of astrocyte density vs logP5 and Molecular Volume. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation.

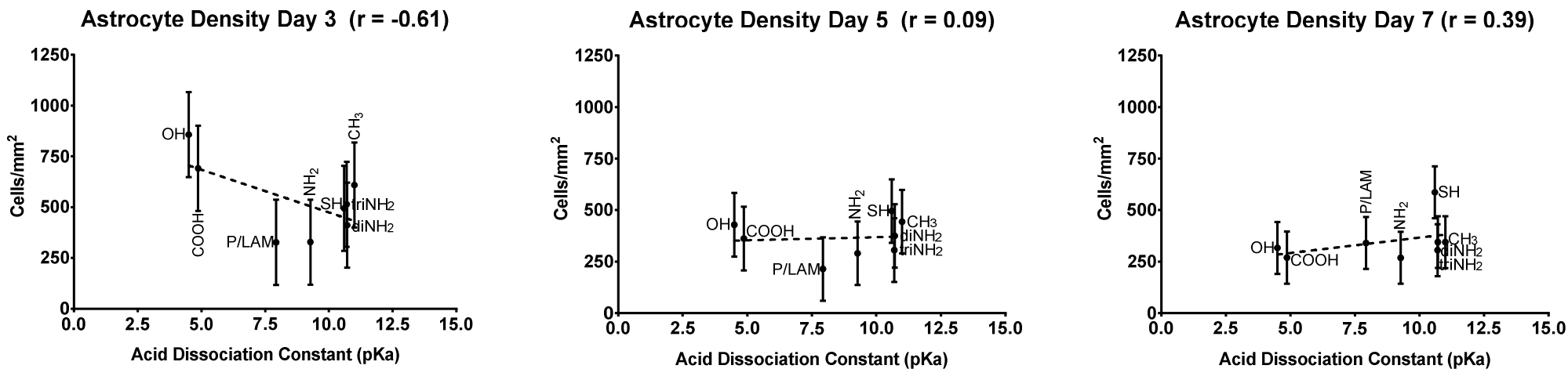


Figure 3.17: Data plots of astrocyte density vs pKa. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$ median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

Table 3.6: Astrocyte density rank in culture environments. Ranks are calculated with Bray & Curtis dissimilarity on related cell variables from all 3 time points. The lower a chemistry's rank is to 0 the closer its cell performance is that of laminin's. Median absolute deviation is  $\pm 163$  cells/mm<sup>2</sup>.

Environment	Median value (cells/mm <sup>2</sup> )	Rank
P/LAM	293.78	0.00
NH <sub>2</sub>	295.60	0.08
diNH <sub>2</sub>	376.93	0.12
triNH <sub>2</sub>	375.20	0.16
COOH	440.68	0.21
CH <sub>3</sub>	465.65	0.23
SH	525.31	0.28
Ph	556.93	0.31
OH	534.02	0.31

Lower cell density indicates good cell interaction with their environment. Cell migration is indicative of cell differentiation to neurons and glia. Referring to Table 3.6, highest performers are laminin (P/LAM) and amine (NH<sub>2</sub>). Lowest performers are phenol (Ph), hydroxyl (OH) and thiol (SH). Median absolute deviation for these measurements is  $\pm 163$  cells/mm<sup>2</sup>.

Astrocyte density (AD) correlates with the logP group (+)strong on day 5 and 7. AD also correlates (–)strong with both molecular volume and acid dissociation constant (pKa) of the terminal group on day 3. This effect vanished on day 5 and by day 7, (+)weak correlations are observed instead. This means, initially AD decreases on more acidic surfaces. The interesting part here is the pKa correlation changes to (+)weak by day 5 and by day 7, it changes to (+)moderate.

### 3.2.3.3 Neuron/astrocyte ratio (NAR)

The key challenge in cell therapy translation is controlling the proportion of neurons and the purity of transplant population is a critical quality attribute (281). An imbalance in the proportion and migration of cells can have adverse effects for transplant recipients. Such effects include increase in uncontrolled movement due to the production of serotonin in excess or at the wrong location in the transplant (282). Another effect is teratomas from progenitors or stem cells if they are present in the transplant tissue (91). Generally, glial cells dominate cultures compared to neurons, which are of interest as the functional component of the nervous system. This cell proportion imbalance likely occurs due to asymmetric cell division of neurons and glia progeny (282).

From cell density data, it is possible to obtain the ratio of cells expressed as  $\frac{\text{Neuron Density}}{\text{Astrocyte Density}}$ .

Low cell density coupled together with high neuron percentage is preferred. The former means cells have migrated away from the neurosphere and differentiated and the latter is an indication of the proportion of neurons. Both help understand the relationship between neural cell division and time across different environments. Below is a figure with significant correlations between neuron/astrocyte ratio vs chemical parameters followed by a table with cell performance ranks:

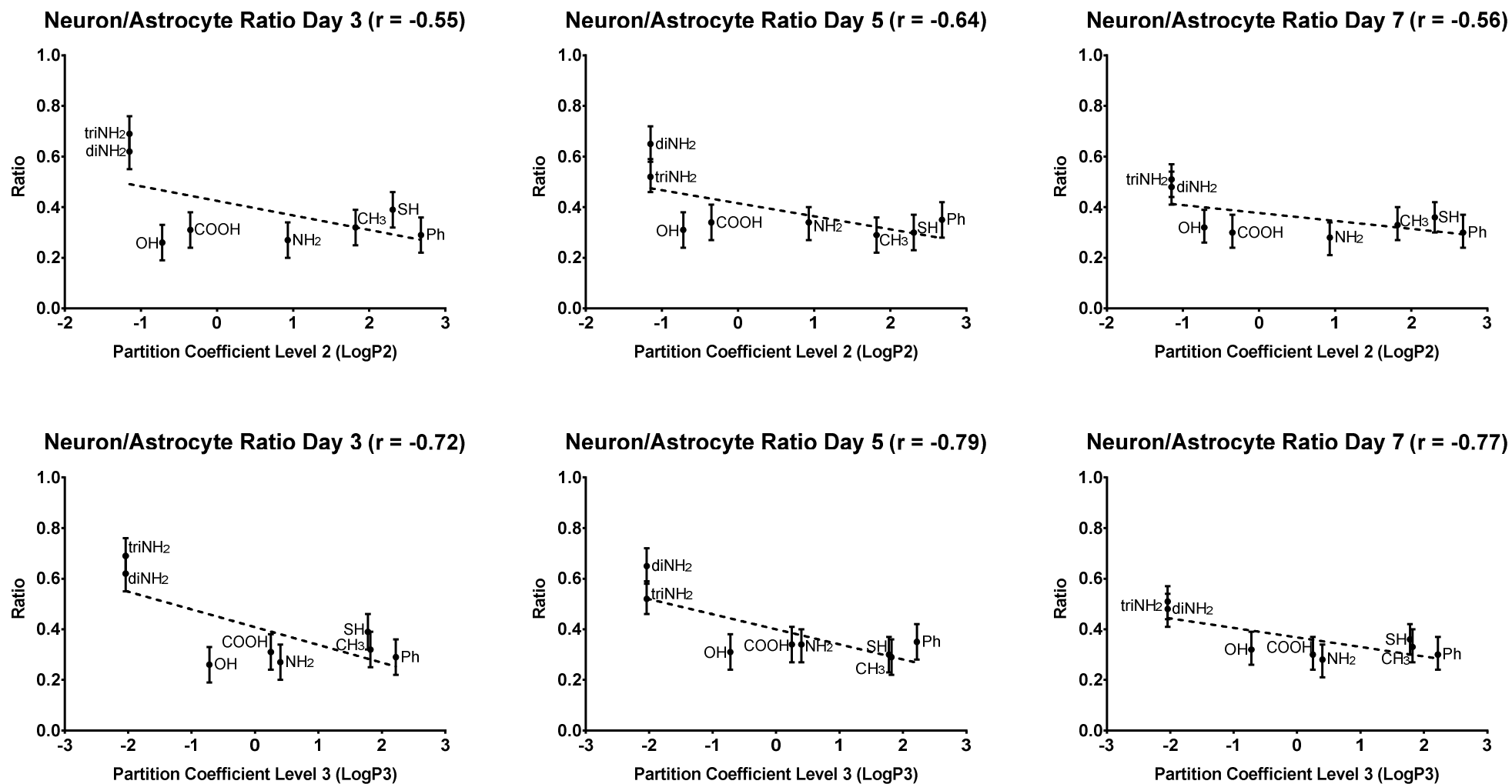


Figure 3.18: Data plots of neuron/astrocyte ratio vs logP2 and logP3. Column 1 (left) represents cell data from day 3; column 2, data form day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$ median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

Table 3.7: Neuron/astrocyte ratio rank in culture environments. Ranks are calculated with Bray & Curtis dissimilarity on related cell variables from all 3 time points. The lower a chemistry's rank is to 0 the closer its cell performance is that of laminin's. Median absolute deviation for the ratio is  $\pm 0.07$ .

Environment	Ratio (Neuron/Astrocyte)	Rank
P/LAM	0.35	0.00
SH	0.35	0.04
CH <sub>3</sub>	0.31	0.05
COOH	0.32	0.06
Ph	0.32	0.07
OH	0.30	0.08
NH <sub>2</sub>	0.29	0.09
triNH <sub>2</sub>	0.57	0.24
diNH <sub>2</sub>	0.58	0.25

The proportion of neurons standardised by the number of astrocytes expresses the ratio between them. A high value of this ratio means more differentiation to neurons than astrocytes. The ranks indicate the similarity to laminin's obtained value for neuron/astrocyte ratio. Thiol (SH) has the same neuron percentage (NAR) as laminin. Diamine (diNH<sub>2</sub>) and triamine (triNH<sub>2</sub>) exhibit the highest differentiation to neurons. Apart from amines, remaining environments stacked up favourably to the gold standard laminin.

Diamine (diNH<sub>2</sub>) and triamine (triNH<sub>2</sub>) are changing the relationship on day 3 but in later time points, they are in line with the rest of the data. Neuron/astrocyte ratio (NAR) correlates (–)strong with logP2 and logP3 on day 3,5, and 7 (Figure 3.18).

#### 3.2.3.4 Neuron axon length (NAL)

Functional nerve tissues consist of neural projections to communicate with neighbouring cells using electrical conduction across large distances. Neuron axon length is a good indicator of this *in vitro*. One aim of neuro-regenerative biomaterials is to grow and guide neurons to specific injury areas and re-wire compromised neural circuits to restore function. Biomaterials have been used to successfully guide neuron contact where they followed surface cues (283). In a more recent example, neurons have been aligned to



nanofiber surfaces (284). The challenge here is to use simple means to control the axon length to allow effective re-wiring of a neural circuit for stem cell therapies. Measurements were taken for 300 neurons per surface from clearly labelled cells (tuj1) with the entire neurite length visible (41). As previously, below is a figure with significant correlations between neuron axon length vs chemical parameters followed by a table with cell performance ranks:

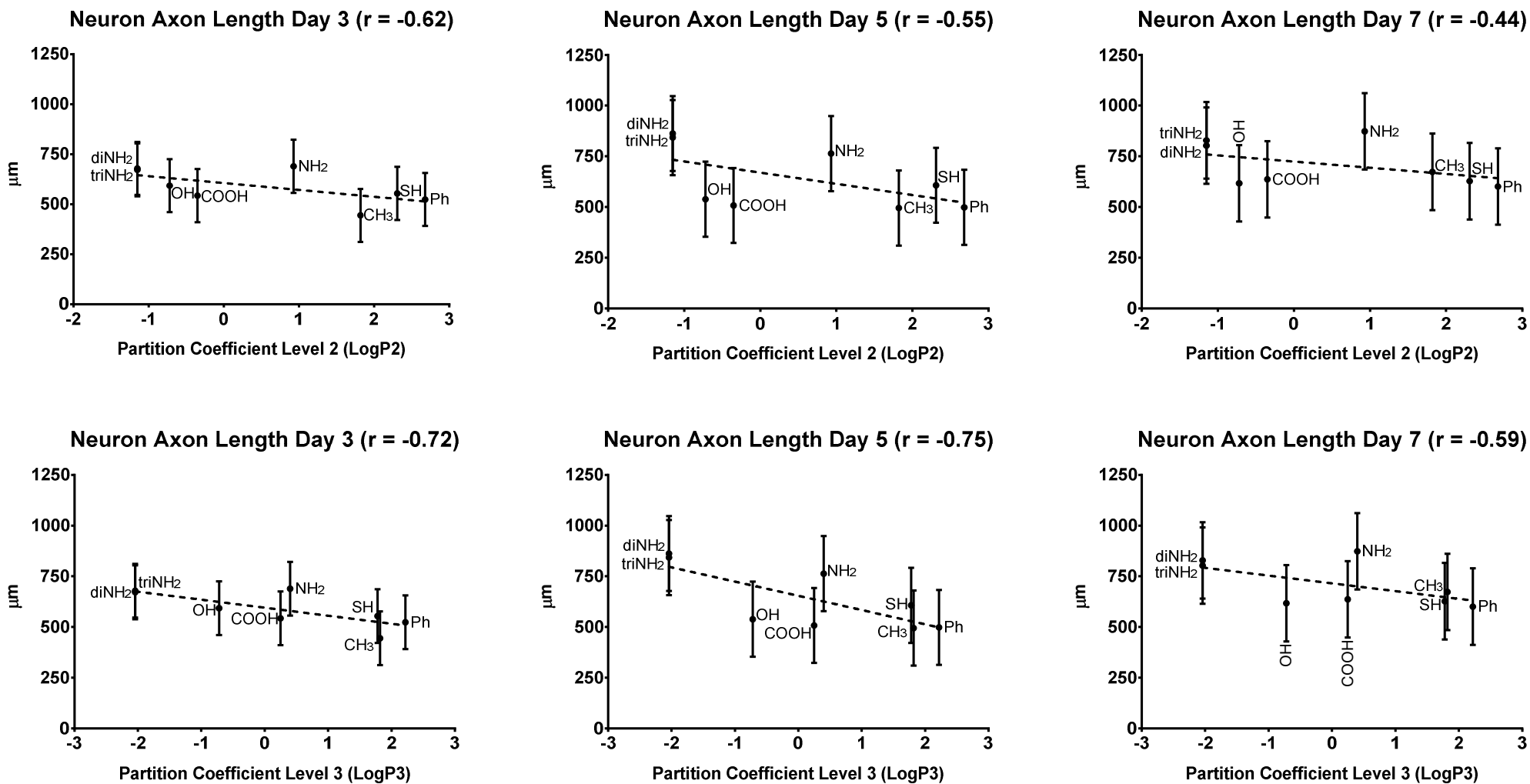


Figure 3.19: Data plots of neuron axon length vs logP2 and logP3. Column 1 (left) represents cell data from day 3; column 2, data form day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

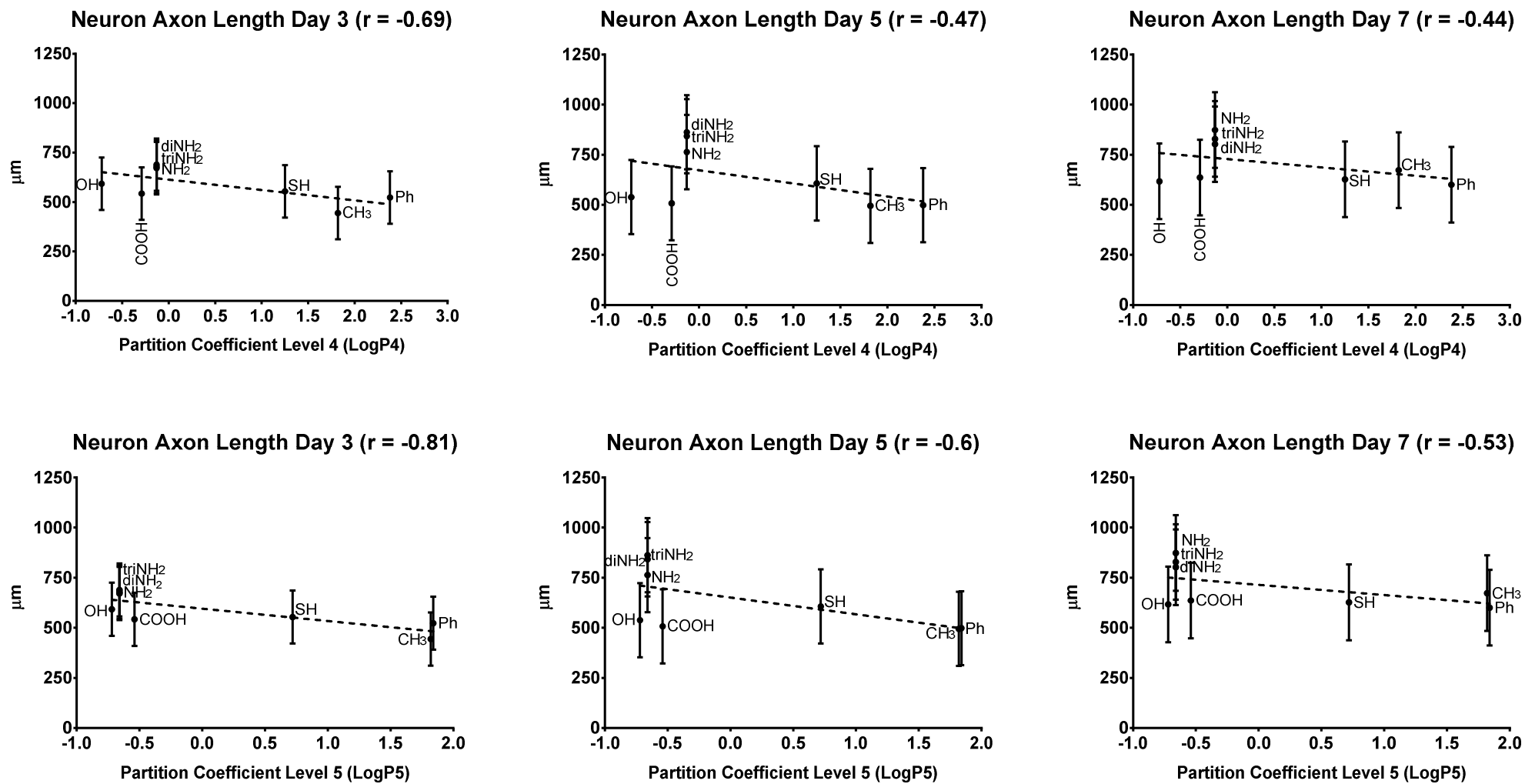


Figure 3.20: Data plots of neuron axon length vs logP4 and logP5. Column 1 (left) represents cell data from day 3; column 2, data form day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

Table 3.8: Neuron axon length rank in culture environments. Ranks are calculated with Bray & Curtis dissimilarity on related cell variables from all 3 time points. The lower a chemistry's rank is to 0 the closer its cell performance is that of laminin's. Median absolute deviation is  $\pm 168 \mu\text{m}$ .

Environment	Median value ( $\mu\text{m}$ )	Rank
P/LAM	746.97	0.00
NH <sub>2</sub>	775.23	0.02
triNH <sub>2</sub>	780.87	0.04
diNH <sub>2</sub>	781.16	0.05
SH	595.98	0.11
OH	582.68	0.12
COOH	562.12	0.14
Ph	540.69	0.16
CH <sub>3</sub>	537.28	0.16

The order of the results in Table 3.8 is the mathematical distance to laminin's neuron axon length. The amine group ( $diNH_2 > triNH_2 > NH_2$ ) showed the longest axons followed closely by laminin. Remaining environments had similar axon lengths over all time points. Median absolute deviation for these measurements is  $\pm 168 \mu\text{m}$ .

Neuron axon length (NAL) correlates with the logP group. The correlations are (–)strong with logP3 and logP5 for all time points. LogP1, logP2 and logP4 correlate (–)moderate to strong on day 3 and day 5 (Figure 3.19, Figure 3.20).

### 3.2.3.5 Astrocyte fibre length (AFL)

Astrocytes have several roles in the nervous system. They manage neurotransmitters, ionic regulation, synaptic processing, anti-inflammatory response, antioxidant defences, and energy metabolism (275). During development, ependymal cells and astrocytes form glial tubes used by migrating neuron precursors. Astrocytes also insulate neurons from chemical and electrical signals released from surrounding cells. In addition, conditioned media with biomolecules produced by astrocytes increase NSCs' proliferation, differentiation, and participate in modulating the cells (280).

*In vivo*, astrocyte processes mediate between blood capillaries and other cells transporting energy substrates as metabolic fuel for brain activity (285). Astrocyte processes play a key role in glial/axonal interactions (279). They contact neuron bodies (somata) and enclose active neuron connection (synaptic) terminals (286). They are also associated with another glial cell type, oligodendrocytes, in shielding neuron axons (myelination). We know this as more astrocytes appear during development in the normal developmental period of myelination in the spinal cord (287). *In vitro* studies show astrocytes induce oligodendrocytes to align their processes with axons thereby controlling the onset of axon insulation (myelination) (288). Myelination is an important attribute of oligodendrocytes (induced by astrocytes) for developing functional neural circuits.

Astrocyte spreading is related with fibre length as astrocytes extend protrusions to interact with other cells and with the surface for migration and attachment (93). For this project, astrocyte fibre length is an indicator of the indirect relationship astrocytes have with the culture environment. It says more about neuron availability and migration as astrocyte processes reach out further for neurons that are sparse or distant. Therefore, the shorter the processes the more likely neurons are within the vicinity. Below is a figure with significant correlations between astrocyte fibre length vs chemical parameters followed by a table with cell performance ranks:

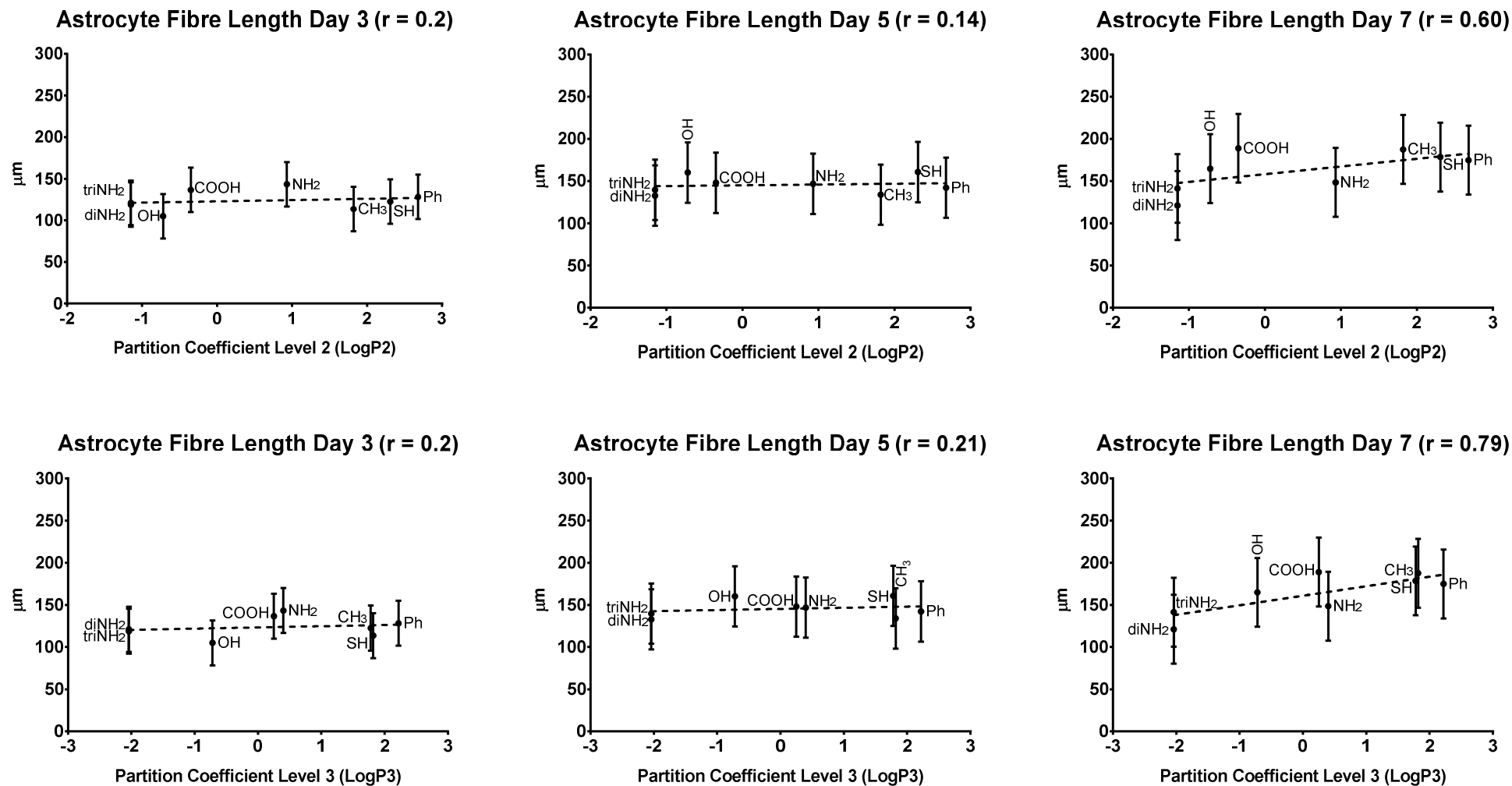


Figure 3.21: Data plots of astrocyte fibre length vs logP2 and logP3. Column 1 (left) represents cell data from day 3; column 2, data from day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$  median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

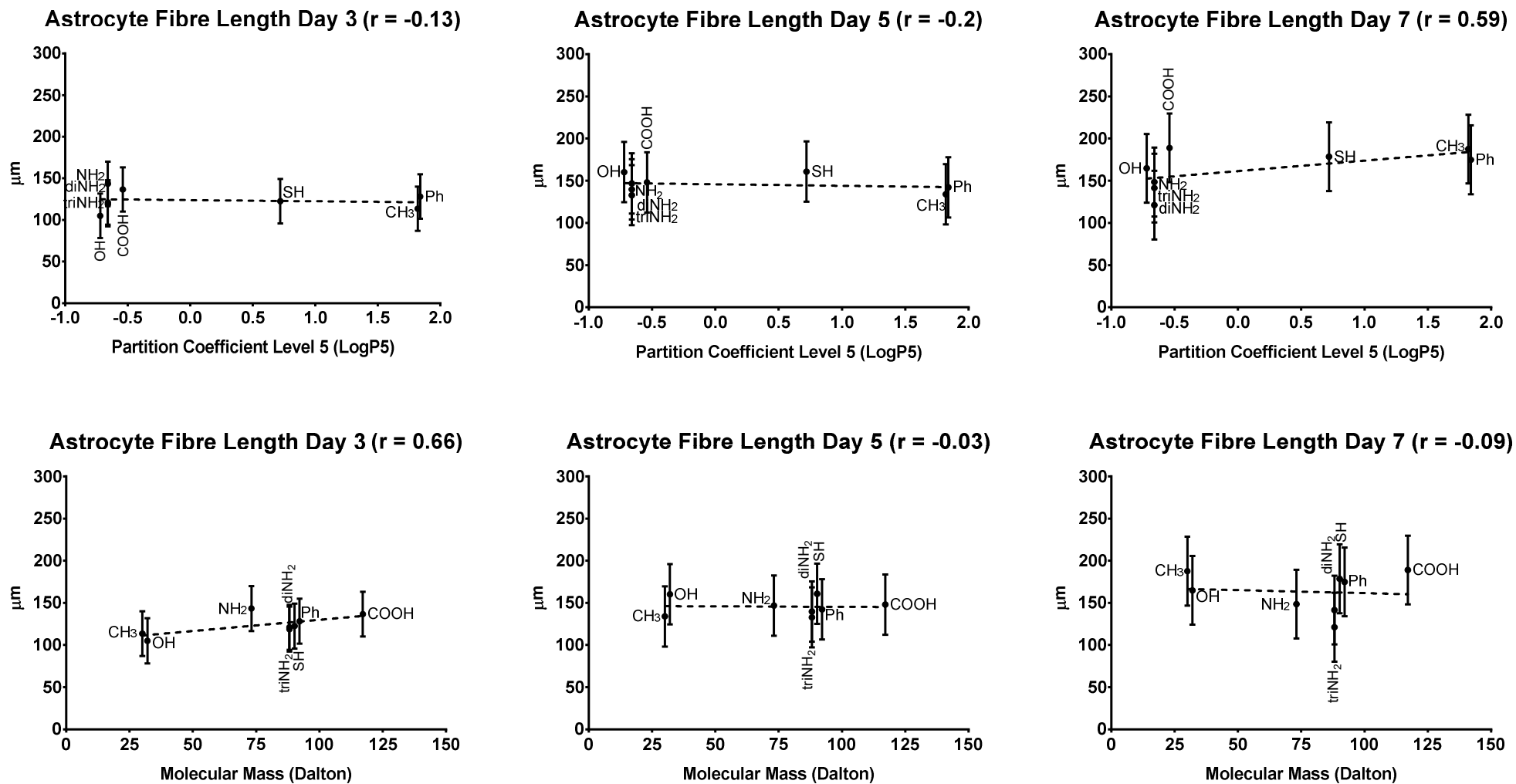


Figure 3.22: Data plots of astrocyte fibre length vs logP5 and molecular mass. Column 1 (left) represents cell data from day 3; column 2, data form day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$ median absolute deviation. The standard curve is a linear regression model fitted on data as a reference for linear relationships.

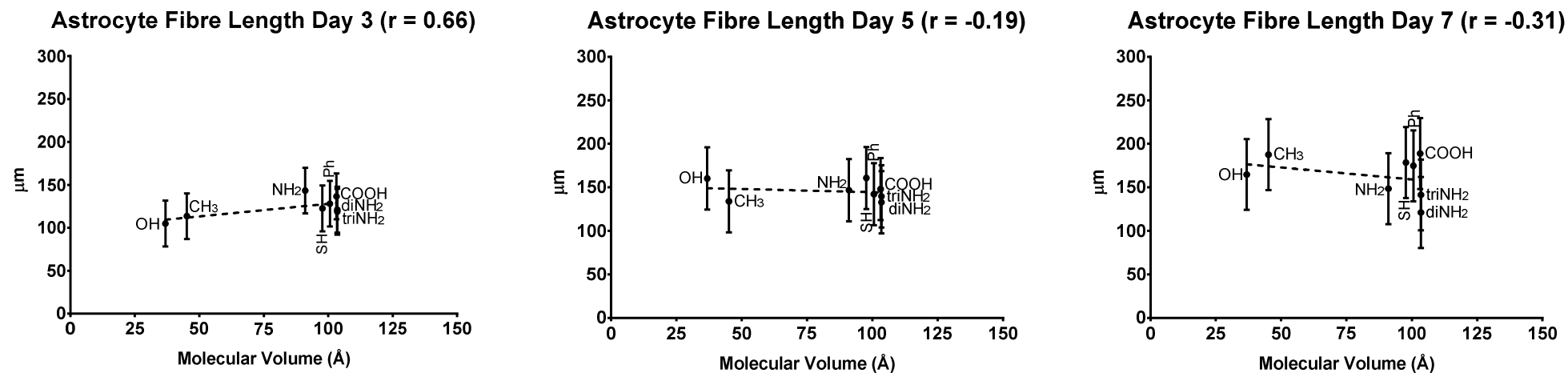


Figure 3.23: Data plots of astrocyte fibre length vs molecular volume. Column 1 (left) represents cell data from day 3; column 2, data form day 5; and column 3, data from day 7. Data point labels represent the abbreviations of synthetic chemistries used. The bars indicate the  $\pm$ median absolute deviation. The dashed line is linear regression model fitted to indicate linear relationships.

Table 3.9: Astrocyte fibre length rank in culture environments. Ranks are calculated with Bray & Curtis dissimilarity on related cell variables from all 3 time points. The lower a chemistry's rank is to 0 the closer its cell performance is that of laminin's. Median absolute deviation is  $\pm 34 \mu\text{m}$ .

Environment	Median value ( $\mu\text{m}$ )	Rank
P/LAM	132.72	0.00
triNH <sub>2</sub>	134.02	0.06
diNH <sub>2</sub>	124.29	0.07
OH	143.35	0.08
CH <sub>3</sub>	145.01	0.08
NH <sub>2</sub>	146.19	0.08
Ph	148.44	0.08
SH	153.91	0.08
COOH	157.88	0.09



Best performers are diamine ( $\text{diNH}_2$ ), laminin, and triamine ( $\text{triNH}_2$ ). On the lower end are phenol (Ph), thiol (SH), and carboxylic acid (COOH). Visually, the astrocyte processes are not significantly different in the environments. Median absolute deviation for these measurements is  $\pm 168 \mu\text{m}$ .

Astrocyte fibre length (AFL) correlates with the logP group. The correlations at time-point day 7 are (+)strong and significant with logP2, logP4 and logP5 (Figure 3.21, Figure 3.22). In addition, there is a (+)strong correlation with molecular mass on day 3. A similar (+) strong correlation appears with molecular volume on day 3 but by day 7, this relationship changes to (–)moderate (Figure 3.23).

### 3.2.4 LogP correlations

The logP group is interesting as strong relationships appear for all cell variables. In this section, the interest is in understanding the depth of lipophobicity effect on cell performance in modified culture surfaces. The partition coefficient (logP) is a measure of compound solubility when placed in settled solutions that are incapable of mixing (immiscible) such octanol and water. LogP refers to the concentration ratio of un-ionised species of compound. In pharmaceutical sciences, this measure is useful in estimating the drug distribution in the body. Lipophilic drugs with high logP are administered to lipophilic areas such as the skin (289), gastrointestinal tract (290), and blood-brain barrier (291). Lipophobic drugs with low logP are administered in aqueous regions e.g. intravenously. The logP is defined experimentally as:

$$\log P_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{octanol}}^{\text{un-ionised}}}{[\text{solute}]_{\text{water}}^{\text{un-ionised}}} \right)$$

Equation 3.1: Calculating partition coefficient experimentally.

Nowadays, the logP is estimated computationally using a variety of methods such as atom-based and fragment-based, among others. Fragment methods are better suited for larger molecules compared to atomic methods (106,209). Atomic methods are more accurate for smaller molecules (106).

The logP appears in the literature in investigations for its effect on cell adhesion (292), attachment and spreading (77). In cell adhesion studies, the logP serves as one of the molecular descriptors (input) modelling embryoid body cell adhesion. It is one of the most relevant input to the author's predictive model as it can explain the outcome well (292). In cell spreading studies, osteoblast cell spreading was found to correlate with calculated logP with a (+)strong relationship ( $r = 0.88$ ) (77). These reports compelled the investigation of logP and its effect on neural stem cells.

Typically, in protein and cell studies the logP is estimated for a fragment of the surface chemistry with multiple constituents. In this project, the logP value granularity will be increased by using logP values for up to the top 6 constituents of the chemistry (Figure 3.24). To our knowledge, this is the first study investigating the logP in this fashion as a biological descriptor.

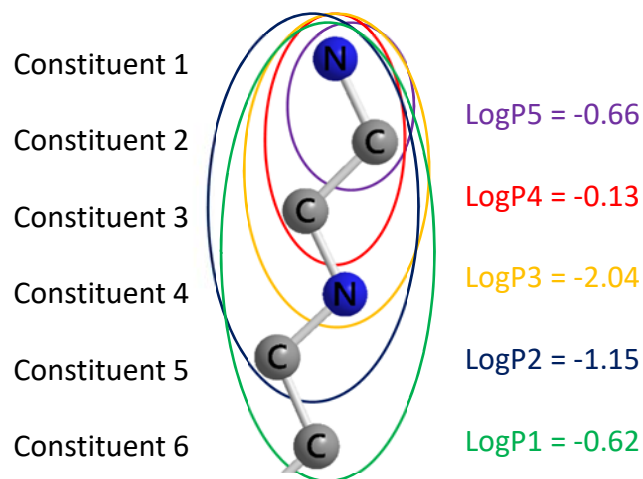


Figure 3.24: LogP values for molecule constituents. LogP5 value is the logP value for the terminal group. Moving down to logP1 being the logp value for up to the top 6 constituents of the molecule.

LogP values for synthetic chemistries used in this project were calculated for up to the top 6 constituents of the presenting chemistry. The logP calculation has been extensively compared among many others in (106,209) performing very well compared to real logP values. Evidence in the literature found cells sense up to 10 nm of surface characteristics (293,294). Accounting for the adsorbed protein layer, we believe this to be up to the top 6 constituents of the surface chemistry. This will be investigated further in a later chapter. Below are correlation count graphs followed by plots of individual correlations between logP of molecule constituents and cell responses. A short discussion for each cell variable group follows the graphs. Plots with the data providing these correlations are shown in previous sections within this chapter. Below is a figure with correlation frequencies:

### 3.2.4.1 Correlation frequency

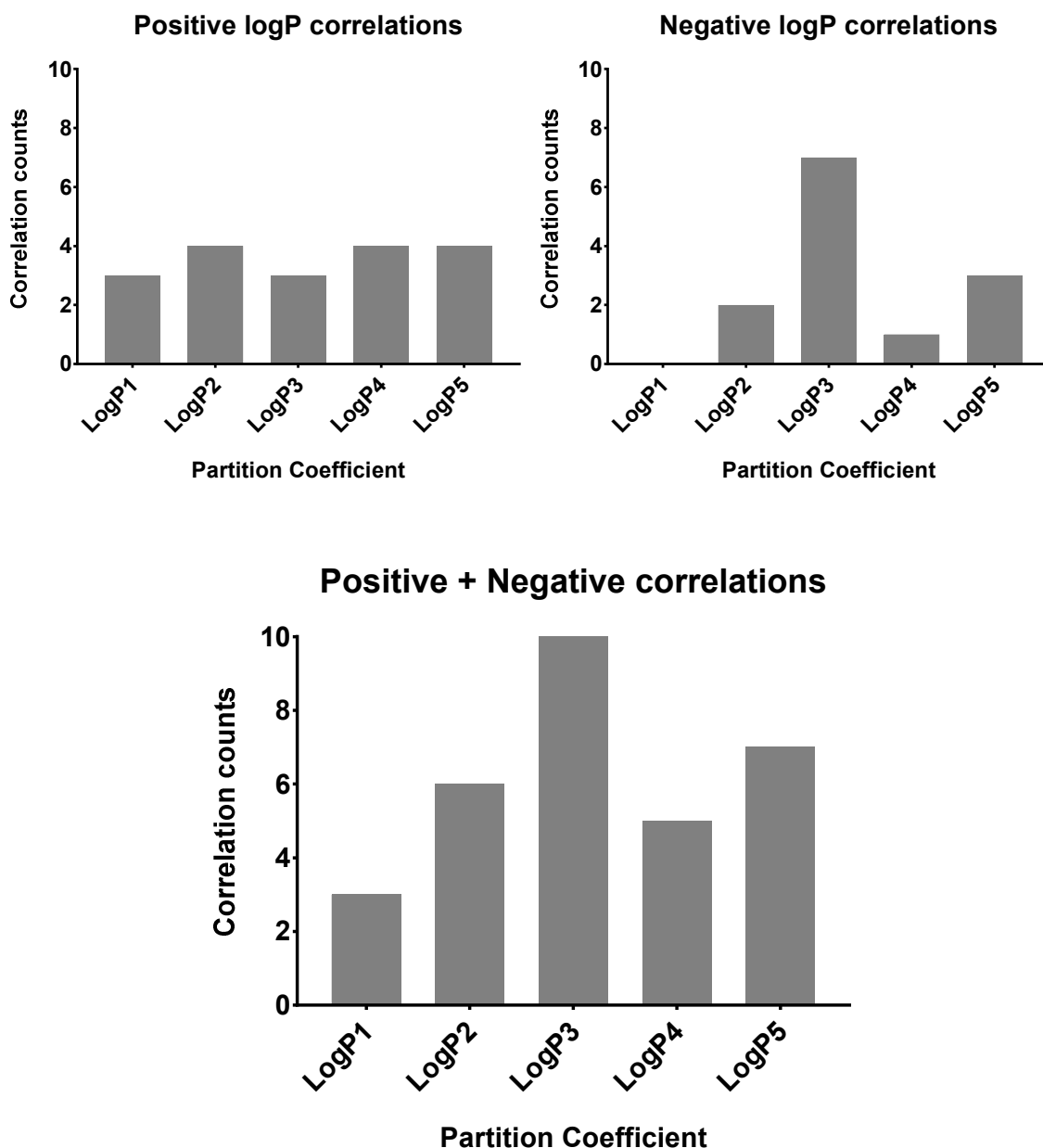


Figure 3.25: LogP significant correlation counts.  $x$  axis represents the chemical variables and the  $y$  axis represents correlation counts. Top left: positive correlation counts, top right: negative correlation counts, and bottom: total correlation counts.

Positive logP constituent correlation count is similar throughout the constituents whereas, logP3 and logP5 prevails for negative correlations. From the bottom graph, both positive and negative correlation counts were put together. The important parameters for logP are in this order  $LogP3 > LogP5 > LogP4 > LogP1$ . Below are graphs of correlations coefficients with their standard error for each cell variables of all 3 time points. Below is a

figure with significant correlations between cell cluster area and neuron density vs all logP parameters:

### 3.2.4.2 Cell cluster area and neuron density

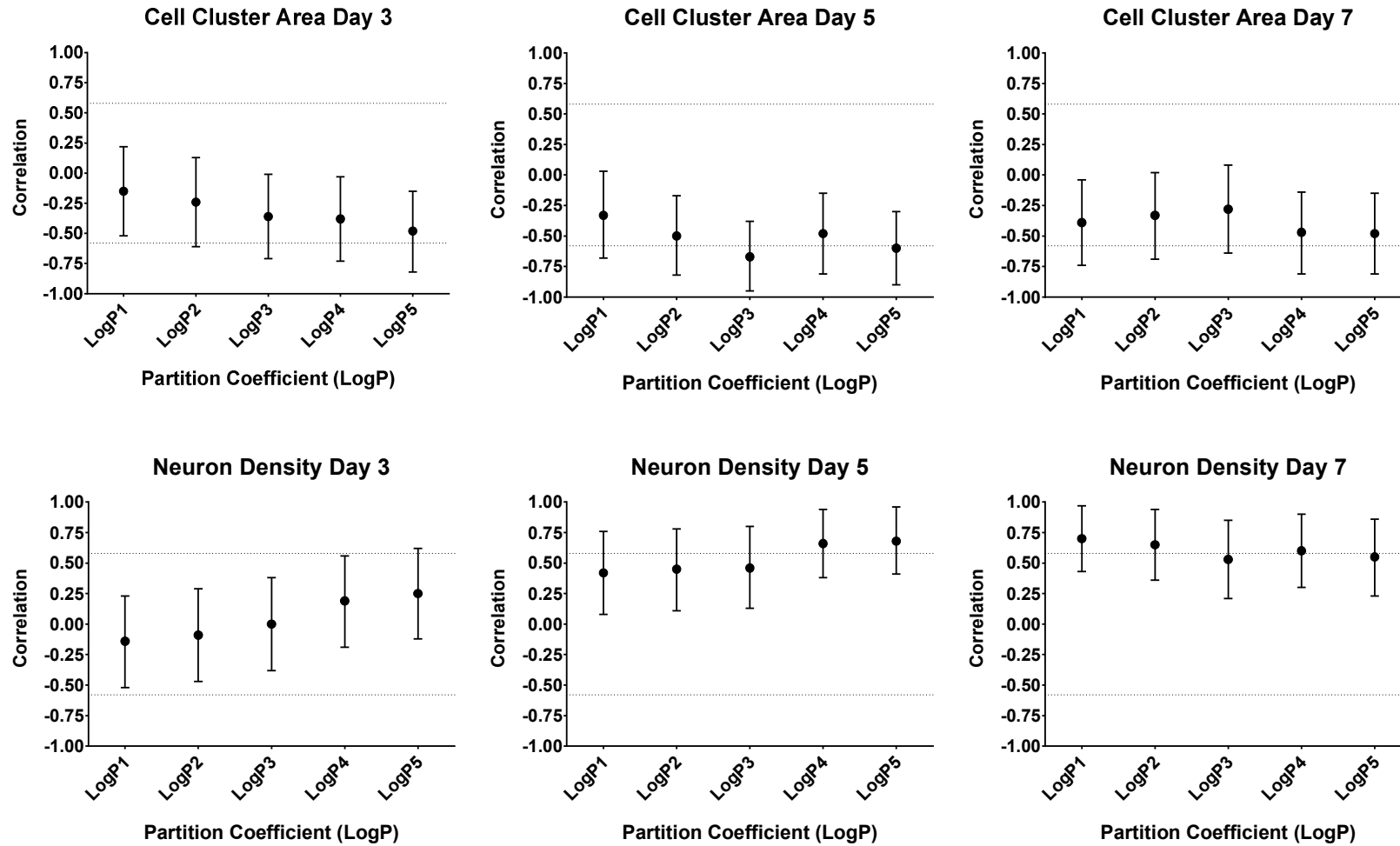


Figure 3.26: Correlation graphs. Cell cluster area and neuron density vs partition coefficient (logP) group. Bars are the standard error of correlations. Data points outside the thresholds indicated by dashed lines are significant. No significant differences between any correlations for any time point (Kruskal-Wallis test).

Cell cluster area on time point day 3 informs of cell differentiation, migration, and proliferation. In later time points, CCA informs on cell migration and proliferation (101). On time point 3, (–)correlations are observed and the relationship's strength increases to moderate as it moves to the terminal group. On day 5, (–)correlations appear ordered by strength  $r_{\log P3} = -0.67 > r_{\log P5} = -0.6$  both of which are significant ( $< -0.58$ ). On day 7, CCA correlates (–)moderately with  $\log P5 > \log P4$  and  $\log P3$ 's correlation changes to (–)weak.  $\log P5$  and  $\log P4$  are consistent in their relationship with CCA throughout the time points. The important constituents for the effect of lipophilicity on cell cluster area are  $\log P4 > \log P5 > \log P1 > \log P2$ . Here,  $\log P3$  needs additional evidence to support its effect.

Neuron density (ND) on time point 3 informs on neural differentiation. High density here means cells are densely packed. Low neural density is a strong indicator of cell differentiation. Day 5 and 7 time-points are good indicators of proliferation (101). On time point day 3, there are mostly very weak correlations in both directions (+) and (–). On day 5, (+)strong and significant correlations appear with  $\log P5 > \log P4$ . By day 7, ND correlates with the entire  $\log P$  group (+)strong. The order in correlation strength at day 7 is  $\log P1 > \log P2 > \log P4$ , all of which are significant.  $\log P5 > \log P3$  have (+)strong relationship although their correlation strength is just below the high critical value ( $< 0.58$ ) therefore not significant. Almost all constituents are important for the lipophilic effect on neuron density.  $\log P3$  needs additional evidence to support its effect. Below is a figure with significant correlations between astrocyte density and neuron/astrocyte ratio vs all  $\log P$  parameters:

### 3.2.4.3 Astrocyte density and neuron/astrocyte ratio

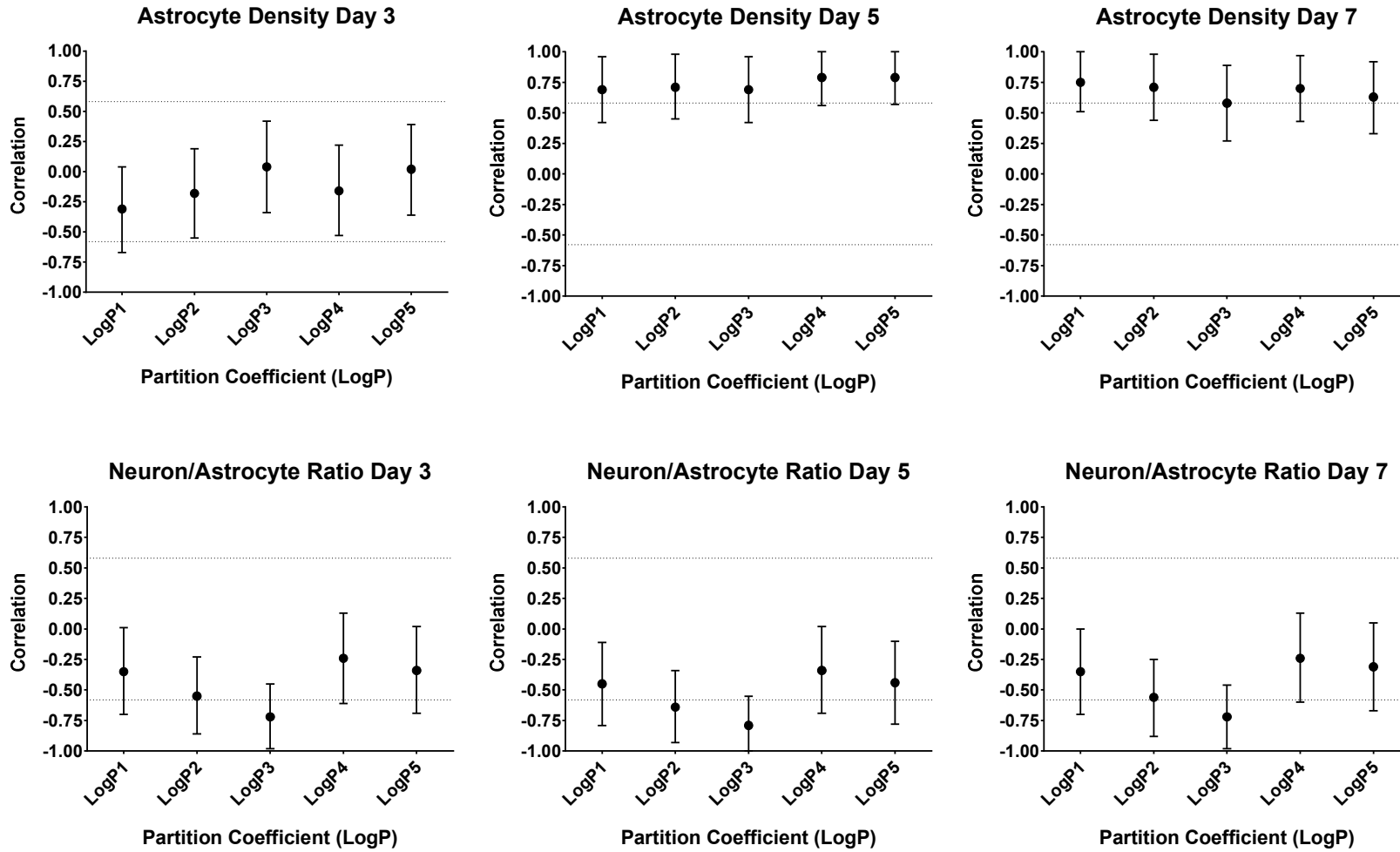


Figure 3.27: Correlation graphs. Astrocyte density and neuron/astrocyte ratio vs partition coefficient (logP) group. Bars are the standard error of correlations. Data points outside the thresholds indicated by dashed lines are significant. No significant differences between any correlations for any time point (Kruskal-Wallis test).



Astrocyte density at day 3 informs on differentiation. High density means cells interact more among themselves whereas low astrocyte density indicates differentiation and migration. Day 5 and 7 time-points are good indicators of proliferation (101). AD has (–)weak correlations with logP1, logP2, and logP4 on day 3. However, on day 5 and 7 (+)strong correlations appear and all of them are significant ( $r \geq 0.58$ ). This means all constituents contribute to the lipophobic effect on astrocyte density. The evidence of the relationship is strong since correlations appear on both time points related to proliferation. The order of correlations in strength for both time points is  $LogP4 > LogP1 > LogP2 > LogP5 > LogP3$ .

Controlling the proportion of neurons and the purity of the transplant population is a critical quality attribute for cell therapy translation (281). An imbalance in the proportion and migration of cells can have adverse effects for transplant recipients such as uncontrolled movement (282) and teratomas (91). Cell densities were used to calculate neuron proportion expressed as  $\frac{Neuron\ Density}{Astrocyte\ Density}$ . Low cell density coupled together with high neuron proportion is preferred. Neuron/astrocyte ratio (NAR) has a significant (–)strong correlation with the second constituent (logP2) of the surface chemistry on day 5. On day 3 and day 7, (–)strong correlations appear as well but are not significant ( $> -0.58$ ). LogP3 has the strongest correlations appearing in all time points. For NAR, the order of logP correlations in strength appearing is  $LogP3 > LogP2$ . Below is a figure with significant correlations between neuron axon and astrocyte fibre length vs all logP parameters:

### 3.2.4.4 Neuron axon and astrocyte fibre length

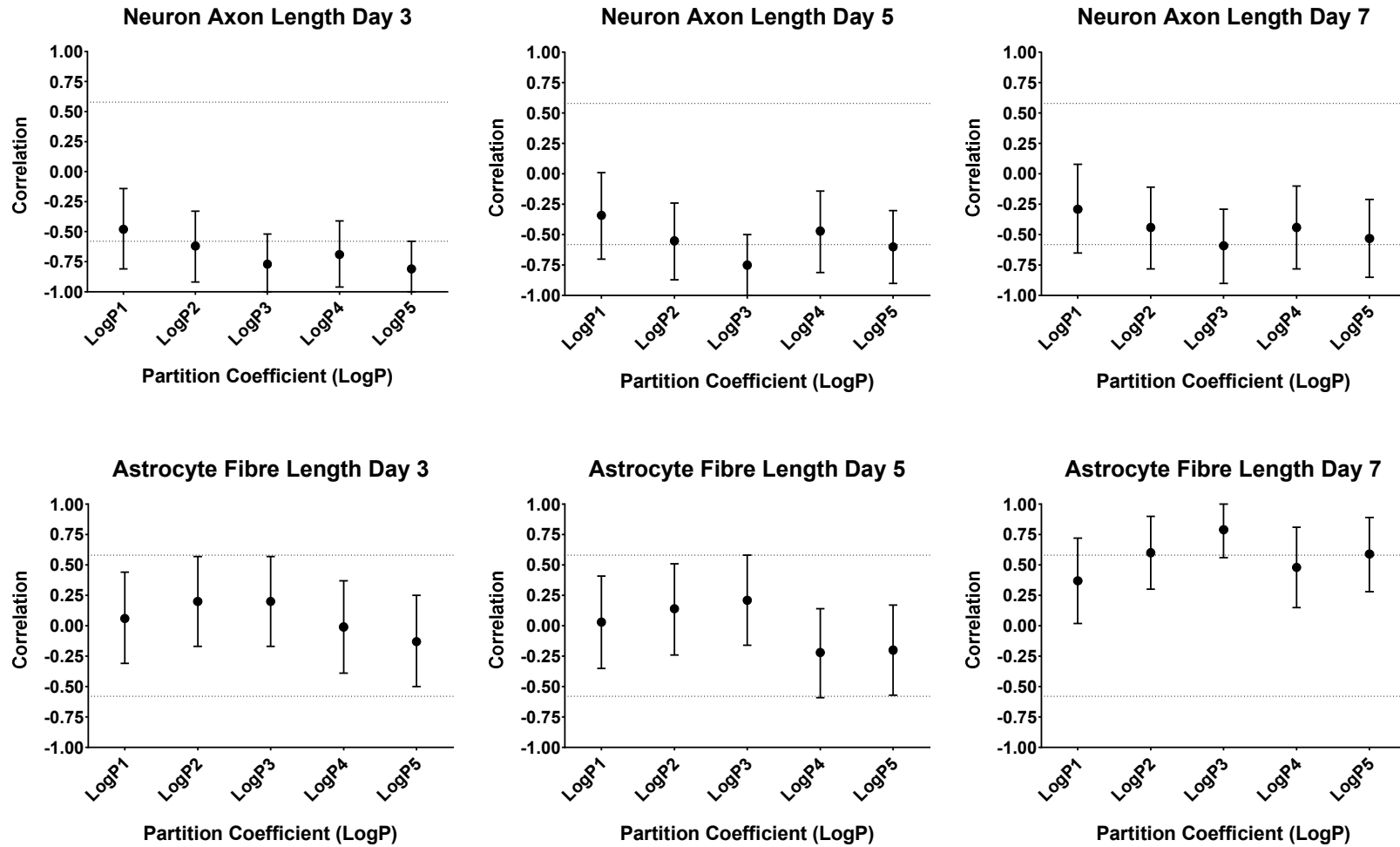


Figure 3.28: Correlation graphs. Neuron axon and astrocyte fibre length vs partition coefficient (logP) group. Bars are the standard error of correlations. Data points outside the thresholds indicated by dashed lines are significant. No significant differences between any correlations for any time point (Kruskal-Wallis test).

Functionary nerve tissue consists of neural projections to communicate with neighbouring cells using electrical conduction across large sections of tissue. Neuron axon length (NAL) is a good indicator of this *in vitro*. One aim of neuro-regenerative biomaterials is to grow and guide neurons to specific injury areas and re-wire compromised neural circuit to restore function. NAL correlates (–) strong with logP3 on all time points and significant ( $r < -0.58$ ). The next one is logP5 with (–)strong correlations on day 3 and 5 that are significant. On day 7, there is (–)strong relationship but this is not significant ( $r > 0.58$ ). LogP2 and logP4 both have a (–)strong relationship with NAL on day 3 that is significant and another (–)strong relationship on day 5 that is not significant. The important constituents for the lipophobic effect on NAL are  $LogP3 > LogP5 > LogP4 > LogP2$ , ordered by correlation strength.

Astrocyte fibres *in vitro* work with oligodendrocytes to align their processes with axons thereby controlling the onset of shielding their terminals (neuron-myelination) (288). It has been proposed that reactive astrocytes are linked to axon elongation in spinal axons as well (295). Myelination is an important attribute of oligodendrocytes (induced by astrocytes) for developing function neural circuits *in vitro*. For this project, astrocyte fibre length (AFL) is an indicator for neuron availability and migration. The shorter astrocyte processes the more likely neurons are within the vicinity. On time points day 3 and 5, there are mostly very weak correlations in both directions (+) and (–). Strong (+)correlations appear on day 7 with logP3, logP2, and logP5 all of which are significant. The important constituents for the lipophobic effect on AFL at day 7 are  $LogP3 > LogP2 > LogP5$ , ordered by correlation strength.

### 3.3 DISCUSSION

Here we investigated synthetic environments to control cell performance and relate it with the attributes of the biological environment. Cell performance data used in this chapter are from E12 Sprague-Dawley rat ventral mesencephalon chosen to maximise the potential of forming dopaminergic neurons (41,296).

#### 3.3.1 Cell cluster area (CCA)

The evidence in Figure 3.10 and The table below shows the cell cluster area ranks for all cell culture environments for comparison:

Table 3.4 and suggest that cell clusters spread more on lipophobic ( $\log P$  0.4 to -0.66) surfaces. There are some indications for advantageous cell cluster spreading on moderately hydrophilic to borderline hydrophobic surfaces ( $60-89^\circ$  WCA) but more evidence is required in support for this claim. A note the water contact angles obtained with the in-house instrument (OneAttension Theta Lite) are higher for all surfaces compared to those found in the literature. The chemical properties lipophobicity and wettability are not always inversely related. In fact, hydrophobic and lipophobic surfaces such as teflon support long-term neural cultures because they reduce evaporation and pathogen contamination (297). Teflon is essentially chained fluorinated carbons. The interesting part is that such materials are known to be non-fouling resisting protein adsorption (298).

It is known on hydrophobic surfaces ( $WCA > 90^\circ$ ), proteins adapt their secondary structure (conformation) to maximise interaction with their hydrophobic parts (hydrophobic effect of protein adsorption) (95,116). This means irreversible protein adsorption of smaller non-adhesive proteins (299), such as the abundant albumin, leaving

less potential to displace these with adhesion-mediating proteins, such as vitronectin and fibronectin (300). Neuronal survival and differentiation depends on cell adhesion. No adhesion means cell death and a study found cells will die within 2 days if this is the case (301). In lipophobic and hydrophobic environments, the hydrophobic effect on protein adsorption is inhibited. Adsorption strength is weakened although protein denaturation is inhibited by a large degree (300,302). As a result, the exchange of non-adhesive serum-proteins for adhesion-mediating proteins and the accessibility of adhesion sites for the cells integrins is significantly improved. Initial cell adhesion is comparable to that observed on more wettable surfaces although it is slightly retarded (300).

On the other hand, in lipophilic and super-hydrophobic environments remarkable attachment of mesenchymal stem cell (MSC) has been observed but the effect disappeared after 3 days in culture (303). Six hours after seeding, they also found cell adhesion was not significantly improved on hydrophobic surfaces. After the 24-hour mark, the effect was strong and apparent. Reports indicate the temporarily enhanced cell adhesion likely occurs due to hydrocarbon chains of self-assembled monolayer (SAMs) interacting with cell membranes (304). Adsorption experiments with extra-cellular matrix (ECM) proteins (type I collagen, fibronectin, laminin) showed that specific adsorption of these proteins on methyl-terminated self-assembled monolayers cannot account for this temporary effect.

### 3.3.2 Neuron density (ND)

The evidence in Figure 3.11 to Figure 3.13 suggest neuron density (ND) is lower in lipophobic ( $\log P$  0.1 to -0.66) and moderately hydrophilic/borderline hydrophobic surfaces with water contact angle ( $60^\circ$  to  $89^\circ$  WCA). In hydrophilic surfaces, lower ND could be attributed more to cell death instead of migration. This is speculated as cells on hydroxyl

(OH) and carboxylic acid (COOH) had increased type I astrocyte spreading, they were larger and more elongated which are signs of stress. A similar finding was recorded in (305) with embryonic cortical neurons on hydrophilic surfaces.

Generally, hydrophilic surfaces offer enhanced cell adhesion (306) and cell spreading (307). We know this as moderately hydrophilic surfaces (40°-60° WCA) were found to be conducive to protein adsorption (95,299,308) and cell adhesion (107). There are many examples supporting this finding but there are some exceptions as well. In a study, moderately hydrophilic surfaces (58° to 74° WCA) were made with “sticky” (PEI or PDL) compounds coated with laminin or fibronectin proteins (301). In these, cells attached well and neuron processes emerged from 2 days in culture. That is 15° WCA above the previously decided boundary. On the other hand, hydrophobic surfaces (108° WCA) had no neuronal adhesion and they died after 2 days (301).

In another study (309) poly(l-lactic acid) (PLLA) nanofiber and film surfaces were made hydrophilic with plasma etching then coated with polylysine for cell attachment (39° and 46° WCA). Although within the moderately wettable range, these worsened motor neuron survival when seeded in low cell density (50 cells/mm<sup>2</sup>) compared to plasma untreated PLLA surfaces (43°, 58° and 68° WCA).

Protein studies found increased adsorption attributed to WCA does not necessarily mean increased interaction (“activity”) with cells for cell adhesion (94,310). Clearly, there is more to the story than wettability alone as the main chemical property controlling biological responses. Studies and reviews are emerging stating this fact after extensive investigations involving 20000 samples (311,312).

### 3.3.3 Astrocyte density (AD)

From the evidence in Figure 3.14 to Figure 3.16, astrocyte density (AD) decreases in lipophobic surfaces (logP 0.1 to -0.66). In 2D *in vitro* neural cultures, astrocytes are below neurons cushioning them (305). Following the same trend as neurons, they migrate better on lipophobic surfaces. In the literature, there are similar findings where glial cells preferentially attached and proliferated on hydrophilic surfaces (45° to 60° WCA) instead of hydrophobic environments (90° to 108° WCA).

Regarding the (–)correlation of AD with acid dissociation constant (pKa), this was only for time point day 3. This suggests acidic surfaces will lower AD but this is likely due to cell death rather than migration. With correlations, causality is not implied and there is no trend in other time points. There is no correlation on day 5 and by day 7, the correlation changes to (+)moderate. What can be said from the best performers in lowering AD (Figure 3.17), molecules with a terminal group around 8-11 pKa perform better. This pKa range is similar to what appears in the drug discovery literature (282). On more acidic surfaces such as hydroxyl (OH, pKa 4.5) there is more dehydration and less potential for hydrogen bonding. One would expect less competition with water for protein adsorption, however hydrogen bonding is one of four processes fundamental to proteins adsorbing (115). On these surfaces, protein degradation and limited protein adsorption is expected therefore less binding sites for cells limiting cell attachment, differentiation, and viability.

Molecules with low mass and volume (< 33 Da and 45 Å) (e.g. methyl, CH<sub>3</sub> and hydroxyl, OH) and/or more acidic (OH) do not perform well in lowering AD by day 3 and limit future astrocyte migration. It is possible cells interact with silanol groups from glass and this is not an ideal environment for cell culture likely due to the concentration of electronegative

charge. When the surface charge is distributed, desirable cell responses are observed. For example, the carboxylic acid (COOH) surface has a pKa of 4.87 and has 8 constituents in its backbone chain whereas, hydroxyl (OH) has pKa value of 4.5 and 1 constituent. The former performs better in lowering astrocyte density although their terminal groups are almost equally acidic. This adds to the evidence of the importance of the terminal group chemistry and the depth of the chemistry allowing more control over cell responses.

Regarding the correlation with astrocyte density (AD) and molecular volume, thiol (SH) and phenol (Ph) are changing the relationship from negative to positive on day 5 and 7. Negative relationships as shown on day 3 were expected for all time points. This means as molecular volume increases, astrocyte density decreases. In self-assembly molecules (SAMs), increasing molecular volume (and mass) means adding side chains on any molecule constituent in the backbone or increasing SAM chain length. The former will also change SAM packing density, which is out of scope for this project so the latter is assumed.

From our results, we know the lower boundary (3 constituents) and the literature has set the upper one. SAM chain length was investigated with milk allergen protein binding ( $\beta$ -lactoglobulin and apo-transferrin) (313). These proteins were bound by activating their acid groups and there was higher protein binding on surfaces with shorter amine SAMs (4 constituents) compared to longer (8 constituents). Others found similar outcomes with long-chain SAMs resisting protein adsorption and cell adhesion with long chains of a non-fouling material (PEG, 11-13 constituents) (314,315). The authors proposed this occurs due to the enhancement of the repulsive interaction forces minimising protein aggregation (steric stabilisation) (134).



### 3.3.4 Neuron/Astrocyte ratio (NAR)

The evidence in Figure 3.18 and Table 3.7 suggest lipophobic ( $\log P$  -1 to -2.3) surfaces will provide higher neuron to astrocyte ratio. There are indications hydrophilic (WCA  $60^\circ$  to  $70^\circ$ ) surfaces with reduced terminal group acidity ( $pK_a$  8-11) may help increasing neuron proportion but these findings require additional evidence. There are larger populations of neurons in hydrophilic environments. This is proof that neuron density decreased due to cell migration. Our previous hypothesis of lower neuron density attributed to cell death is now proven not to be true.

Contrary, in related work, embryonic stem cells from mouse and humans differentiated to neurons 2.4 and 1.6 fold respectively on hydrophobic PDMS surfaces ( $111^\circ$  WCA) compared to the neutral ultra-low-attachment plates (LAC,  $23^\circ$  WCA) (316). In another study, carbon nanotubes (CNTs) were treated with nitric acid to turn them more hydrophilic ( $<90^\circ$  WCA). The authors did not provide exact wettability measures for the surfaces used in their study. On these surfaces, laminin adsorption, cell adhesion, and neuron differentiation were enhanced compared to the typical standard surface (poly-L-ornithine) commonly used for neuron culture (317). Here, the hydrophobicity of CNTs was modulated after acid treatment and their topographical effect better mimics the extracellular matrix providing enhanced cell responses.

Another contradiction with our findings is the hydrophilic carboxylic acid (COOH) groups found to be negative cues for neuron differentiation (317,318). This interpretation depends on the application. If maximising neuron population is desirable then yes, COOH surfaces are not great compared to diamine ( $diNH_2$ ) and triamine ( $triNH_2$ ). From our findings (Table 3.7), COOH's neuron differentiation potential is close to that of laminin's.

Regarding the acid dissociation constant (pKa) correlation, relevant work contradicts our findings with good reason. Carbon nanotube (CNTs) scaffolds offer large surfaces and after grafting them with a compound (4.65 pKa) called poly(methacrylic acid), they present an ideal moiety for protein adsorption. These modified CNTs, “cage” proteins and capture growth factors from the cells’ immediate microenvironment. In effect, this regulates cell behaviour and enhance differentiation of human embryonic stem cell into neuronal cells (313,319). Our culture environments do not possess this feature as they are flat coverslips with a much smaller topographical effect. The high ordered structure of self-assembly molecules provide a better method to assess the effect of the chemistry on cell behaviour.

### 3.3.5 Neuron axon length (NAL)

The evidence in Figure 3.19 to Table 3.8 suggest lipophobic (logP -1 to -2.3) chemistries will maximise NAL possibly due to enhanced cell migration previously discovered. This finding is expected as earlier it was discovered cell clusters are larger and neurons migrate to a larger degree on lipophobic surfaces. The hypothesis is neuron axons will have to reach out further for other neurons to synapse. Regardless of the hypothesis, the literature agrees with the finding. In a study, neurite formation of rat noncancerous tumour cells was studied on polymer surfaces with a wettability gradient. Neurite volume and length increased on the gradient with moderate hydrophilicity (55° WCA) instead of more hydrophobic or hydrophilic areas (320).

In another study, the growth and axon length of hippocampal neurons was investigated on surfaces with different materials (321). The materials’ wettability ranged from hydrophobic (110° WCA) to hydrophilic (35° WCA). All of these were coated with poly-L-lysine (PLL) to

create membranes with the same functional groups interacting with cells. As a result, their wettability changed to moderately hydrophilic (64° WCA). Axon length was high in smoother membranes (175 µm) compared to rougher surfaces with 3x shorter axons (55 µm). The author suggests smooth membranes modulate the development process of neurons hence the longer neurites. The surface roughness may influence cell motility, or hinder the extension and ramification of neuronal processes emerging from the cell soma. Perhaps the surface roughness guides the adsorption of adhesion proteins necessary for the interaction with membrane surfaces (321).

One of the surfaces providing longer axons is fluorocarbon with PLL coating on top (321). Surfaces with fluorocarbon are lipophobic and hydrophobic. Such surfaces have been successfully used in long-term neural cultures as they reduce evaporation and pathogen contamination (297). From previous findings (section 3.2.3.3), a relationship was discovered with cell cluster areas increasing in lipophobic and moderately hydrophilic/borderline hydrophobic surfaces. This suggests such surfaces with this configuration are interesting to investigate further. This agrees with our own and findings from related work (95,299,308) that cell migration increases on moderately hydrophilic surfaces. There are exceptions in the literature with examples of good cell migration on more hydrophilic surfaces (<60° WCA) (301,309).

### 3.3.6 Astrocyte fibre length (AFL)

The evidence in Figure 3.21 to Figure 3.23 suggests lipophobic (logP -0.6 to -2) environments will decrease astrocyte fibre length. This finding is expected as cell cluster area is larger, neuron and astrocyte densities are lower in lipophobic surfaces. This means cells migrate further away in such environments and astrocytes will reach out further for

other cells. There are few studies investigating surface wettability with astrocyte fibre length (78,322). Several studies have demonstrated preferential cell adhesion (306) and cell spreading (307) on hydrophilic surfaces. In a study, cell spreading was investigated on thin film polymers (323). The films were made of poly(acrylonitrile-vinylchloride) (PAN-PVC) and are moderately hydrophilic (76° WCA). The investigators confirmed the reports of others (324–326) that cell migration rate is reciprocally related with cell spreading. Astrocytes displayed significantly lower migration rate ( $15\pm 4 \mu\text{m/hr}$ ) relative to meningeal cells ( $42\pm 5 \mu\text{m/hr}$ ). In addition, astrocyte spread cell area and processes were significantly greater ( $4250\pm 1000 \mu\text{m}^2$ ) compared to meningeal cells ( $2000\pm 300 \mu\text{m}^2$ ).

The author attributed these astrocytic responses to differential expression of integrin receptors among different cell types. This makes sense as integrin expression has been shown to vary among different cell populations within the central nervous system (327). Essentially, astrocytes were tightly bound on the surface and there was more effort required to migrate hence the large cell spreading. We found astrocyte fibres are larger in surfaces with limited cell migration (phenol, Ph and thiol, SH) (Table 3.5, Table 3.6). The opposite is observed as well - astrocyte fibres are generally smaller in surfaces with good cell migration (amines and laminin).

There is evidence that chemistries with molecular mass 62-90 da and molecular volume 76-105 Å in their untethered state will do better in lowering astrocyte fibre length (AFL). In self-assembly molecules (SAMs), increasing molecular volume (and mass) means adding side chains in any molecule constituent in the backbone or increasing SAM chain length. The former will also change SAM packing density, which is out of the scope of this project so the latter is assumed. In a related study, differences in the chain length of molecules used to prepare model surfaces directly influences cell attachment and cell spreading (328).

From our results, we found the minimum chain length for self-assembly molecules (SAMs) to minimise AFL is 3 constituents. The upper boundary was found to be around 8 constituents found by protein adsorption and cell adhesion studies (134,313–315).

### 3.4 NOVELTY

- 1) Cell clusters spread more on lipophobic (logP 0.4 to -0.66) surfaces
- 2) Neuron density is lower in lipophobic (logP 0.1 to -0.66) and moderately hydrophilic/borderline hydrophobic surfaces with water contact angle (60° to 89° WCA)
- 3) Astrocyte (AD) density decreases in lipophobic surfaces (logP 0.1 to -0.66)
- 4) Acidic surfaces will lower AD but this is likely due to cell death rather than migration
- 5) Molecules with low mass and volume (< 33 Da and 45 Å) such as methyl (CH<sub>3</sub>) and hydroxyl (OH) and/or more acidic (OH) do not perform well in lowering AD by day 3 and limit future astrocyte migration
- 6) Lipophobic (logP -1 to -2.3) surfaces will provide higher neuron to astrocyte ratio. There are indications hydrophilic (WCA 60° to 70°) surfaces with reduced terminal group acidity (pKa 8-11) may help increasing neuron proportion
- 7) Neural stem cells on carboxylic acid terminated surfaces differentiate to the same degree as on laminin surfaces
- 8) Lipophobic (logP -1 to -2.3) chemistries increases neuron axon length likely due to reduced cell migration
- 9) In lipophobic (logP -0.6 to -2) environments, decreased astrocyte fibre length is observed
- 10) Chemistries with molecular mass 62-90 da and molecular volume 76-105 Å (in their untethered state) will lower astrocyte fibre length (AFL)

## 4 DESCRIBING RELATIONSHIPS COMPUTATIONALLY

---

### 4.1 INTRODUCTION

The relationship between the surface-cell has not been described yet. This is because the relationship is multi-dimensional (multiple inputs) and traditional experimental methodologies are limited to a few inputs (311,312). The main aim of this chapter is to describe the relationship between the surface-cell in the form of computational model(s) where the chemical inputs are in multiple dimensions. After testing these model(s) for their predictive “goodness”, these will be used to screen future chemical designs without performing time consuming and costly cell culture experiments. In this way, animal use is reduced (3R's).

The idea is to feed chemical designs in numerical form into predictive models and these will provide cell performance estimates. The chemical parameters we chose to investigate are partition coefficient (lipophilicity), acid dissociation constant (acidity), molecular volume and mass. The values of these parameters define chemical designs. Surface topography and stiffness are two more families of surface properties investigated in the past (90) but they fall out of scope for this project. The cell performance measures are morphological changes in cells such as cell cluster and cell spreading, cell type proportion, and cell projection elongation.

This chapter is about computationally modelling the relationship between the chemical parameters and cell performance. The models are found using machine learning techniques or more precisely supervised learning where the data were labelled with a header (column name). Computational models are frequently used to support the understanding of complex systems and optimise industrial processes in engineering and

physical sciences (329). A computational model is a set of equations that describe how a system changes as a function of some variable, such as time. They involve variables representing things that change over time and parameters where their values are static or change on longer time scales (330).

Machine learning is the automation of building analytical models using algorithms that iteratively learn from previous data without being explicitly programmed (331). Machine learning evolved from the study of computational learning theory and pattern recognition. The idea here is to construct algorithms that can learn from data and make predictions without following static instructions. They make data-driven decisions or predictions by building a model from sample inputs. Supervised learning is the machine-learning task of inferring a function from training data that we know what they are and label them with a header (column name). The training data consist of the training examples and each example is a pair of an input object and an output value. A supervised learning algorithm analyses the training data and produces an inferred function used to map new examples. The ideal situation is where the learnt algorithm determined the class labels for new examples and this is referred to as generalisation from training data.

The objectives for this chapter include:

- To discover chemical designs to modify cell culture surfaces. Inspiration comes from the literature, laboratory experiments and previous work (41)
- To perform additional cell culture experiments for more data to establish a stronger analytical foundation
- To computationally model cell performance as a function of cell environment properties using machine learning techniques

- To perform sensitivity analysis for models to understand the importance each of their inputs play in the estimation of the outputs.

Figure 4.1 below shows the above in a logical order. Steps 5 and 6 belong in the next experimental chapter. The criteria for stopping the above cycle is the discovery of synthetic environments whose cell performance is better than our current best synthetic environment, amine (NH<sub>2</sub>). Better cell performance means “closer” to that of biological environments such as our gold standard, laminin.

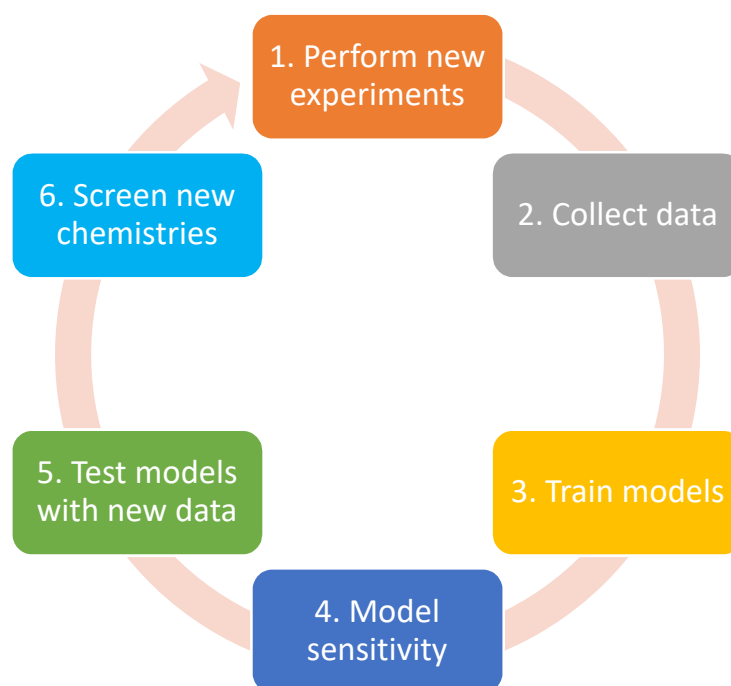


Figure 4.1: Workflow to discover better synthetic environments for nervous tissue engineering using data science methods and techniques. The process starts from performing experiments.

#### 4.1.1 Model performance and cross-validation

Model performance is a model’s predictive ability expressed through an error measure of how “wrong” the predictions are compared to real values. A good way to obtain this error measure is by using a trained model on a test dataset made with data that have not been used to train the model. Training data will be “sacrificed” and this could be a problem if there are not enough data to work with. On the other hand, more testing data is better for robustness on predictive performance on future examples. The model error on the



training set (re-substitution error) is calculated by resubstituting training instances into a classifier that was constructed from them. It is not a reliable predictor of the true error on new data, but it is useful to know.

A better way to assess model performance is through cross-validation (CV). CV maximises data used for training and testing by splitting data in  $k$ -parts then fit  $k$ -models. In each iteration, training data is made up of  $k - 1$  parts and the remaining part is used for testing. The predictive performance here is the average for all  $k$ -iterations. CV performance is still better than the training model performance.

Modelling with cross-validation usually means training data are shuffled for good reasons. This is undesirable for our purposes as the data representing each environment is in 9 parts (instances). This was performed to retain as much cell output variation as possible without data sparsity. This maximises the amount of data used for machine learning but also means the data need to stay in their respective group otherwise we risk additional data leak in model validation during training. Each cross-validation part is forced to be cell data from 1 environment but the order of these was randomised in each cross-validation iteration.

Prediction error is the sum from three terms, irreducible error, bias, and variance. Irreducible error results from noise from the problem itself and it owes its name to the fact there is nothing to be done about it. Bias is the error from violated assumptions made by the learning algorithm. High bias means the algorithm will miss the relevant relations between features and target (underfitting) e.g. shallow decision trees, low-order/linear regression polynomials. The other source of error is variance, and this is the sensitivity to small fluctuations in the training set. Here, high variance means trained models will be fit on random noise of the training data rather than the intended outputs (overfitting) e.g.

deep decision trees, high-order regression polynomials. The goal is to capture the regularities in the training data (no underfitting) and simultaneously generalise well in unseen data (no overfitting). The reason there is a trade-off is because underfitting is the inverse of overfitting and the balance between the two is usually preferred. In cross-validation, the average error is a measure of bias and the standard deviation of errors is the variance.

In practice, there is no analytical way to find the “sweet-spot” of the bias-variance trade-off. Instead, we must use a suitable measure of prediction error, explore differing levels of model complexity (features), and then choose the complexity level that minimises the overall error. The key here lies in the selection of an accurate error measure as often grossly inaccurate measures are used and these can be deceptive. For this project, the mean absolute error (MAE) is selected as the measure of prediction error. This is the mean of the absolute difference between real values and estimates. On its own, the MEA cannot tell us if the error is acceptable or not for our cause. Here, we will inject domain knowledge and define the boundary of “acceptable” error as 1 standard deviation. This new performance metric is coined as the model performance ratio (MPR) and is the absolute difference between predictions and real values standardised by the average standard deviation. A ratio of 1 means the prediction is outside the natural bounds of cell performance and classed as unacceptable. We prefer a ratio closer to 0 meaning the “closer” the prediction is to real values.

#### 4.1.2 Linear regression

Below is the classic approach in modelling for regression problems, linear regression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$$

Equation 4.1: Cell cluster area modelling with linear regression.  $\hat{y}$  is the estimate response,  $\hat{\beta}_1 \dots \hat{\beta}_p$  are estimated using least squares minimisation and  $x_1 \dots x_p$  are the predictors (input variables).

Since we know the values of the input variables, the coefficients  $\hat{\beta}_1 \dots \hat{\beta}_p$  are to be estimated by choosing values to these where the model minimises the  $\sum residuals^2$  (least squares method). The coefficients determine the slope of the model fit and tell us the “importance” of each predictor. The closer to zero the coefficient is the less important the predictor. In other words, the predictor cannot explain the response in linear regression.

The goal is to find a function of  $y$  with respect to  $x$  using a smaller sample of data where this is as similar, or ideally, identical to the function fit on the population data. Obtaining the population data is almost impossible but there is a way to measure the model fit “variation”. The standard error of the coefficients tells us how much sampling variation there is if we were to re-sample and re-estimate the coefficients. It is calculated with:

$$\widehat{se}_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}} \quad \text{where:} \quad \hat{\sigma}^2 = \frac{1}{n-2} \sum_i \hat{\epsilon}_i^2$$

Equation 4.2: Coefficient standard error and standard deviation.

We can use a t-test to evaluate a sample regression coefficient in relation to its standard error (332). The t-test here is  $t = \frac{b_{yx} - \beta_{yx}}{\hat{\sigma}_b}$ , where  $\beta_{yx}$  is equal to 0. The t-test is the ratio of the sample regression coefficient to its standard error. This ratio shows the probability of the regression coefficient not being 0 if the  $p$  value is equal to or smaller than 0.05. This means there is 5% or lower chance accepting a false positive. This test can also show signs of collinear predictors. Linear regression assumes the opposite, that there is little to no correlation between your input variables (multi/collinearity). Another method to detect collinearity is a model producing very high  $R^2$  but most of the coefficients are insignificant

according to the  $p$  values of the t-test. This test is not definitive but it is an indication of collinearity.

#### 4.1.2.1 Collinearity

The simplest method to begin with regression problems is linear regression. To use this method, certain assumptions are made. Classical Linear Regression Model assumes there is no exact collinearity between explanatory (predictor/input) variables. If the predictor variables are perfectly correlated then regression coefficients become indeterminate making model interpretation difficult. This is because changes to the data will produce wildly different coefficients.

Perfect or near-perfect collinearity will return a singular or near-singular matrix with a determinant of zero. This means the rows or columns of this matrix are proportionally interrelated. In other words, one or more of its rows/columns is expressible as a linear combination of all or some of rows/columns with the combination being without a constant term. This can cause statistical analysis issues such as:

- Coefficient estimates become less certain and more variable as training data changes
- Prediction intervals are much wider therefore the hypothesis the true coefficient is zero cannot be rejected
- Prediction estimates are not affected and the  $R^2$  can still be very high

#### 4.1.2.2 Variance inflation factor

Variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance

(standard deviation<sup>2</sup> of the estimate) of an estimated regression coefficient is increased because of collinearity. The magnitude of multicollinearity can be analysed by considering the size of the  $VIF_{\hat{\beta}_i}$ . It is calculated by taking the ratio of the variance of all model's coefficients divide by the variance of a single coefficient if it were fit alone. VIF is the reciprocal of tolerance  $1/Tolerance$  where tolerance is  $(1 - R_i^2)$  and this represents the proportion of variance in the  $i$ th independent variable that is not related to other independent variables in the model.

The  $\sqrt[2]{VIF}$  indicates how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model. For example, if the variance inflation factor of a predictor variable were 10.54 ( $\sqrt{10.54} = 3.2$ ) this means that the standard error for the coefficient of that predictor variable is 3.2 times as large as it would be if that predictor variable were uncorrelated with the other predictor variables. The rule of thumb for the maximum values of VIF range from 4 to 40 but these are set arbitrarily and no best threshold exists (333).

### 4.1.3 Chapter related literature

Relevant applications of data mining and machine learning appears in the literature from cell migration and adhesion, gene expression profiling, antifungal solution discovery and cancer diagnostics.

An example of machine learning (ML) work is the computational model of cell migration in three-dimensional matrices (173). The authors used a force-based dynamics approach. The model determines overall locomotion velocity vector for speed and direction for individual cells based on internally generated forces transmitted into external traction forces. The

model also considers timescales where multiple attachment and detachment events are integrated. Model predictions agreed well with experimental findings for both 2D substrata and 3D natural tissues and synthetic gels. The logP appears in the literature in investigations for its effect on cell adhesion (292), attachment and spreading (77). In cell adhesion studies, the logP serves as one of the molecular descriptors (input) modelling embryoid body cell adhesion. It is one of the most relevant inputs to the author's predictive model as it can explain the outcome well (292).

In another work (334), different learning algorithms were evaluated for classification and prediction of antifungal peptides for use in medicine and agriculture. Antifungal peptides are safer and more effective drug candidates against fungal threats. Using computational techniques, the authors overcame costly and time-consuming screening new peptides. Support vector machines and bagged decision tree (C4.5) had the higher performance among other classifiers. Model performance measures were above 80% and for the authors this was acceptable for deployment to screen new antifungal peptides. The authors did not specify which model performance metrics they have used explicitly but they did mention accuracy in model validation.

Remaining articles mentioned below used a machine learning technique called artificial neural networks (ANNs). These are computer-based algorithms which are modelled on the structure and behaviour of neurons in the human brain and can be taught to recognise and categorise complex patterns (335). Pattern recognition is achieved by adjusting the parameters of ANNs in the process of minimising prediction error through a process resembling learning from experience. ANNs can be adapted to use any type of input data and the number of output categories can be specified.

Artificial neural networks (ANNs) among others methods, were used to optimise physical conditions of bacterial cultures (336). The fermentation conditions serving as the inputs to the model are multiple and these are pH, temperature, and inoculum volume (biological material that triggers immune response). The output is enzyme production (protease) as an indicator of fermentation. Fermentation produces organic acids, gases, or alcohol serving as bacterial energy. ANN with radial basis function network was chosen as this has a feed-forwards structure excelling in function optimisation for bioprocesses (336). Among other methods used, ANNs had better model fit and accuracy as it can represent nonlinearities better for this optimisation problem. The authors found a redundant input in the process. Inoculum volume did not explain protease production well, despite its strong presence in the literature and other relevant work.

Cancer classification often presents diagnostic dilemmas in clinical practice. It is believed that the answer lies in cancer gene expression. Standard practice is limited to the detection of single gene expression (immunohistochemistry) and thousands of genes are in play, typically. The authors collected 6567 gene data from 91 samples. Molecular techniques able to handle more genes such as RT-PCR sometimes provide non-definitive diagnosis. Artificial neural networks (ANNs) were presented with multi-dimensional gene expression data from round blue-cell tumours (337). This tumour data is classed in four distinct diagnostic categories serving as the model's output. The ANNs correctly classified all samples and identified genes that are most relevant to the classification. The model has been tested with data not used in the training procedure and all cases were correctly classified. Correct diagnosis of cancers literally means saving lives and this work did not stop there; potential genes as targets for therapy were also discovered.

Below are the experimental results for this chapter. The next section is the chemical characterisation for the cell culture environments used in experiments. After that, follow cell images and graphs of morphological cell performance. The section after that is about the capturing the relationship between surface-cell in the form of computational models. Sensitivity analysis of the chemical inputs for the cell models is investigated followed by a discussion and conclusions for this chapter.

## 4.2 RESULTS

### 4.2.1 Surface characterisation

Surface characterisation is a collection of techniques used to verify the presenting chemistry of modified surfaces. Some instruments and techniques are better suited for certain bonds and atoms defining the chemistries of these surfaces therefore 2-3 methods are necessary.

A list of self-assembly molecules is shown in Table 8.5. Methyl ( $\text{CH}_3$ ), carboxyl ( $\text{COOH}$ ), amine ( $\text{NH}_2$ ), hydroxyl ( $\text{OH}$ ), phenyl ( $\text{Ph}$ ), thiol ( $\text{SH}$ ) have been chosen as a starting point as these surface chemistries appeared in different literature (82,83,264,323,338–340). Adding to this list aminohexyl ( $\text{I-diNH}_2$ ), butylamine ( $\text{butylNH}_2$ ), propamine ( $\text{propylNH}_2$ ), 3-methoxy, and carbomethoxy (CBM). These have amine and oxy groups lower down the backbone of self-assembly molecules and they were used to investigate the cell sensing depth (341,342). The final synthetic surface chemistry terminates with a nitrile group and this is to investigate the effect of terminal group bonding on cell behaviour. For a benchmark environment, a biological control is necessary and protein (laminin) coated environments (P/LAM) have been chosen (41).



#### 4.2.1.1 Contact angle

Contact angle measurements (CAMs) reflect chemical and topographical characteristics of a material such as surface roughness, polarity, interfacial tension, and surface free energy. CAMs give an indication of biological response to materials such as proteins of interest adsorbing advantageously to direct cell adhesion and signalling as desired.

Using sessile drop technique of a solvent such as water (polar) or decanol (lipid) is released on a surface and the interfacial contact at the edges of the solvent is investigated. Static contact angle measurements were performed. In this method, a drop of solvent is released, and contact angle measurements are taken immediately after the drop stabilised on the surface. A high hydrophilic surface gives a CAM between 0 and 90° (e.g. hydroxyl surface, OH) and an angle between 90 and 180 gives low hydrophilicity/high hydrophobicity (e.g. methyl surface, CH<sub>3</sub>). Solvents that form a ball on the surface indicate a high contact angle. Surfaces can be hydrophilic and lipophilic at the same time (e.g. teflon). Hydrophilic surfaces attract proteins like albumin and hydrophobic materials attract proteins such as c3 fibronectin and vitronectin. Below are graphs and a table of water and decanol (lipid) contact angles for each environment:

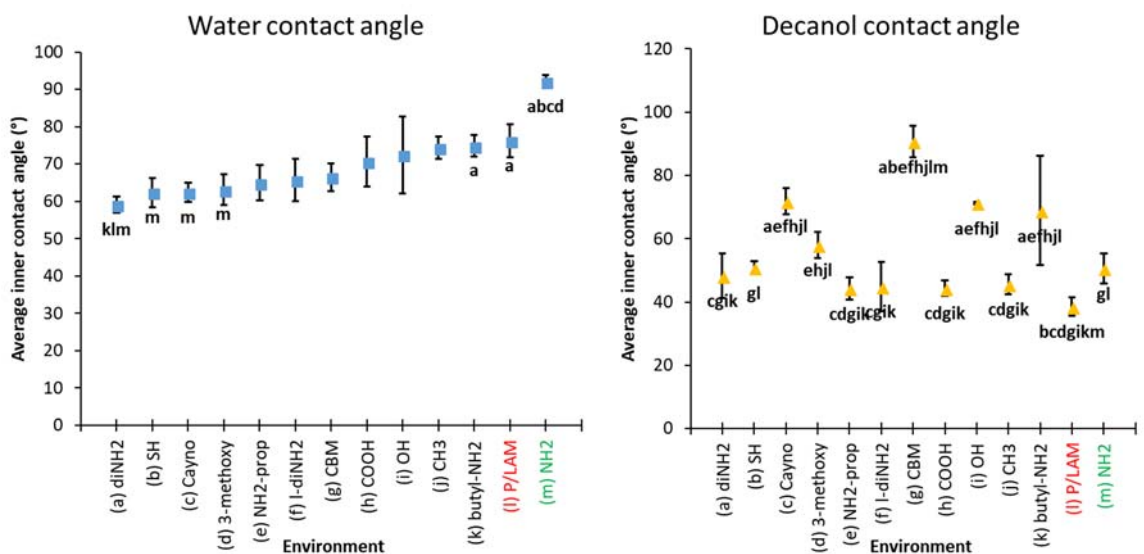


Figure 4.2: Contact angle measurements for environments used in this experimental chapter. The inner contact angle of a solvent (water/octanol) when it contacts a solid surface can be used as wettability/lipophilicity measures. The error bars indicate 1 standard deviation. No significant differences between samples were found (Kruskal-Wallis test with Conover-Iman pairwise comparison and Bonferroni correction).

The contact angles were captured near-zero contact time of solvent and surface with a camera at 160 frames per second. For both solvent types, this resulted in higher contact angles compared to the ones reported in the literature. Methyl ( $\text{CH}_3$ ) considered hydrophobic has lower contact angles compared to amine ( $\text{NH}_2$ ). Since  $\text{COOH}$  is synthesised on top of amine, the lower contact angles compared to amine are expected but this also disagrees with the literature (343). Hydroxyl provides similar contact angles with both solvents. Although this chemistry is hydrophilic, once soaked with water the contact angle rises (344). The diamine ( $\text{diNH}_2$ ) with an additional amine (replacing a carbon) has a lower water contact angle (WCA) compared to amine but similar decanol contact angle (DCA). This is expected as diamine has two polar atoms. Long diamine ( $\text{l-diNH}_2$ ) shows a similar pattern with higher WCA than diamine due to additional carbon atoms in the backbone hence the name long diamine. The cyano surface has a similar WCA with diamine but its DCA is higher compared to other nitrogen containing chemistries. This is because the solvent retention behaviour is altered due to more lipophilic interactions and less potential for hydrogen bonding compared to amines (345). Poly-d-lysine (PDL) and laminin surfaces are moderately hydrophilic and moderately lipophilic agreeing with the finding of the previous experimental chapter.

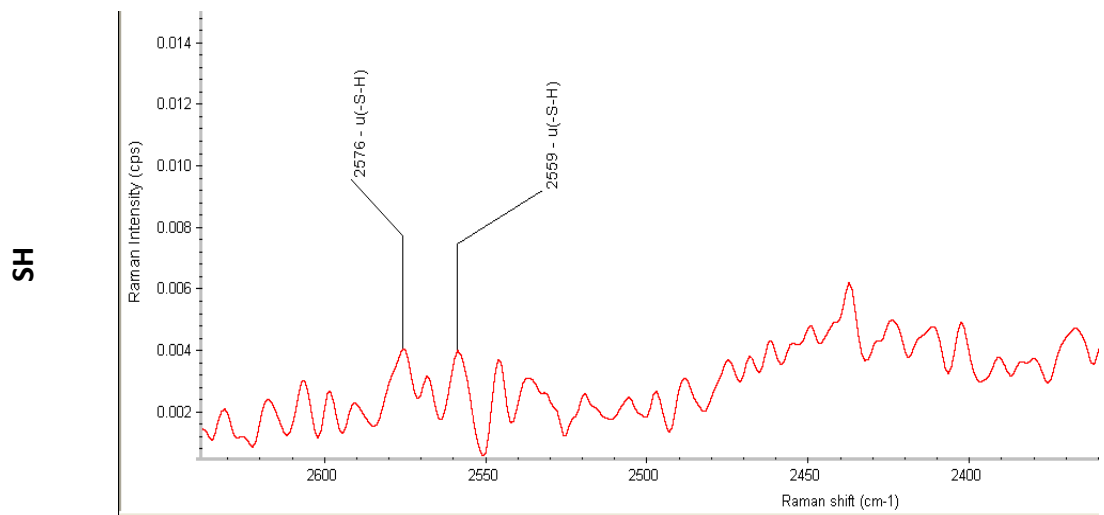
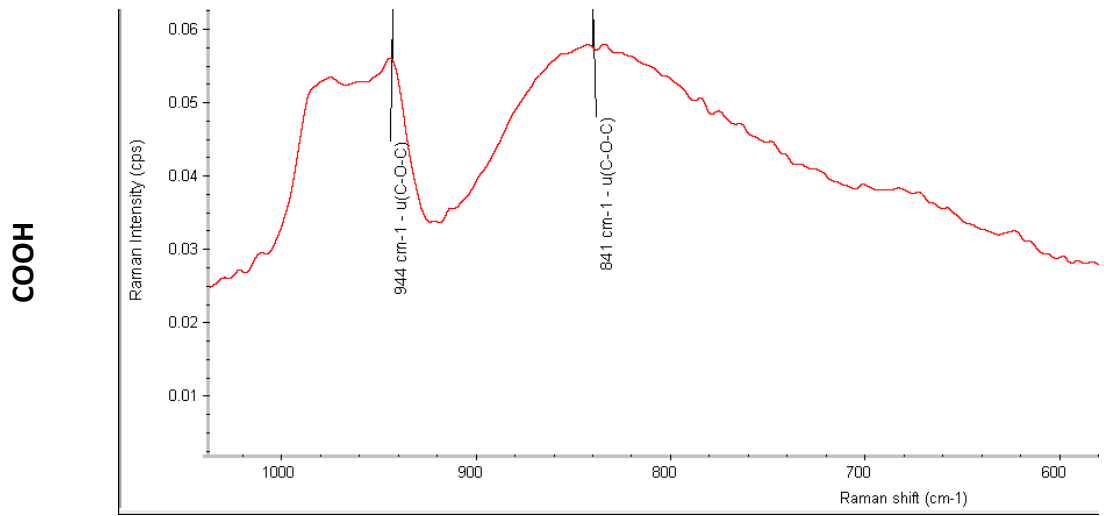
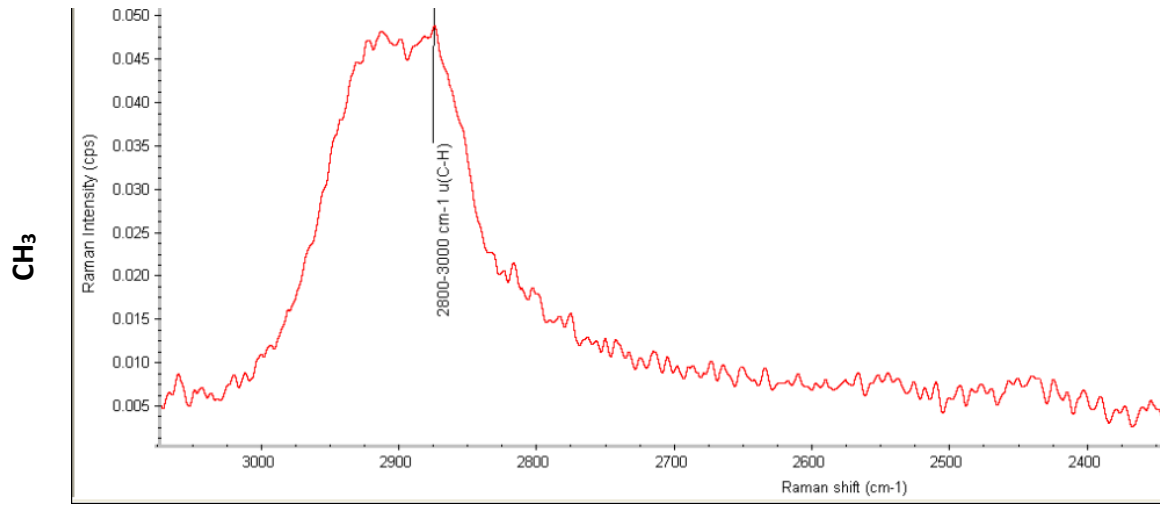
Table 8.5 in appendices shows the data used in the graphs above in addition to literature reported water contact angles. Butylamine's WCA sits between aminohexyl and amine but it is higher than cyano's. On the other hand, butylamine's DCA is like cyano's. This suggests the terminal group has an effect on wettability also found before (95). This is evident as well with propamine, where its amine group is closer the terminal group giving the water

contact angles like aminohexyl's. Carbomethoxy has similar WCA with 3-methoxy but is the most lipophobic (high DCA) from all other chemicals used.

#### 4.2.1.2 **Surface enhanced Raman Spectroscopy (SERS)**

Raman spectroscopy can be used to identify the chemical bonds of a sample surface by exciting it with light. Light deflects when it interacts with matter in the same way that particles scatter through collisions with other particles. Like infrared spectroscopy, this a vibrational technique to collect unique chemical fingerprints (346). Raman spectroscopy works with scattered light by the vibrating molecules. Raman has the advantage on low-frequency modes and water can be used as a solvent.

Surface Enhanced Raman Spectroscopy (SERS) technique is used for bond identification on samples where the signal would otherwise be weak. By adding metal particles (e.g. gold) and using the Raman instrument on particle pockets adsorbed on the surface the Raman signal increases in intensity (347). This technique is termed SERS and it works by combining electromagnetic, charge-transfer, and resonance signal enhancement mechanisms (348). The existence of this charge-transfer state increases the probability of a Raman transition by providing a pathway for resonant excitation. This mechanism is site-specific and analyte-dependent (349). Below are spectra of modified silicon wafers. Only the regions of interest are shown as the silicon peak is so high no other bond excitations are visible.



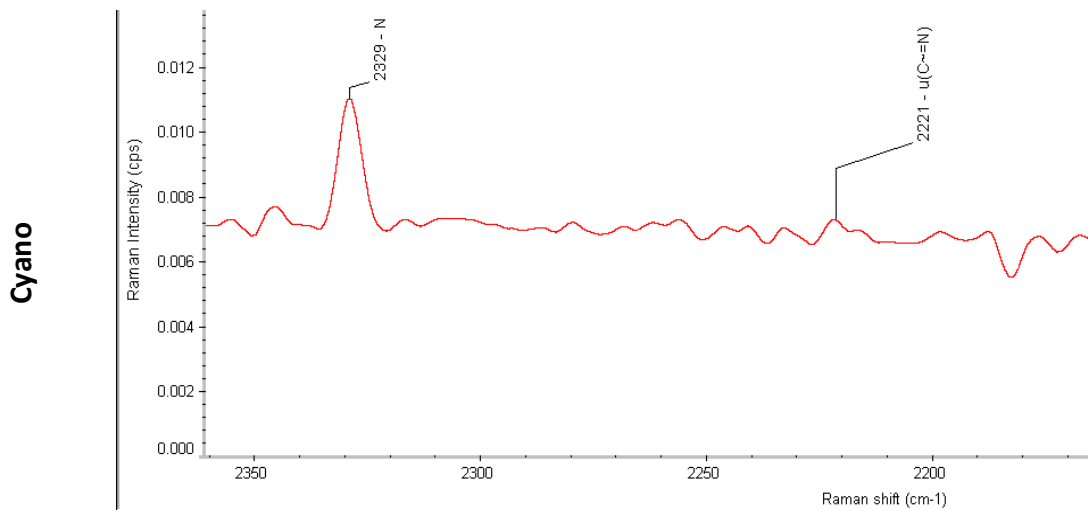
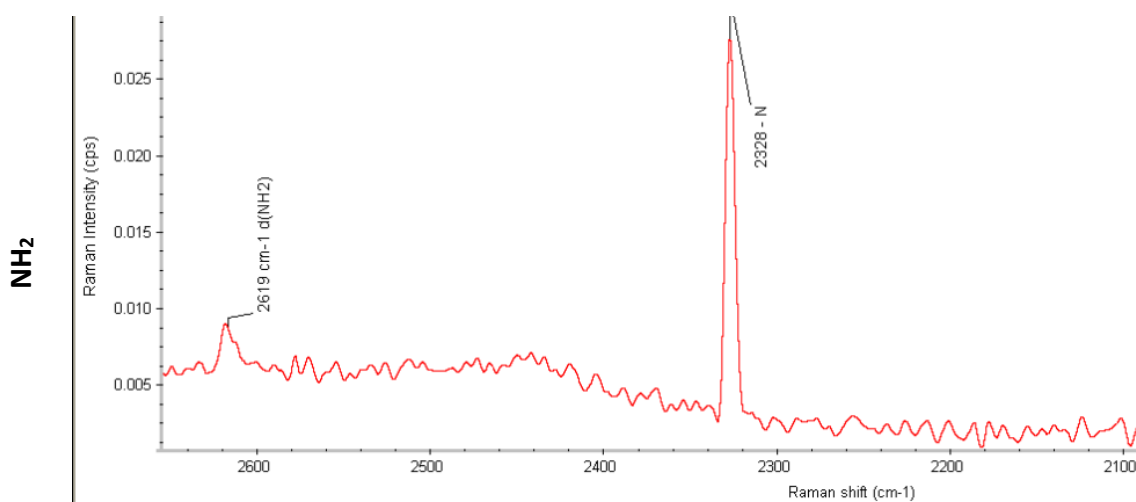


Figure 4.3: Surface enhanced Raman spectroscopy used for chemical characterisation of functionalised surfaces. The peaks indicate molecular bonding excitation as shown in the spectra.

The frequencies of vibration depend on the masses of atoms involved and the strength of the bonds between them. Light atoms and strong bonds have higher Raman shifts ( $x$  axis), and heavy and weak bonds have low Raman shifts. From the results above, methyl ( $\text{CH}_3$ ) has a peak at high frequency (between  $2800\text{-}3000\text{ cm}^{-1}$ ) indicating a carbon-hydrogen vibration. For carboxyl surfaces, the C-O-C vibration around  $80$  and  $950\text{ cm}^{-1}$  Raman shift is a good indicator as reported in this study (350). Thiol (SH) has a weak signal but the expected peaks are present and indicated in the graph. Cayno surfaces showed a peak around  $2330\text{ cm}^{-1}$  for the presence of nitrogen (351). Below are SERS spectra of the remaining surfaces:



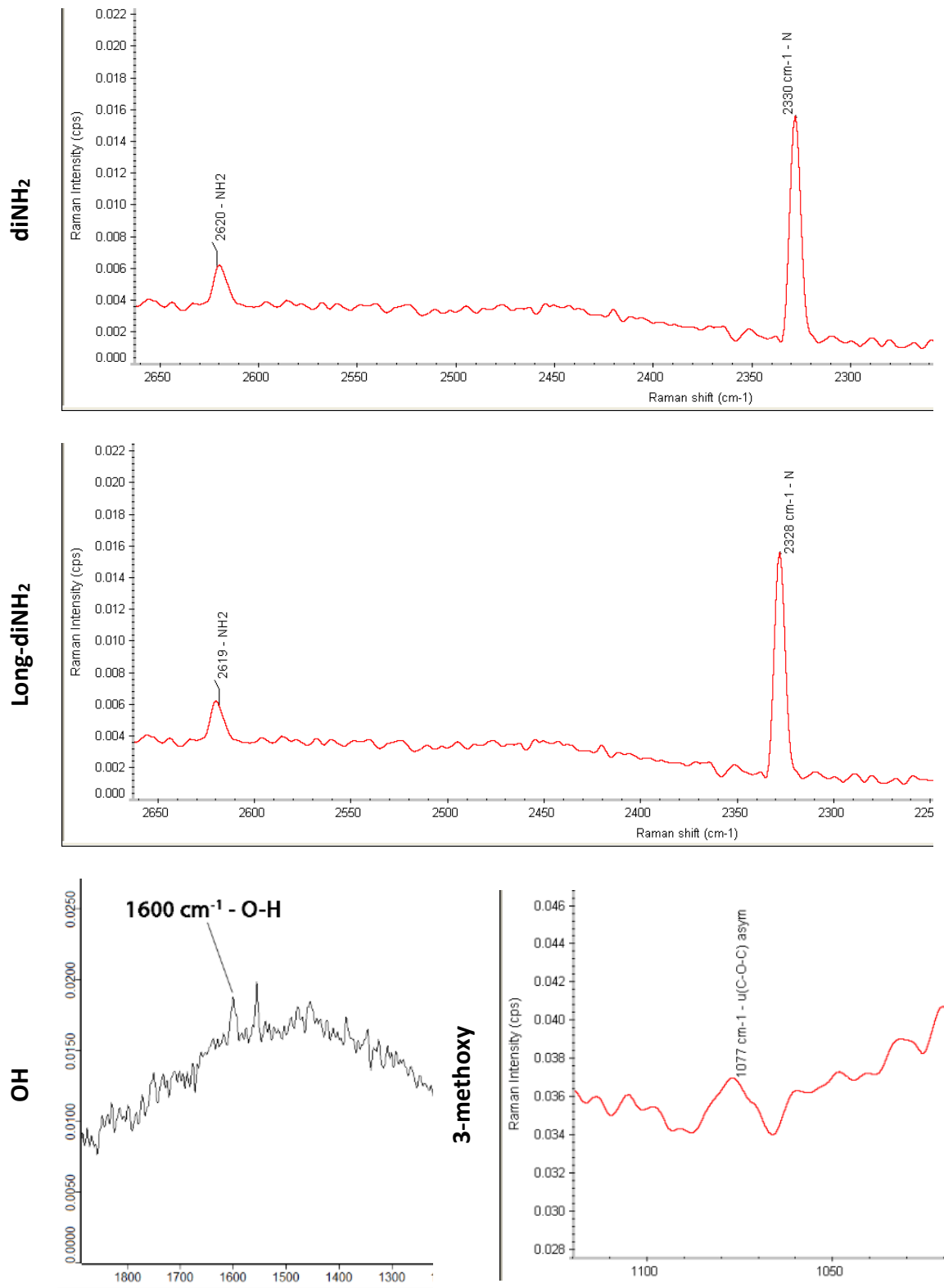


Figure 4.4: Surface enhanced Raman spectroscopy used for chemical characterisation of functionalised surfaces. The peaks indicate molecular bonding excitation as shown in the spectra.

From the above results, amines (NH<sub>2</sub>, diNH<sub>2</sub>, and long diNH<sub>2</sub>) have two characteristic peaks around 2330 and 2620 cm<sup>-1</sup> Raman shift that is also found in another study (352) investigating single amine and diamines self-assembled monolayers. The hydroxyl surface

showed a characteristic O-H vibration at  $1600\text{ cm}^{-1}$  and 3-methoxy showed asymmetric C-O-C vibration with a peak around  $1080\text{ cm}^{-1}$ .

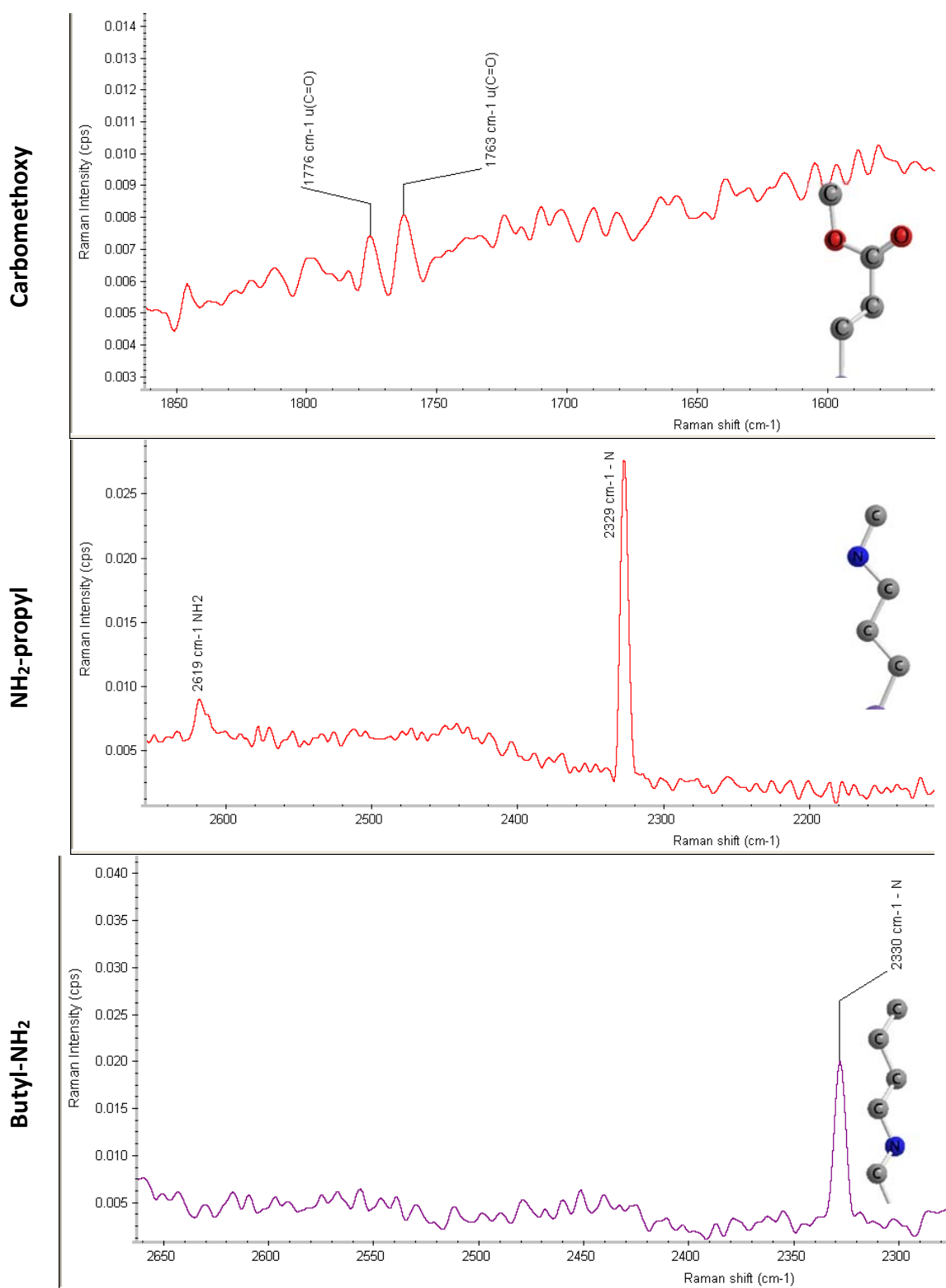


Figure 4.5: Surface enhanced Raman spectroscopy used for chemical characterisation of functionalised surfaces. The peaks indicate molecular bonding excitation marked in the spectra.

From the results above, carbomethoxy has two Raman emissions between 1763 and 1776  $\text{cm}^{-1}$  indicating a carbon-oxygen double bond vibration. Propamine ( $\text{NH}_2\text{-prop}$ ) surfaces, have nitrogen bond excitation around 2330  $\text{cm}^{-1}$  and around 2620  $\text{cm}^{-1}$  Raman shift both found in another study (352) as well. On the other hand, butylamine ( $\text{butylNH}_2$ ) has nitrogen bond vibration around 2330  $\text{cm}^{-1}$  Raman shift at a lower intensity as well. An explanation for this is the amine group being closer to the labile group (silane head) requiring more energy to detect a signal at 2620  $\text{cm}^{-1}$ .

#### 4.2.1.3 X-ray Photoelectron Spectroscopy

XPS is a surface chemistry characterisation technique allowing the investigation of elemental composition of solid samples. XPS does not require much sample preparation but it is performed in vacuum conditions to avoid atmospheric noise issues. The principle of XPS is to irradiate a sample with narrow wavelength band (monochromatic, non-destructive) x-ray beam. When the atom or molecule absorbs x-ray photon, electrons eject. The kinetic energy of electrons depends on the photon energy and the binding energy of the electron (i.e. the energy required to remove the electron from the surface). By measuring the kinetic energy of the emitted electrons, it is possible to determine which elements are near a material's surface, their chemical states and the binding energy of the electron (in electronvolts, eV) (353). Below are XPS spectra of environments used in cell culture for this chapter:



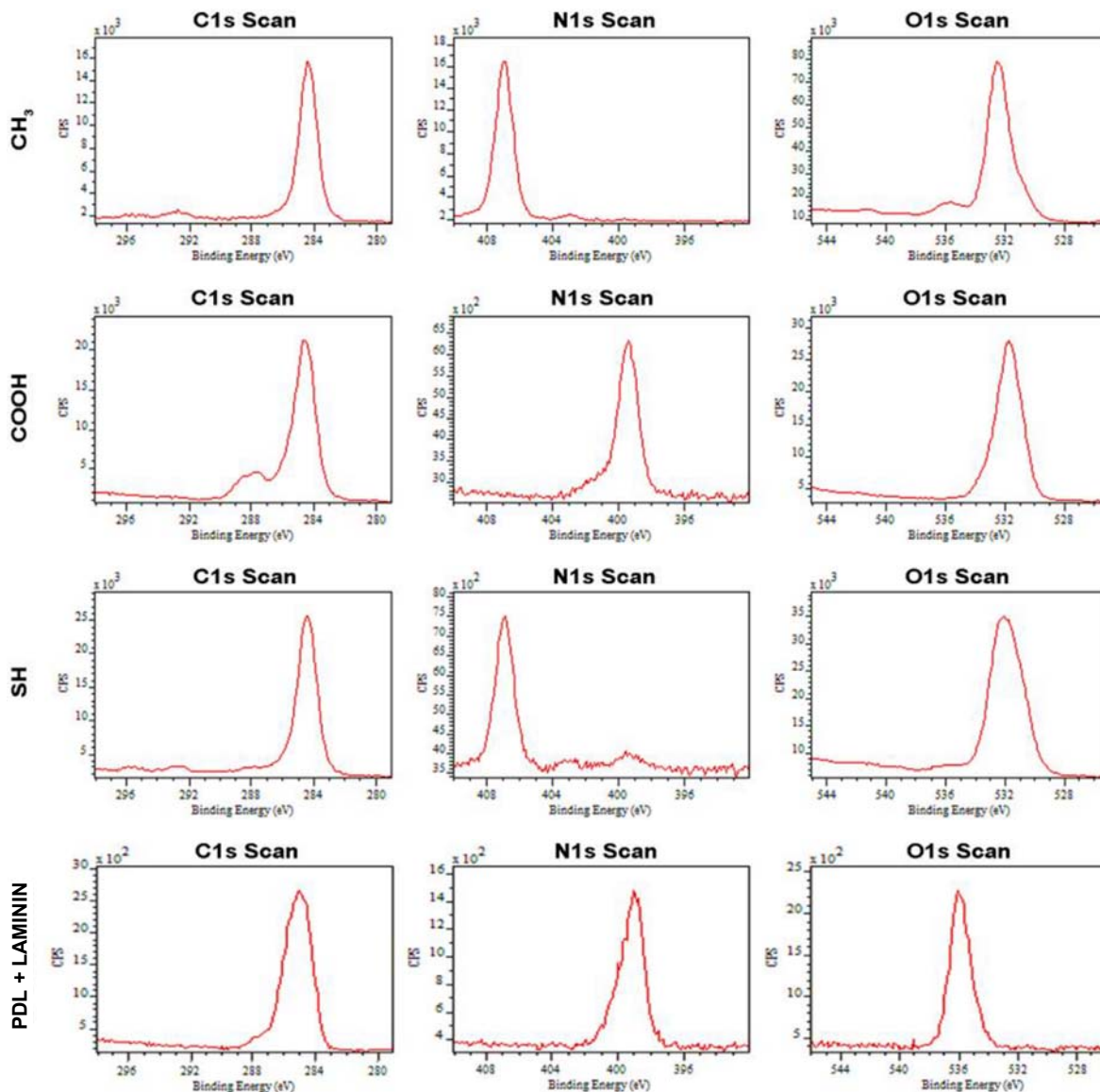


Figure 4.6: XPS spectra used for chemical characterisation of functionalised surfaces. Spectra in this table are common with previous work. Acquired from (41). y axis is electron count per second and x axis is binding energy in electronvolts (eV). Service provided by the National EPSRC XPS Users' Service (NEXUS) at Newcastle University.

From the image above, peaks indicating the presence of elements from modified surfaces are shown. The presence of carbon, nitrogen, and oxygen is sought as these are contained in the self-assembly molecules selected for modifying cell culture surfaces. Methyl has a strong carbon signal, and so does carboxyl (COOH) surfaces with the longest carbon (alkyl) backbone. COOH has the smoothest nitrogen and oxygen spectra. Biological control surfaces with poly-d-lysine and laminin have strong peaks for carbon, nitrogen, and oxygen all being common in proteins. Below is the next batch of XPS spectra of environments:

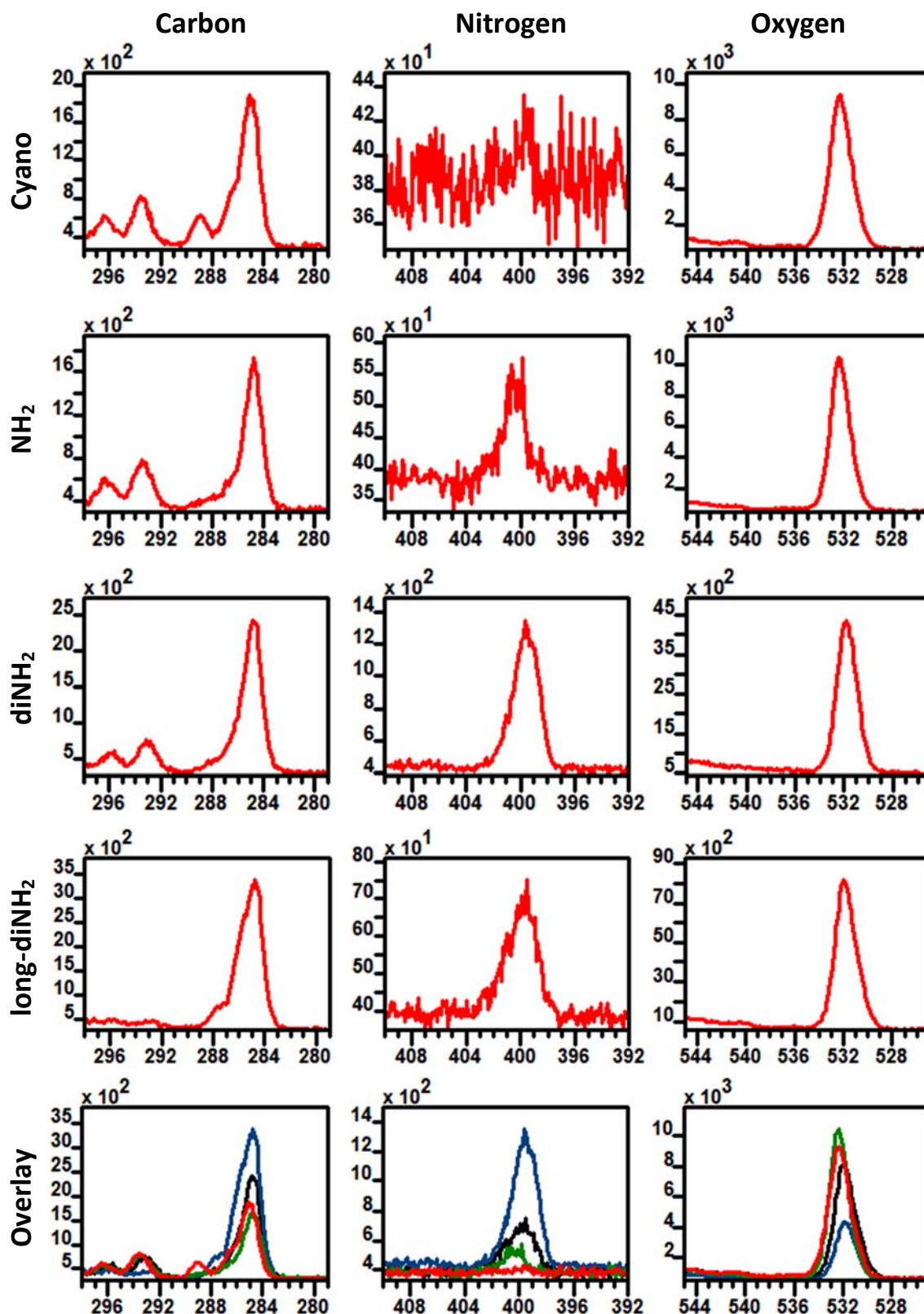


Figure 4.7: XPS spectra of nitrogen containing surface chemistries. In the overlay, in blue is long diamine, black is diamine, green in amine, and red is cyano. y axis is electron count per second and x axis is binding energy in electronvolts (eV). Service provided by the National EPSRC XPS Users' Service (NEXUS) at Newcastle University.

From the graphs above, almost all nitrogen containing surfaces have smooth spectra. The exception is cyano ( $R-C\equiv N$ ) being noisy. It is suspected this is due products from x-ray

irradiation or impurities. From the overlay, the intensities for carbon, nitrogen, and oxygen are as expected for all the nitrogen-containing chemicals. There is a trend with carbon intensity increasing with carbon content in functionalities (*long diamine* > *diamine* > *cyano* > *amine*). A similar trend is observed with nitrogen intensity increasing as amine content increases (*long diamine* > *diamine* > *amine* > *cyano*). On the other hand, oxygen intensity has an inverse relationship compared to nitrogen. As nitrogen content increases, oxygen intensity decreases also found in previous results (264) (*cyano* > *amine* > *diamine* > *long – diamine*). This could be from bonding (x-ray irradiation), glass, or contamination. Below is the last batch of XPS spectra for remaining environments for this chapter:

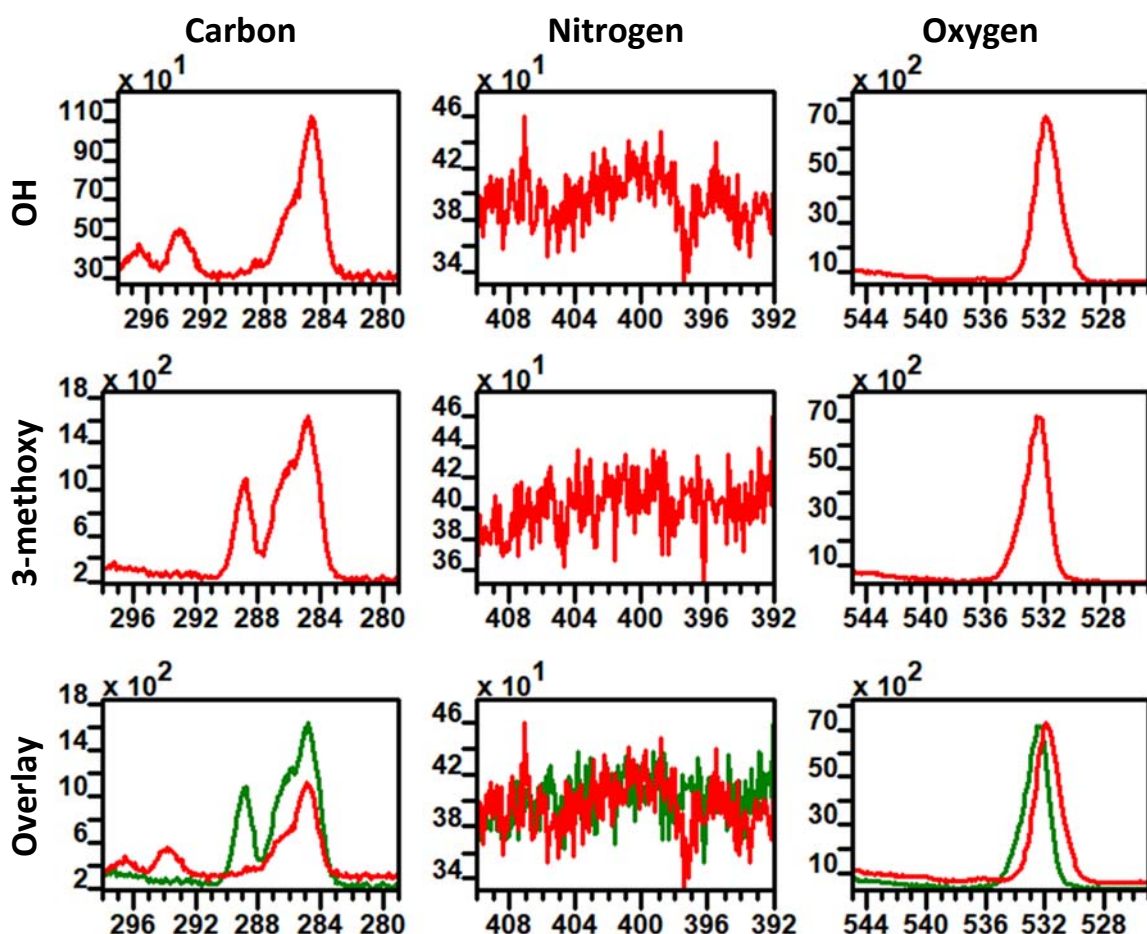


Figure 4.8: XPS spectra of oxygen containing surface chemistries. In the overlay, in green is 3-methoxy and in red is hydroxyl (OH). y axis is electron count per second and x axis is binding energy in electronvolts (eV). Service provided by the National EPSRC XPS Users' Service (NEXUS) at Newcastle University.

On oxygen containing chemistries, strong carbon and oxygen peaks are present as expected. 3-methoxy has a longer carbon backbone hence the increased carbon intensity in the XPS spectra. The noisy nitrogen spectra showing for both hydroxyl and cyano surfaces is attributed to bonding (x-ray irradiation), glass, or contaminants.

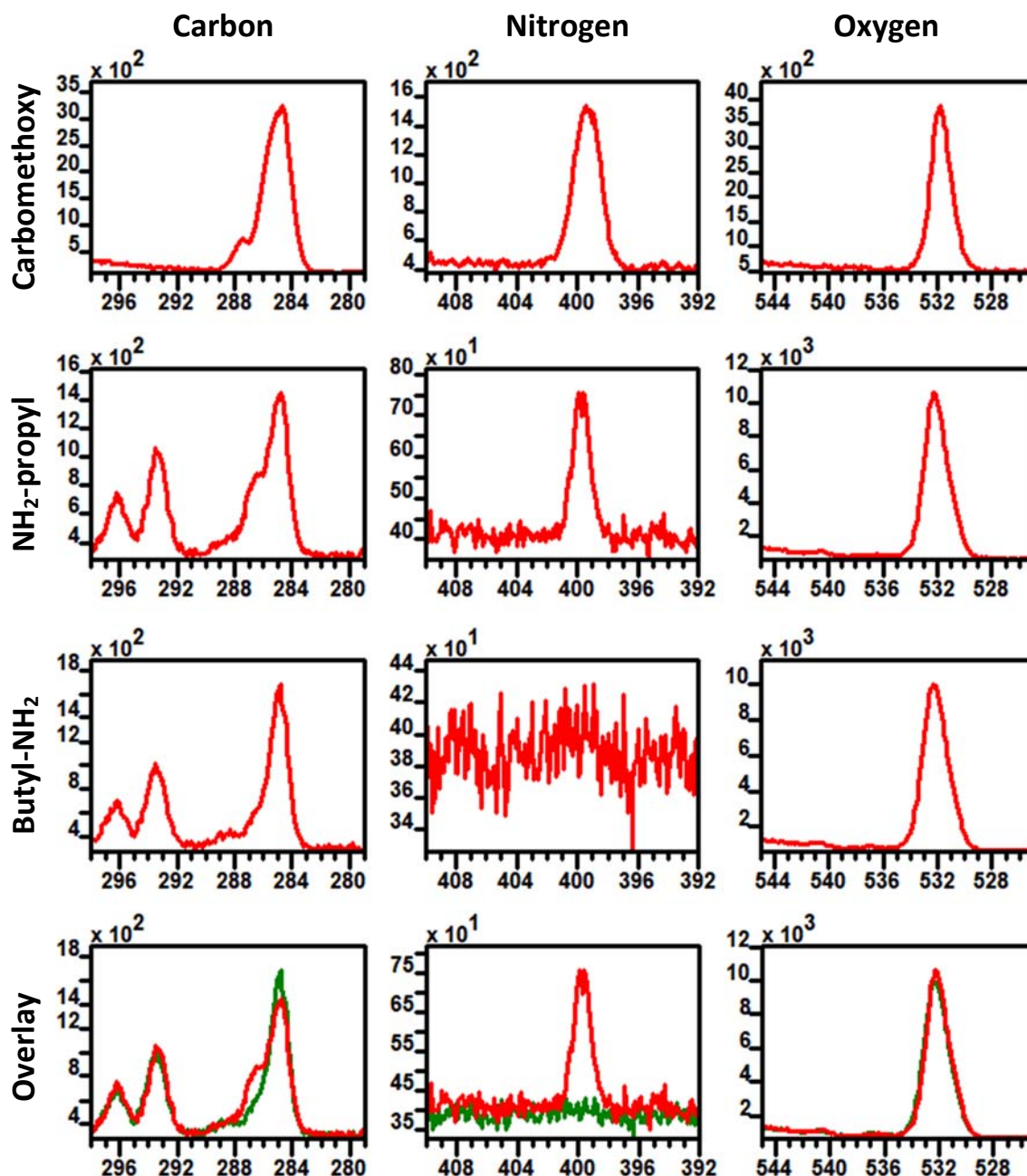


Figure 4.9: XPS spectra used for chemical characterisation of functionalised surfaces. In the overlay, in red is aminopropyl (NH<sub>2</sub>propyl), and in green is butylamino (butylNH<sub>2</sub>). y axis is electron count per second and x axis is binding energy in electronvolts (eV). Service provided by the National EPSRC XPS Users' Service (NEXUS) at Newcastle University.

From the graphs above, the peaks of interest are the highest ones in the nitrogen column. Almost all graphs have smooth spectra except for butylamine with nitrogen. This can be explained by the amine's position in this chemical being lower and closer to the labile group compared to propamine or it is due to x-ray irradiation or impurities. Carbomethoxy has stronger carbon (3500 eV) but weaker oxygen peaks (4000 eV) compared to hydroxyl (OH, 110 eV/7000 eV) and 3-methoxy (1800 eV/7000 eV) shown in the previous chapter (4). Although carbomethoxy has the same carbon atom count as 3-methoxy, its double carbon bond increases carbon intensity in the XPS spectra.

From the overlay, the intensities for carbon, nitrogen, and oxygen are as expected for both the nitrogen-containing chemicals. As in the previous chapter, carbon intensity increasing with carbon content in functionalities (*butylamine > amine – prop*). The nitrogen intensity is suspected to be from position of the amine group in the molecule as both molecules have one amine each. This also explains why both have almost the same oxygen intensity. From previous work (264) and findings from the previous chapter (4), oxygen was found to have an inverse relationship with nitrogen meaning as nitrogen content increases, oxygen intensity decreases.

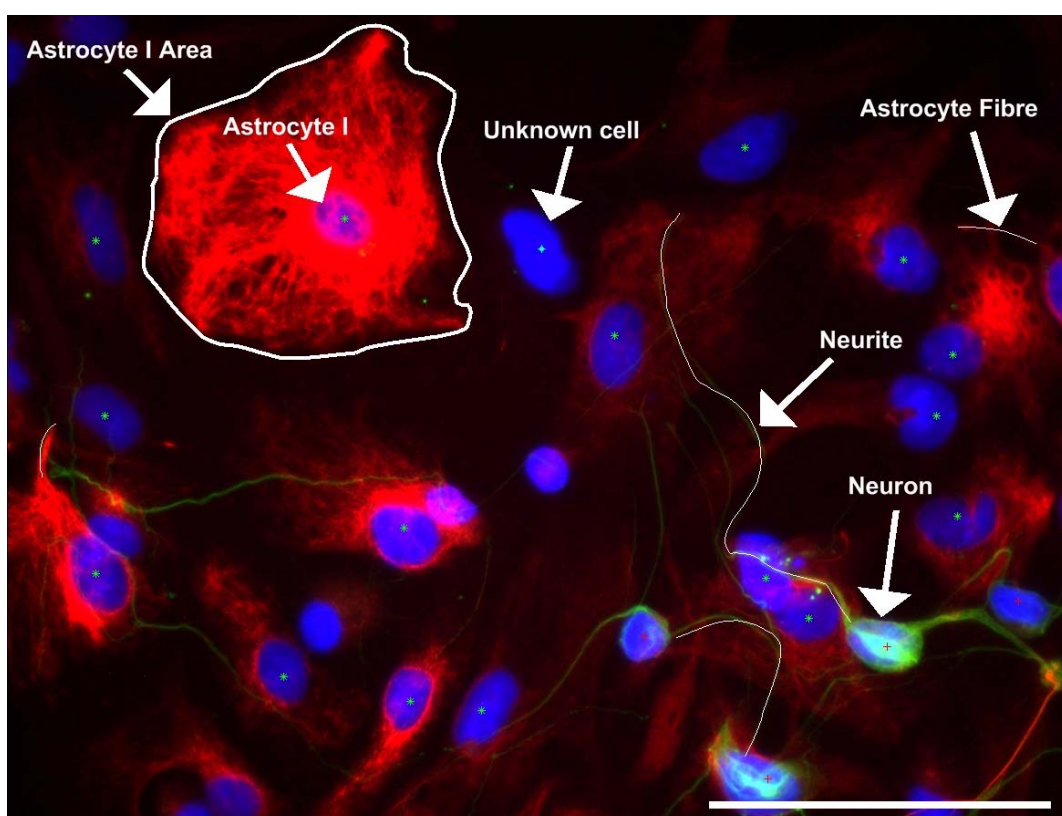
Following from the above presented results the surface chemistry characterisation part concludes here with confidence that the presenting chemistry of each modified surface is what we think it is.

#### 4.2.2 Cell images and measurements

This section is about analysing cell images to quantify cell performance relating to their visual form and structure (morphologically). Cell performance metrics are necessary to

profile each synthetic environment for their effect on tissue formation. Modified surfaces were seeded with cell spheres (neurospheres) and on day 3 and 7 these were “fixed” in place. Fixation preserves cells and tissues and terminates any ongoing biochemical reactions using a cross-linker such as paraformaldehyde typically seen in museums with preserved animals in jars. Fixed cells were tagged with fluorescent markers that selectively bind to cell types of interest. By shining fluorescent light at different wavelengths, it is possible to visualise anatomically target cells. Images of fixed cells were acquired using an automated fluorescent microscope (Nikon Ti).

Cell images were analysed with NIS Elements software bundled with the Nikon Ti microscope. Essentially, measurements such as length and area are in pixels, the building blocks of digital images. Pixels are converted to micrometres and the conversion depends on the camera view area and magnification lens used. This calibration allows for the conversion to real values. Below are images with example cell measurements of morphological performance:



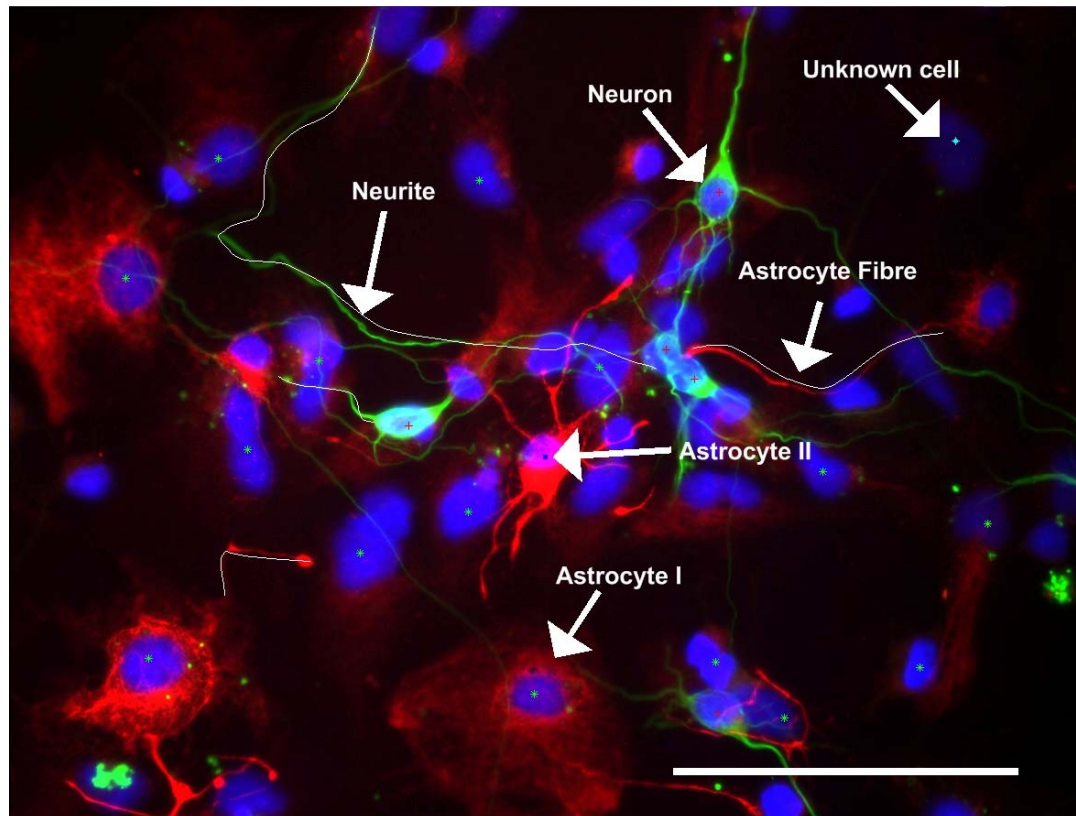


Figure 4.10: Example performance measurements of cell morphology continued. Cell body spreading is evident as type I astrocyte area; cell projection length is shown as neurite and astrocyte fibre; and cell proportion is derived from counts of cell types shown as type I (protoplasmic) and II (fibrous) astrocyte, neuron, and unknown type cells. Blue dots indicate the presence of cells (DAPI), red material indicates astrocytes (GFAP), and green material indicates neurons ( $\beta$  III tubulin). Scale bar is 100  $\mu$ m.

#### 4.2.2.1 Cell cluster area and spreading

Cell cluster area is related to cell sphere (neurosphere) spreading early after seeding them on modified surfaces, and with cell proliferation especially in the later time point (day 7). The effects at play here are both chemical and biological. When neural stem cells and progenitors are cultured as spheroids, a clear indicator of differentiation is cell adhesion and migration away from the sphere causing it to flatten with time. In the first stage of differentiation the neurospheres attach to a high affinity surface. Biological control surfaces with adsorbed laminin protein are “good” because they have plenty of adhesion ligands specific to neural cells. The neurospheres deconstruct and cells differentiate with astrocytes migrating away from the sphere providing the foundation layer for neurons to migrate away (270) and release neuron maintenance factors. Environment permitting,

neurons also make short range migrations away from the neurospheres independently in a process called “chain migration” (271).

Fluorescence microscopy was used with tagging markers to identify astrocyte and neural cell populations. Neurospheres were observed to attach on all surfaces initially between 1-2 hours. Neurospheres attachment on hydrophobic surfaces needed more time compared to less hydrophobic surfaces. Below are images of smallest and largest cell clusters from synthetic environments and below that follow biological control images for comparison.

The cell cluster graphs below show the differences across different environments and at the base of each bar there is an area multiplier. This multiplier is the area increase from the theoretical baseline area calculated from the average neurosphere diameter ( $\varnothing_{NS}$ ) with:  $Area_{NS} = (\pi \frac{\varnothing_{NS}}{2})^2$ . The average neurosphere diameter was obtained from 27 neurospheres (3 experiments) prior to seeding on surfaces.

Here, maximising the cell cluster area/neurosphere spreading is desirable as this increases cell differentiation potential. Large cell clusters are observed in biological control environments therefore the larger the cluster observed in synthetic environments the better. Cell cluster area and graphs are shown below the images:



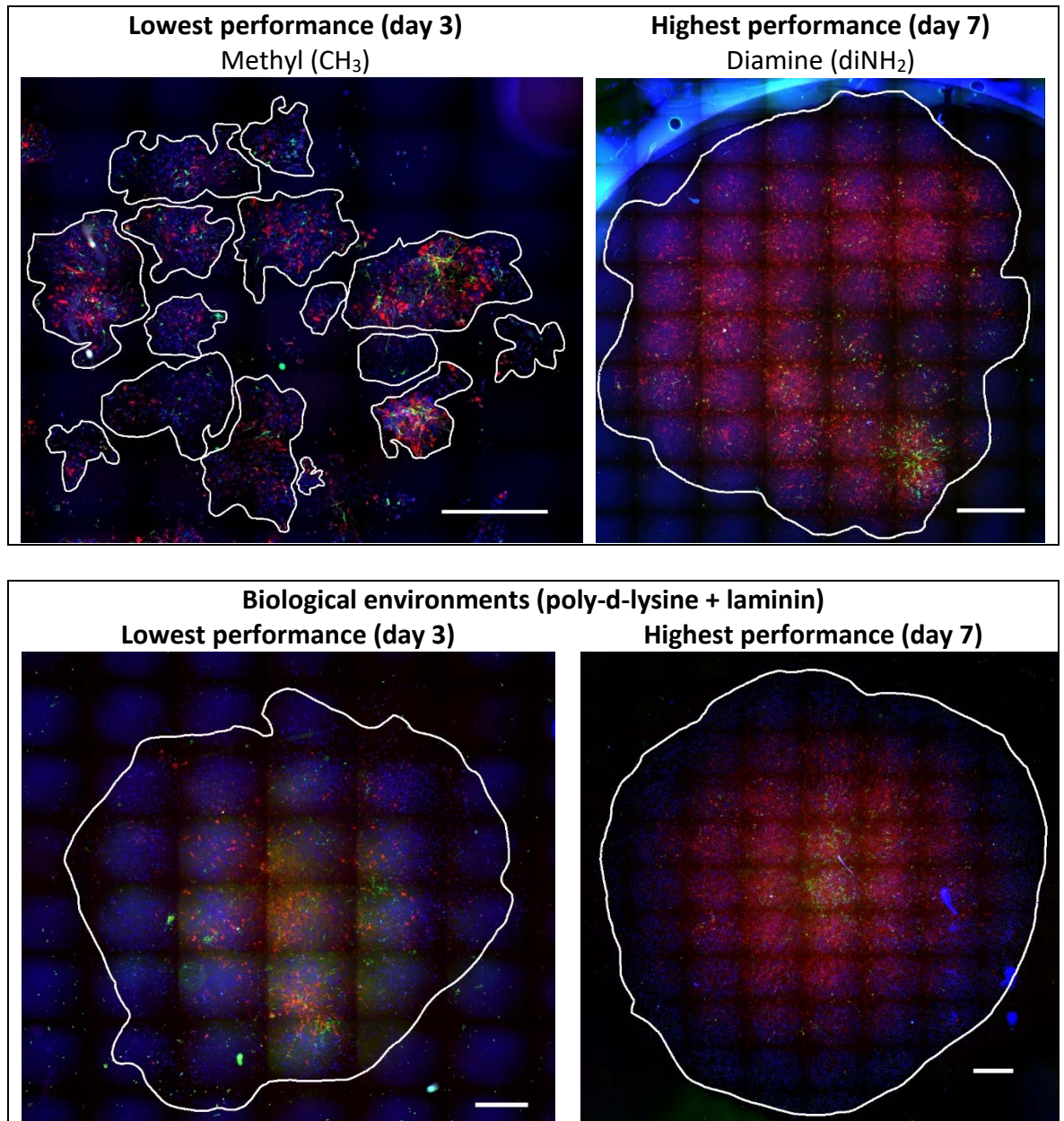


Figure 4.11: Cell cluster images from synthetic (top row) and biological (bottom row) environments. The cell clusters are encased with a white line. Left image shows the smallest cell clusters and the right one the largest. Blue dots indicate the presence of cells (DAPI), red material indicates astrocytes (GFAP), and green material indicates neurons ( $\beta$  III tubulin). Scale bar (bottom right) is 1 mm.

Notice the scale bar size for comparison. The smaller the bar the larger the cluster. From synthetic environments, methyl ( $\text{CH}_3$ ) has the smallest cell clusters in the early time point (day 3) and diamine ( $\text{diNH}_2$ ) has the largest. The biological control has one of the the largest cell clusters in the early time point and by far the largest on the late time point (day 7). The graphs below show the median cell cluster area for all environments on both time points:

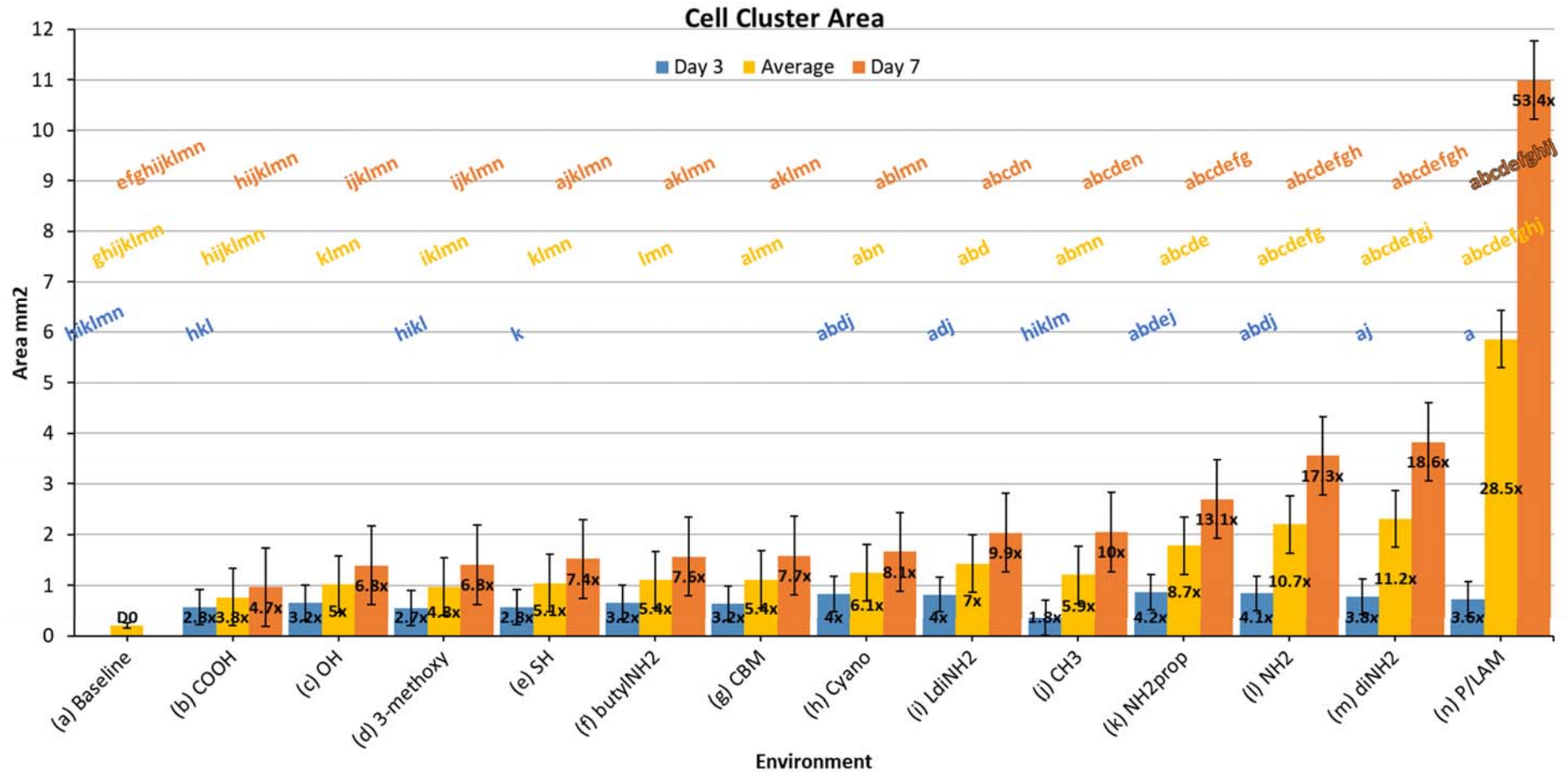


Figure 4.12: Cell cluster area graphs. The graphs show the median cell cluster area from environments used in experiments for both time points side by side with the average between.  $y$  axis for the area is in  $\text{mm}^2$  and  $x$  represents the environments sorted by their smallest to largest cell cluster values from day 7. Baseline area is converted from neurosphere diameter discovered prior to seeding at day 0 (D0). The multiplier at the base of each bar indicates how many times larger the area is compared to the baseline. The error bars on baseline bar indicates the standard deviation and for the rest indicates the median absolute deviation. The labels on top of the bars are the significant differences between groups colour coded by time point.

Referring to the graphs in Figure 4.12, methyl ( $\text{CH}_3$ ) has the smallest cell clusters on the early time point (day 3) with a 1.7-fold increase from baseline area (day 0). Environments containing nitrogen terminations (Cyano,  $\text{l-diNH}_2$ ,  $\text{diNH}_2$ ,  $\text{NH}_2$ ) perform better than biological control environments (P/LAM) with larger cell clusters. That is 4-fold and 3.5-fold increase in cell cluster area from baseline respectively.

In the later time point (day 7), carboxyl ( $\text{COOH}$ ) has the smallest cell clusters with 3.7-fold increase from baseline. Diamine ( $\text{diNH}_2$ ) has the largest with 18.6-fold increase from baseline. Similar findings were observed in previous work (264). From synthetic environments, cell clusters spread better on amine environments. Biological control environments are good with cluster spreading on day 3 but by day 7 the cluster area are almost 3 times larger from the best synthetic scorer, diamine. The lowest performers in this time point are acid terminations ( $\text{COOH}$ ,  $\text{OH}$ , 3-methoxy) which is expected. Their average pKa values is 4.5 being the lowest among all synthetic environments.

The bottom graph of Figure 4.12 shows all results side by side. From synthetic environments of amine and diamine, provide the largest cell clusters and carboxyl ( $\text{COOH}$ ) the smallest. The biological control (P/LAM) was among the highest particularly on the later time point with 53-fold area increases from baseline. The best synthetic environment (Amine) scored 18-fold area increases from baseline.

The remaining environments are propamine ( $\text{NH}_2\text{prop}$ ), carbomethoxy (CBM), and butylamine ( $\text{butylNH}_2$ ). From the graph above, propamine performs almost the same as amine on day 3 but the latter has larger cell cluster area on day 7. Butylamine's performance is mediocre and is outperformed by carbomethoxy overall.

#### 4.2.2.2 Neuron density and proportion

Successful cellular therapies to regenerate nervous tissue depend partly on the amount of neural cells delivered. Neuronal network allows function such as voluntary bodily movement. Controlling the density and proportion of transplant relevant cell populations is a key element in developing and scaling up cell-based therapy. Cell density tells us how close the cells are to each other and cell proportion tells us how many of distinct cell types are there compared to total cell counts. In cell therapy translation, controlling the proportion of cells and the purity of the transplant population is a critical quality attribute (281). An imbalance in the proportion and migration of cells can have adverse effects for transplant recipients such as uncontrolled movement (overproduction of serotonin in the transplant) (282). Another effect is teratomas from progenitors or stem cells if they are present in the transplant tissue (91).

For the purposes of developing therapy grade tissue, a biological benchmark is required and we chose laminin environments (41,97) for this. Unfortunately, biologically derived materials for surface modification cannot be used in tissue engineering for clinical therapies due to concerns over pathogens. Cell culture environments with synthetic chemistry can be made pathogen-free and provide greater degree of control compared to alternatives such as special culture media, and hypoxia as an environmental culture condition among others.

Cell performance measures were obtained from experiments with synthetic environments and neural cells. These measures include cell density (cells/mm<sup>2</sup>) and cell proportion (%) for each cell type stained namely neurons, astrocytes type I and II, and unknown/unstained cells. Cell densities are comparable between different environments by standardising the cell counts with their median cell cluster area. Cell proportion with cell density together

inform on cell differentiation and migration. These two parameters can be used interchangeably to declare the “best” environment depending on the intended use of the tissue. These cell parameters provide additional means to compare environments. For example, in the case of similar cell cluster area, the proportion of a cell type (for example neurons) will inform on the ideal environment.

Cells around the cell cluster but not the dense centre were quantified in random sampling. This is because in the dense centre, cells cannot be distinguished or measured with fluorescence imaging. Cell scores were obtained in two time points. Day 3 neural density and proportion informs on neural differentiation. At this stage, high density means cells reside inside the neurosphere because the environment is not ideal for them. Low neural density is a strong indicator of differentiation if the cells survive.

Day 7 time-point is a good indicator of environment remodelling and cell proliferation due to the duration of the cell culture (101). A situation where neuron density is similar, but the cell cluster area is larger means cells are dividing. In tissue slices and xenografts, higher cell density means smaller extracellular volume and amount suggesting cells use the resources in the vicinity quicker (272,273). Low cell density promotes internal cell signalling for changes within individual cells (autocrine signalling); high cell density promotes cell-cell communication inducing changes in nearby cells (paracrine signalling) (274).

Neuron density is related to neuron proportion and cell cluster area. Cell density is derived by standardising total neuron count with median cell cluster area. This changes the relationship between the cell density and proportion visible as the horizontal distance between the green and red data points on the same  $x$  axis. The ideal environment would minimise neuron density and maximise neuron proportion at the same time. Below are cell

images of synthetic and biological environments and after that, follow graphs of cell density and proportion for each cell type investigated:

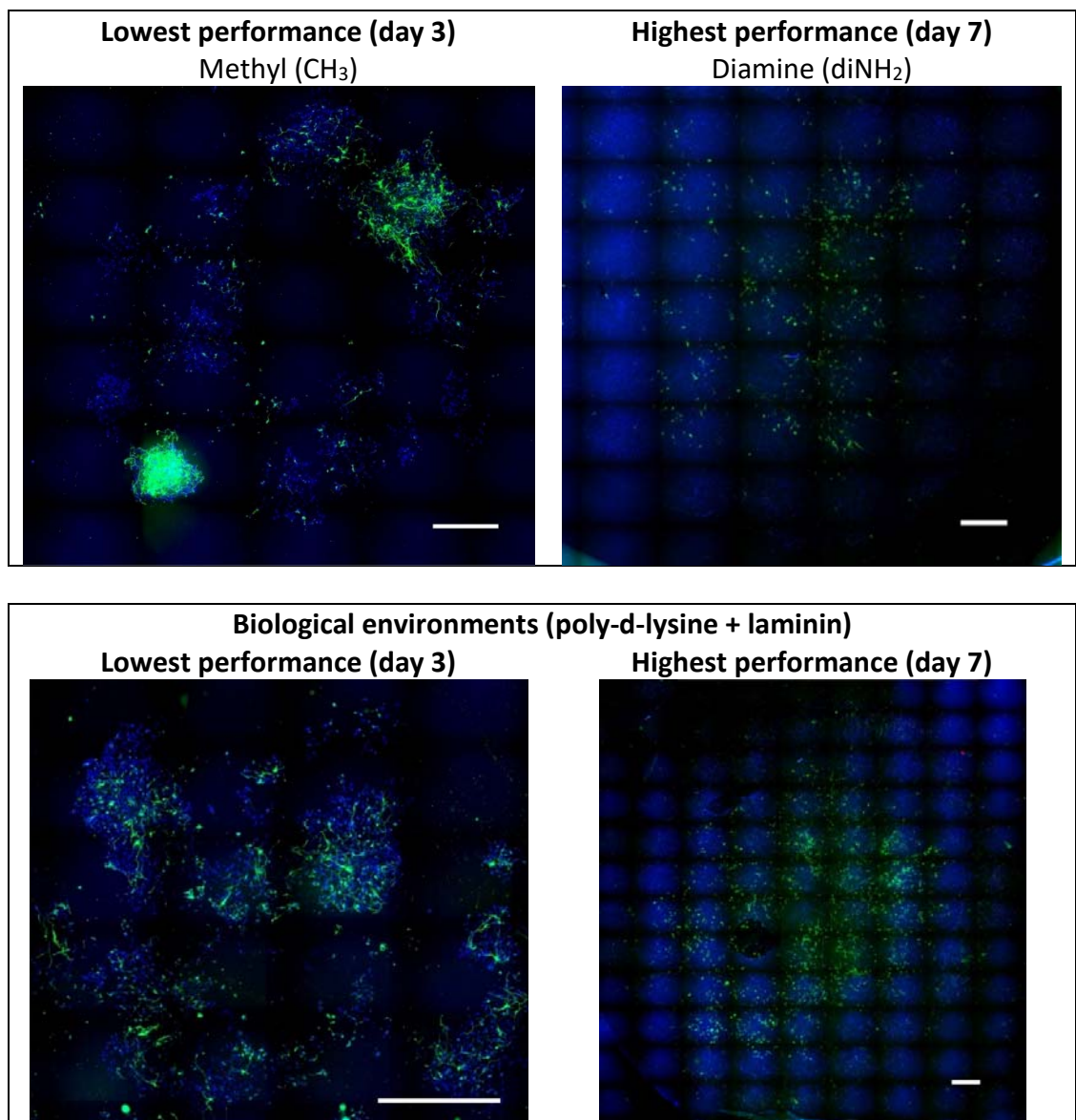
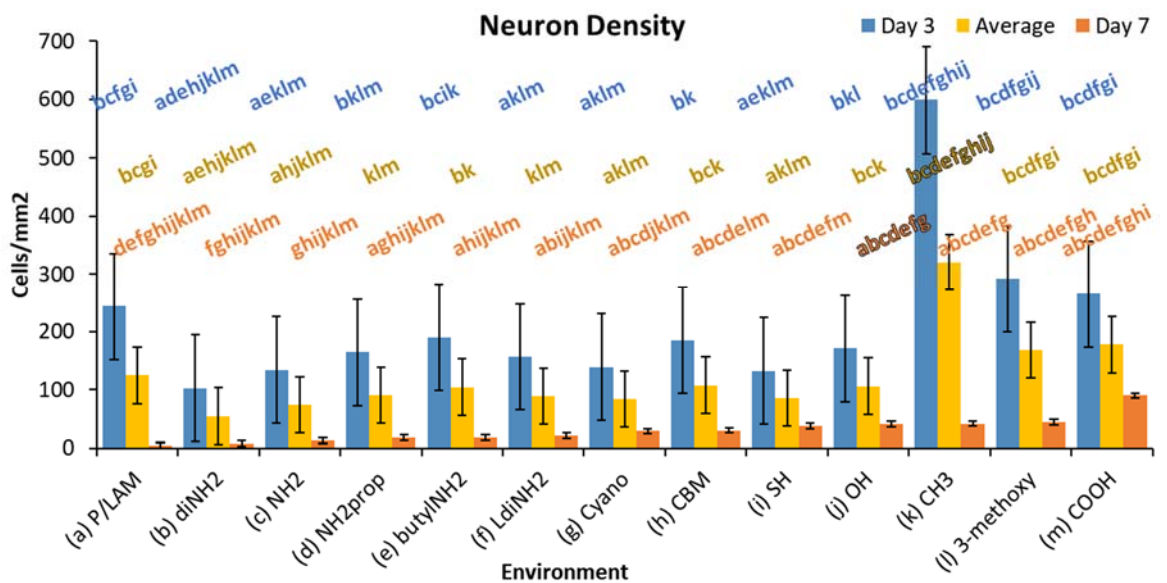


Figure 4.13: Neuron density and proportion images from synthetic (top row) and biological (bottom row) environments. Left columns shows the highest cell density and the right one the lowest. The more intense green the image is, the stronger the presence of neurons. Blue dots indicate the presence of cells (DAPI) and green material indicates neurons ( $\beta$  III tubulin). Scale bar (bottom right) is 1 mm.

A visual method to interpret cell density is the green intensity and cell cluster area. The more intense that the green is and the larger the cell cluster area the better. For cell proportion alone, the greener the image the stronger the presence of neurons regardless of green intensity (cell density). From the top left image of Figure 4.13, methyl (CH<sub>3</sub>) has the highest neuron density in the early time point (day 3) and diamine has the lowest density in the later time point (day 7). The biological control on day 3 has low cell density

and high proportion of neurons. By day 7, laminin has both lower cell density and higher neuron proportion compared to the best synthetic scorer (diamine).

On day 3 images, neuron density is high evident by the green intensity. There are more neurons on the biological control. On day 7 time point, the difference between the synthetic and biological environment is again the green intensity. In the latter, neurons are less dense compared to the former. Neuron proportion on the latter time point (day 7) should drop because this cell type does not proliferate after differentiation (G0 phase) (354,355). The decrease in cell proportion is from lower cell density or it could be from cell death. Below are graphs of neuron density and proportion:



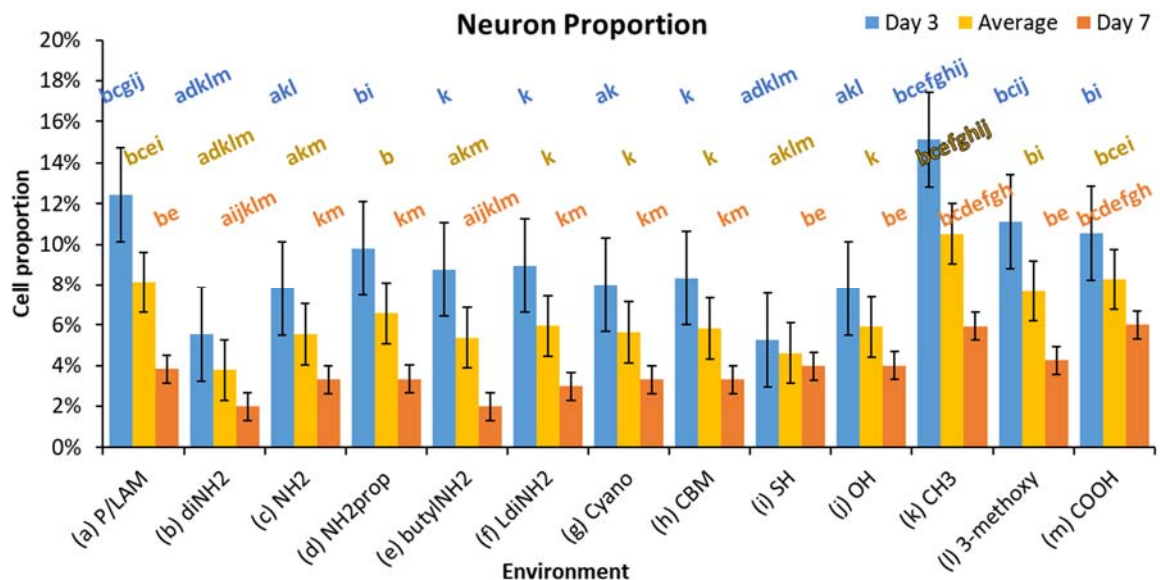


Figure 4.14: Neuron density and proportion results obtained from cell culture experiments.  $y$  axis is the cell density (cells/mm<sup>2</sup>) or cell proportion (%) and  $x$  axis are the environment sorted by lower to higher day 7 cell density for comparison. The error bars indicate the median absolute deviation and the labels on top of the bars are the significant differences between groups colour coded by time point.

Top row graphs (Figure 4.14) show the median neuron density on the left  $y$  axis and the neuron proportion on the right  $y$  axis. Day 3 graph (top left) shows amines (NH<sub>2</sub>, diNH<sub>2</sub>, L-diNH<sub>2</sub>), cyano, thiol (SH) and hydroxyl (OH) to have similar and low neuron density but different neuron proportion. The best performance from synthetic environments for this time point is from diamine and amine environments with a similar finding from previous work (41). Interestingly, the biological control (P/LAM) environments showed slightly higher neuron density and second highest neuron proportion. The environment with highest neuron proportion is methyl (CH<sub>3</sub>) but it also comes with the highest neuron density meaning neurons migrated the least in these environments. 3-methoxy and carboxyl (COOH) are the lowest performers and this outcome is expected. These environments have the most acidic termination as show by their pKa value 4.5.

At the later time point (day 7) graph (middle top), diamine (diNH<sub>2</sub>) is the best performer with lowest neuron density but also has the lowest neuron proportion. This could be from neuronal death, more differentiation to astrocytes and/or higher astrocyte proliferation. A



similar trait is observed with long diamine (I-diNH<sub>2</sub>) although with slightly higher neuron density. Amine (NH<sub>2</sub>) here is the best balance between lower neuron density and proportion after biological environments (P/LAM). Cyano environments perform better than expected for both time points. Although thiol (SH) environments were not great at day 3, they did well in day 7 and methyl environments did even better with similar neuron density and higher neuron proportion. Carboxyl (COOH) is the lowest performer with highest neuron density. This means neurons did not migrate as much compared to other environments.

Carbomethoxy on the early time point (day 3) has among the second lowest performance from the remaining synthetic environments whereas propamine has the best on day 7. The biological control has higher neuron density on day 3 than the new synthetic environments but the lowest on day 7 due to the massive cell cluster area it promotes. The order of cell density performance for day 3 is *prop – NH<sub>2</sub> < CBM < butylNH<sub>2</sub> < P/LAM* and for day 7 is *P/LAM < NH<sub>2</sub>prop < butyl – NH<sub>2</sub> < CBM*. For cell proportion, the best synthetic performer is once again propamine for day 3 and very similar to the biological control on day 7. The performance order for this cell parameter at day 3 is *P/LAM > NH<sub>2</sub>prop > CBM > butylNH<sub>2</sub>* and for day 7 is *P/LAM > NH<sub>2</sub>prop > butylNH<sub>2</sub> > CBM*.

#### 4.2.2.3 Astrocyte density and proportion

There are two types of astrocytes, where type I has fibroblast-like morphology and type II has spindle-like morphology. They are robust glial cells that play several roles in the central nervous system. They manage chemical signals (neurotransmitters) exchanged by neurons,

strengthen neuron connections (synapses) (356) called long-term potentiation (275) among other functions.

We are after lower astrocyte density for all cell types because higher cell density means smaller extracellular volume and amount suggesting cells use the resources in the vicinity quicker (272,273). In addition, low cell density promotes cell differentiation and internal cell signalling for changes within individual cells (autocrine signalling). On the other hand, high cell density promotes cell-cell communication inducing changes in nearby cells (paracrine signalling) which is undesirable (274).

Cell proportion tells us about differentiation (day 3) and proliferation (day 7). Generally *in vitro*, astrocyte type I cells dominate cultures compared to neurons but the degree of dominance can inform on actin stress (357). We know this as after central nervous system trauma, proliferative astrocytes give rise to other astrocytes (358,359). Extrapolating from this, lower type I astrocyte proportion is desirable. On the other hand, astrocytes type II are rare so increasing their proportion is desirable.

Day 3 astrocyte density and proportion leans more on informing on cell differentiation. At this stage, high density means cells reside inside the neurosphere because they are avoiding interacting with their environment. Low astrocyte density is a strong indicator of differentiation. Day 7 time-point is a good indicator of proliferation (101). In a situation where cell density is similar, but the cell cluster area is larger means astrocytes are dividing.

Here, the ideal environment would minimise cell density and astrocyte proportion at the same time as these cells dominate the cell culture environment compared to other cell types. Below are astrocyte type I images, density and proportion graphs:

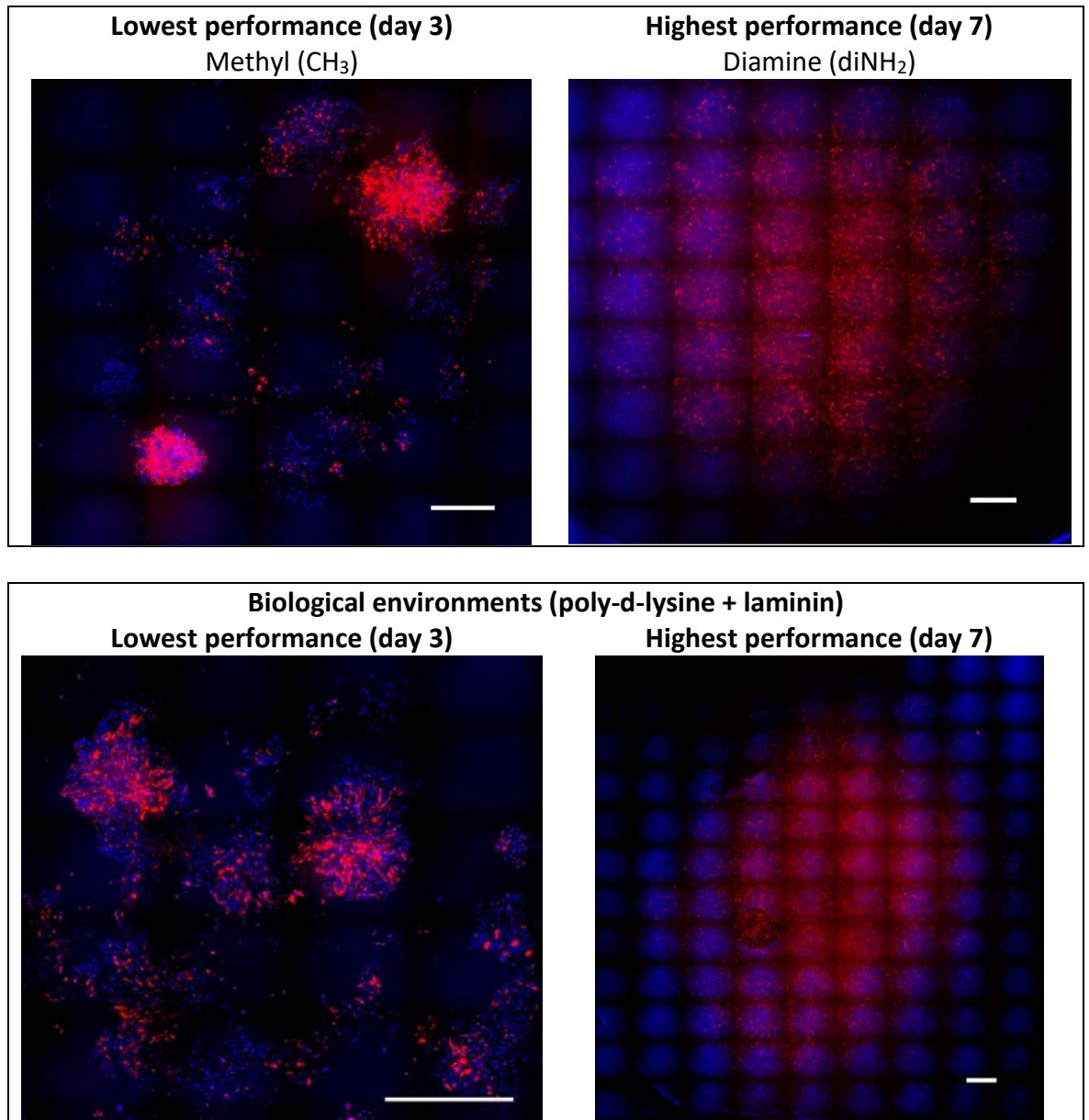


Figure 4.15: Astrocyte density and proportion images from synthetic (top row) and biological (bottom row) environments. Left columns shows the highest cell density (cells/mm<sup>2</sup>) and the right one the lowest. The more intense red the image is, the stronger the presence of astrocytes. Blue dots indicate the presence of cells (DAPI) and red material indicates astrocytes (GFAP). Scale bar (bottom right) is 1 mm.

A visual method to interpret cell density here is the red intensity and cell cluster area. The more intense the red is and the larger the cell cluster area the better. For cell proportion individually, the redder there is in the image the stronger the presence of astrocytes regardless of colour intensity (cell density). From the top left image of Figure 4.15, methyl (CH<sub>3</sub>) has the highest astrocyte density in the early time point (day 3) and diamine (diNH<sub>2</sub>) and amine (NH<sub>2</sub>) have the lowest density in the later time point (day 7). The biological control (P/LAM) on day 3 has low cell density and low proportion of astrocytes compared

to other environments. Carbomethoxy (CBM) though, has smaller cell cluster area making it the lowest performer from the remaining environments. On the later time point (day 7), the biological control (P/LAM) is the best performer with the lowest cell density evident by the lower intensity of red. Below are graphs of type I astrocyte density and proportion for all environments:

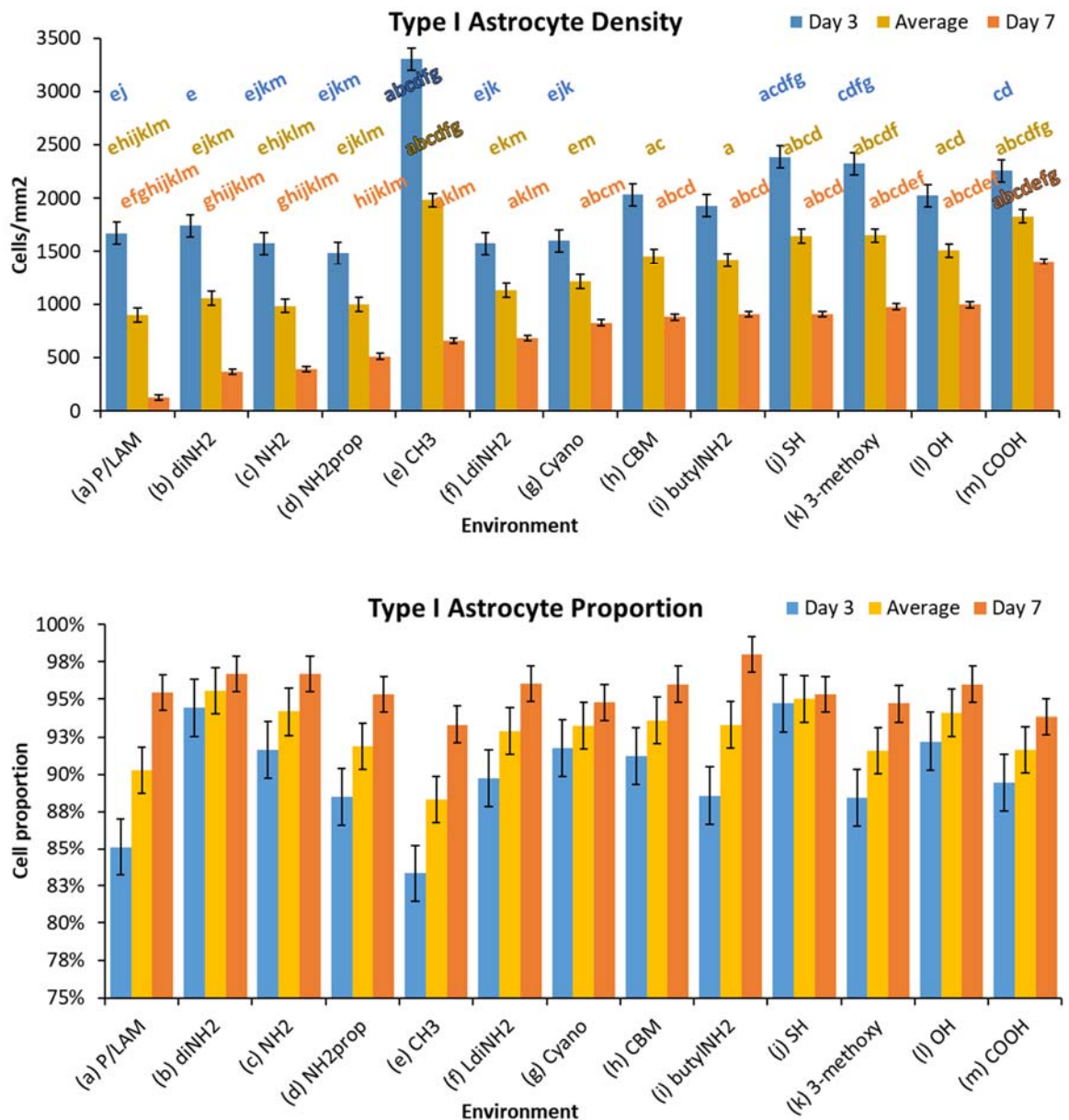


Figure 4.16: Type I astrocyte density and proportion results obtained from cell culture experiments. *y* axis is the cell density (cells/mm<sup>2</sup>) or cell proportion (%) and *x* axis are the environment sorted by lower to higher day 7 cell density for comparison. The error bars indicate the median absolute deviation and the labels on top of the bars are the significant differences between groups colour coded by time point.

Type I astrocyte density is related with type I astrocyte proportion and cell cluster area. Cell density is derived by standardising total type I astrocyte count with median cell cluster area.

This changes the relationship between the cell density and proportion visible as the horizontal distance between the green and red data points on the same  $x$  axis.

Top row graphs (Figure 4.16) show the median type I astrocyte density on the left  $y$  axis and the cell proportion on the right  $y$  axis. At the early time point (day 3) graph (top left) shows  $\text{NH}_2$ ,  $\text{l-diNH}_2$ , and Cyano environments perform similarly and well with low cell density and cell proportion. Diamine has similar cell density, but the astrocyte proportion is higher meaning there is more proliferation/differentiation to astrocytes. Biological control (P/LAM) environments have low type I astrocyte density and proportion and serve as the benchmark here as well. The environments with highest cell density terminate with methyl ( $\text{CH}_3$ ) but they also come with the lowest type I astrocyte proportion meaning there are more cells of a different type such as neurons.

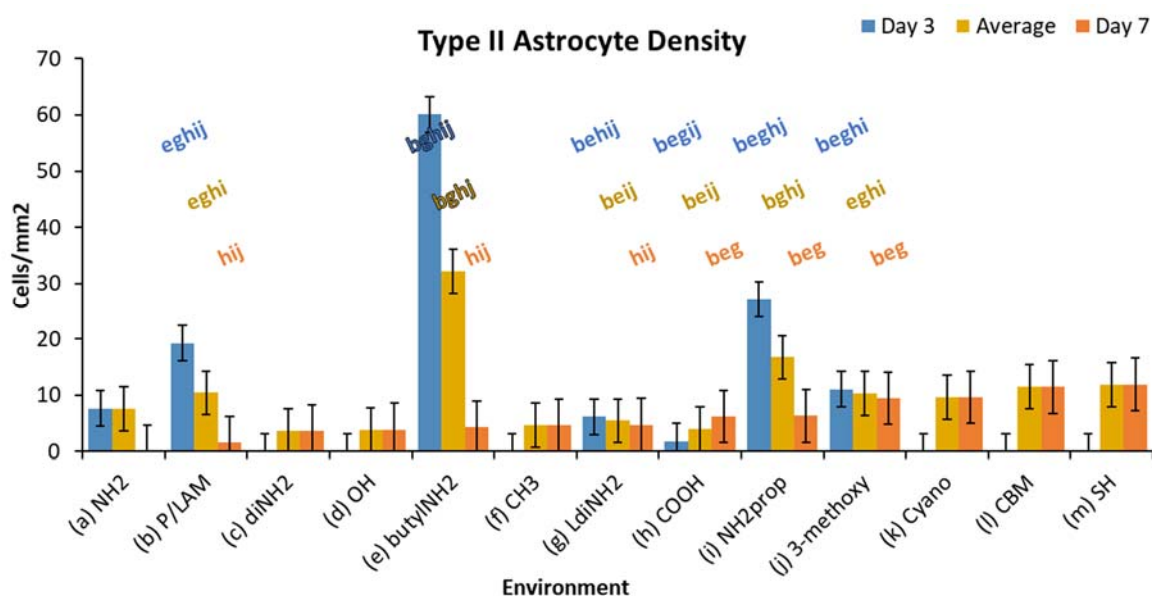
At the later time point (day 7) graph (middle top), diamine and amine ( $\text{diNH}_2$ ,  $\text{NH}_2$ ) are the best performers with lowest type I astrocyte density but they come with the highest cell proportions. This means astrocytes type I dominate the cell clusters in these environments. Methyl ( $\text{CH}_3$ ) is the best performer offering low cell density and proportion and this is unexpected. Carboxyl ( $\text{COOH}$ ) is the lowest performer with highest cell density. This means neurons did not migrate as much compared to other environments. 3-methoxy and hydroxyl ( $\text{OH}$ ) did not do well either meaning acidic terminations are not great for astrocytes type I migration.

The general trend observed is type I astrocyte density starting high on all environments shown on day 3 and drop by day 7. The type I astrocyte density on the left of the top row show propamine ( $\text{NH}_2\text{prop}$ ) as the best performer, among the new environments, having the lowest cell density for the early time point (day 3). The order of performance for this

cell performance metric is  $NH_2prop < P/LAM < butylNH_2 < CBM$ . On the later time point (day 7) the best performer is the biological control and the order of performance is  $P/LAM < NH_2prop < CBM < butylNH_2$ .

In the top right graph, the trend for type I astrocyte proportion is to start high on day 3 and go even higher by day 7. From the remaining environment, propamine has the lowest type I astrocyte proportion. The order of performance on day 3 is  $P/LAM < NH_2prop < butylNH_2 < CBM$ . For day 7, propamine has the lead again with lower cell proportion. The order of performance in this time point is  $P/LAM < NH_2prop < CBM < butylNH_2$ .

Next, we will investigate the second type of astrocyte (II) found in cultures, astrocyte type II. Here, the ideal environment would minimise cell density and maximise astrocyte type II proportion at the same time as this cell type is very rare in *in vitro* cell culture. Below are type II astrocyte graphs:



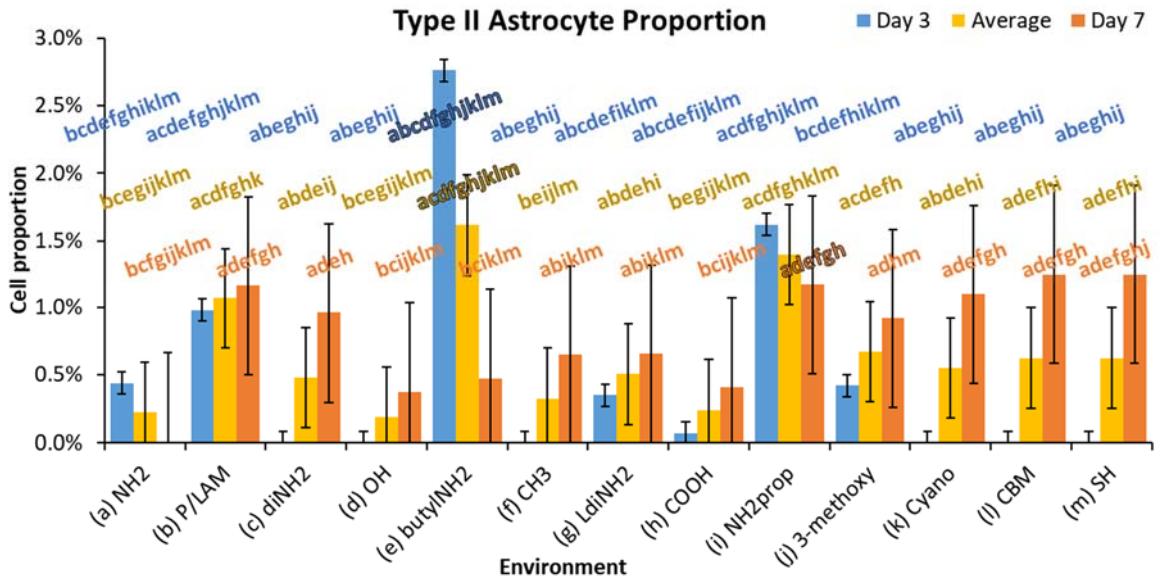


Figure 4.17: Astrocyte type II density and proportion results obtained from cell culture experiments. y axis is the cell density (cells/mm<sup>2</sup>) or cell proportion (%) and x axis are the environment sorted by higher to lower day 7 cell proportion. The error bars indicate the median absolute deviation.

Top row graphs (Figure 4.16) show the median type II astrocyte density on the left y axis and the cell proportion on the right y axis. Both of these cell parameters at zero means there were no astrocytes type II present. At the early time point (day 3) graph (top left) shows carboxyl environments having the lowest cell density but also the lowest cell proportion. There are more type II astrocytes in amine (NH<sub>2</sub>) and long diamine (II-diNH<sub>2</sub>) environments but these have slightly higher cell density. Interestingly, biological control (P/LAM) environments have the highest proportion of type II astrocyte but also the highest cell density. Except for biological environments, there are too few type II astrocytes for a meaningful value of cell density.

At the later time point (day 7) graph (middle top), diamine (diNH<sub>2</sub>) environments perform the best from the group with lowest type II astrocyte density and highest cell proportion. Long diamine (ldiNH<sub>2</sub>) and methyl (CH<sub>3</sub>) are the next best from synthetic environments but this was unexpected from the latter. On the previous time point, methyl environments had very few type II astrocytes. Hydroxyl (OH) is the lowest performer after carboxyl (COOH)

lowest cell density. The clear winner for this time point is biological control environments having both lowest cell density and highest cell proportion of this rare cell type.

The trend observed is this cell type can appear on day 3 or day 7 time-point and this explains the reason why some day 3 cell density bars are down to zero. The left graph above shows astrocyte type II density and as previously, the lower this value the better as resource consumption present in the environment is reduced (272,273). On day 3 time point, the best performer is propamine (NH<sub>2</sub>prop) from the remaining synthetic environments. ButylNH<sub>2</sub> is the worst performer from all synthetic environments tested for cell density but it also has the highest proportion of this rare cell type *in vitro*. The performance order for cell density on day 3 is  $P/LAM < NH_2prop < butylNH_2$ . Carbomethoxy is excluded from the list as no type II astrocytes were found. For day 7, the performance order for cell density is  $P/LAM < butylNH_2 < NH_2prop < CBM$ .

#### 4.2.2.4 Density and proportion of unknown type cells

“Unknown” cells did not test positive for cell type specific tags used in cell culture experiments. The cells tested positive for the generic cell nuclear marker (DAPI) meaning they are cells but their cell type is not known. In images, these cells appear their cell bodies but their nuclei is visible. They could be ependymal cells, oligodendrocytes, or neural stem cells/progenitors. The latter cell group is important to identify since these undifferentiated stem cells/progenitors can create tumours or cell proportion/migration imbalance in developed tissue (91,282).

Neural stem cells/progenitors cannot be present in transplant tissue (91,282). In fact, this cell group is one of the reasons this work focused on synthetic instead of biological



environments. It is possible to add a specific stain for these cell types but there is a limit to the number of stains used concurrently in cell experiments. Going beyond the recommended number of stains (3) increases the chance of false positives. This means cells will test positive for more than one cell specific marker (e.g. both neurons and astrocytes). The alternative is to test cells with the additional marker in repeat experiments to provide definitive answers, but this depends on resources such as time and funding.

For the purposes of this project, neural stem cells/progenitors are the most important to identify from this group as no undifferentiated cells can enter a patient's brain (91,282). It is possible to tag for nestin or sox2 both of which are neural stem cell/progenitor markers but there is a limit to the number of stains used concurrently in experiments. The more stains used the higher the chances of cross-reactivity meaning false positive binding of stains. In other words, cells will appear positive for cell types they do not belong to and the experiment would have been ruined. The alternative is to test cells with other stains in additional experiments to provide definitive answers, resources permitting.

Since unknown type cells could be stem cells or progenitors, reducing the proportion of unknown type cells is desirable. Reducing cell density is also desirable as this increases the chances of cell differentiation to neurons or glia. Below are unknown type cell density and proportion graphs:

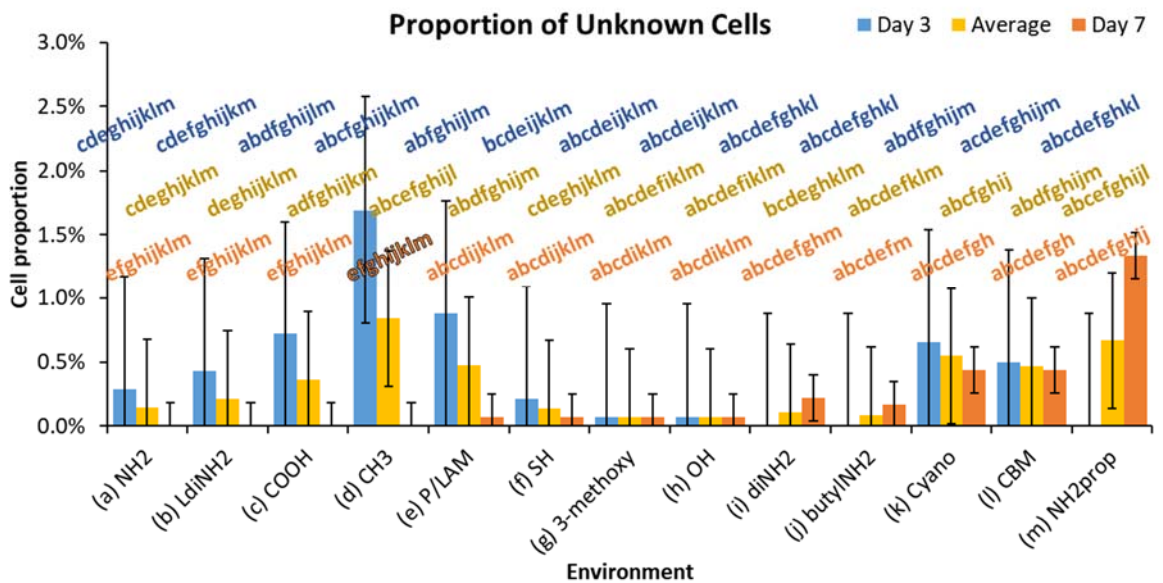
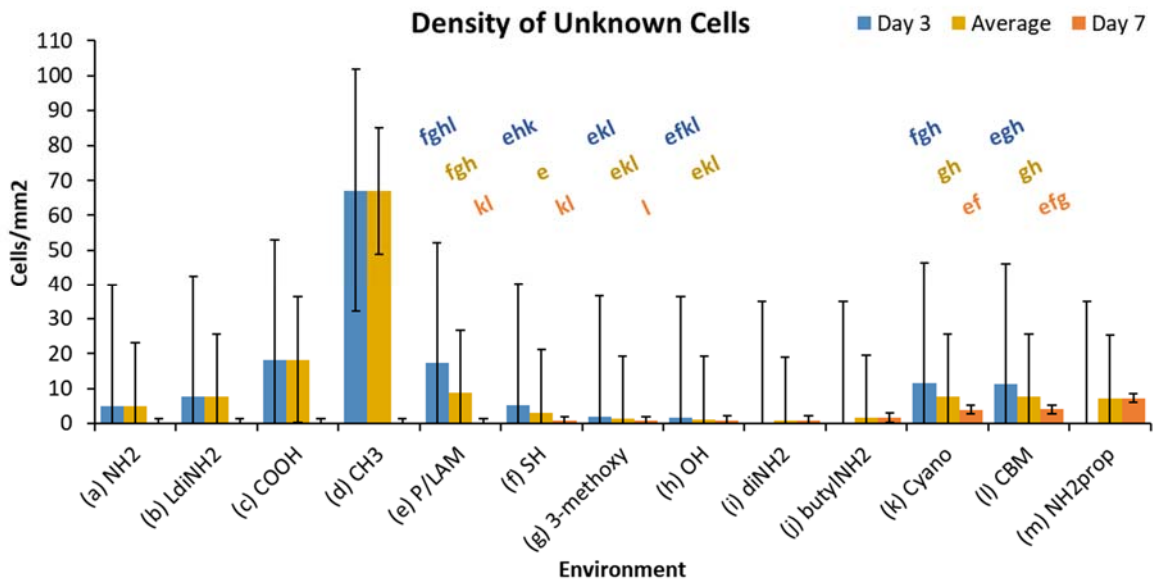


Figure 4.18: Density of unknown type cells and proportion results obtained from cell culture experiments. *y* axis is the cell density (cells/mm<sup>2</sup>) or cell proportion (%) and *x* axis are the environment sorted by lower to higher day 7 cell proportion. The error bars indicate the median absolute deviation.

Top row graphs (Figure 4.18) show the median unknown type cell density on the left *y* axis and the cell proportion on the right *y* axis. For cell density and proportion both being at zero means unknown type cells are not present. At the early time point graph (top left) the best performer is diamine (diNH<sub>2</sub>) with no unknown type cells meaning cells differentiate in these environments. Very close in performance are hydroxyl (OH) and 3-methoxy environments. The lowest performer is methyl (CH<sub>3</sub>) with the highest unknown type cell

density and cell proportion. Interestingly, biological control (P/LAM) is the second lowest performer.

At the later time point graph (middle top), carboxyl (COOH), long diamine (l-diNH<sub>2</sub>) and amine is the best performers with lowest cell density and cell proportion. A similar trend is observed in methyl (CH<sub>3</sub>) and biological control (P/LAM) environments with unknown type cells from the early time point were differentiating to neurons or astrocytes. Surprisingly, diamine at this time point has a small population of unknown type cells. This means the tiny proportion of stem cells/progenitors from the early time point have proliferated. Cyano environments perform the lowest at this time point with highest cell proportion compared to other environments.

As shown in the bottom graphs, overall amine, 3-methoxy, hydroxyl (OH), thiol (SH) provide the lowest unknown type cell density and proportion. Diamine environments follow the same trend but there are also signs of stem cell/progenitor proliferation at the later time point. The lowest performers overall are cyano environments with the same trend in both time points and following are methyl (CH<sub>3</sub>) environments. For the latter environment, cells have differentiated on day 7 despite the higher cell density, compared to other environments.

#### **4.2.2.5 Neurite length**

Neurons projections are electrically conductive and can extend to large sections of nerve tissue. The longer the neurites the better as this provides material to work with therefore increasing the potential of re-wiring damaged circuitry in the damaged tissue.

Neurites are the fine projections outwards from the neuron body. The longest of the neurites usually connecting on other neurons is called the axon. Neurite measurements were taken for 100 neurons per surface from clearly labelled cells (tuj1) with the entire neurite length visible (41). Below are cell images of smallest and largest neurites from synthetic and biological environments:

Here, the ideal environment would maximise neurite length to connect to neighbouring cells and communicate across large sections of tissue for neural circuitry rewiring. Below are cell images of smallest and larger neurites from synthetic and biological environments and after these follow neurite length graphs for all environments:

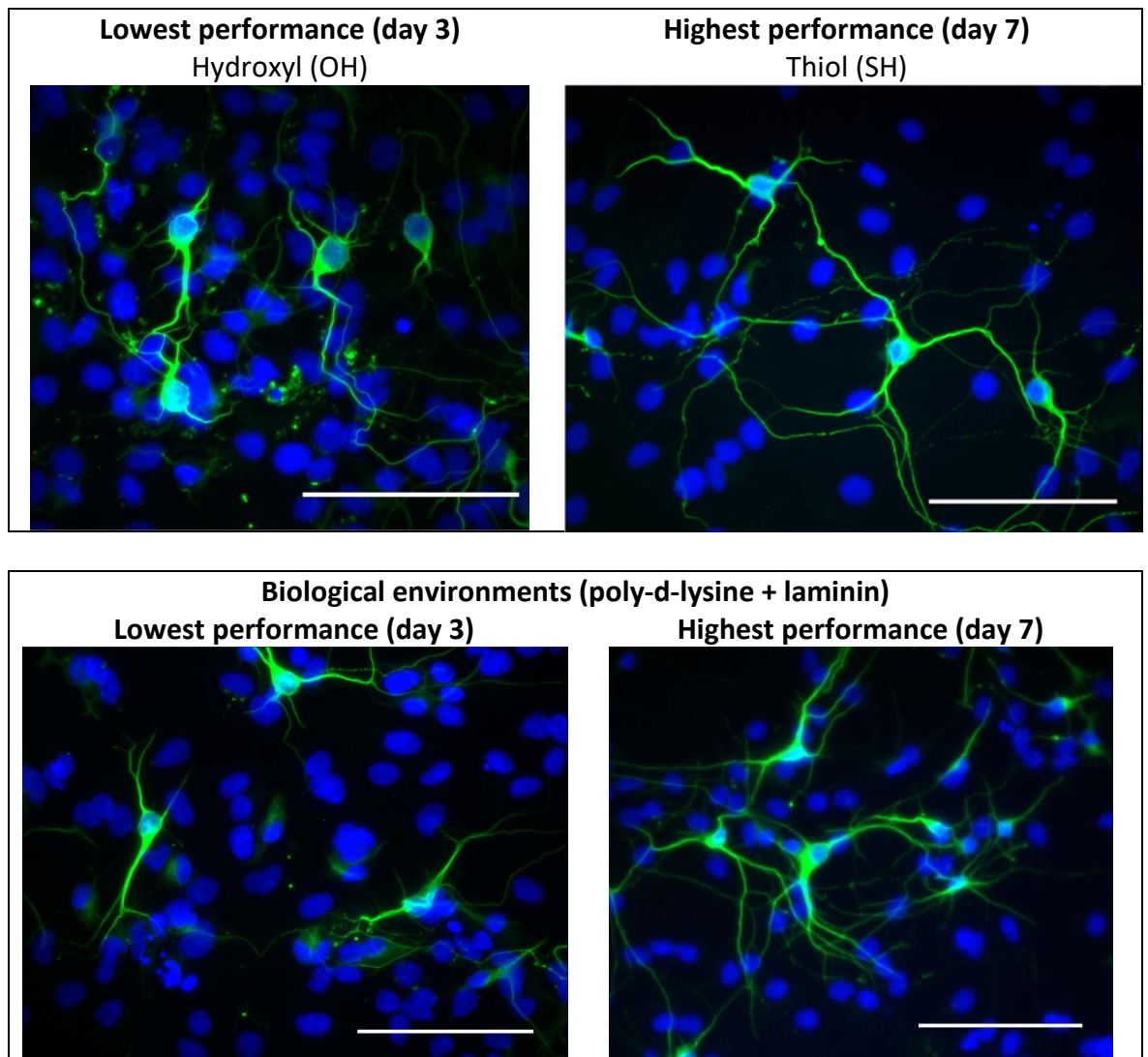


Figure 4.19: Neurite length images from synthetic (top row) and biological (bottom row) environments. Left columns show the shortest neurites and the right one with the longest. Blue dots indicate the presence of cells (DAPI) and green material indicates neurons and neurites ( $\beta$  III tubulin). Scale bar (bottom right) is 100  $\mu$ m.

Neurites are the projections outwards from the neuron body (in green). The longest projection of a neuron usually connecting (synapses) to another cell body is the neuron axon. Neurite measurements were taken for 100 neurons per surface from clearly labelled cells (tuj1) with the entire neurite length visible (41). From the images above (Figure 4.19), the top row shows the shortest neurites recorded in methyl ( $\text{CH}_3$ ) environments in the early time point (day 3). In the top right image are the longest neurites present in thiol (SH) environments in the latter time point (day 7). The biological environments (P/LAM), scored similarly in both time points.

From the remaining environments, carbomethoxy (CBM) has the shortest neurons on day 3 while propamine (NH<sub>2</sub>prop) has the longest on day 7 from all environments (3x laminin's). The bottom row images show neurons and neurites from biological environments for comparison. For this kind of environment, the neurite length was similar. Below are graphs of neurite lengths from all environments used in the study:

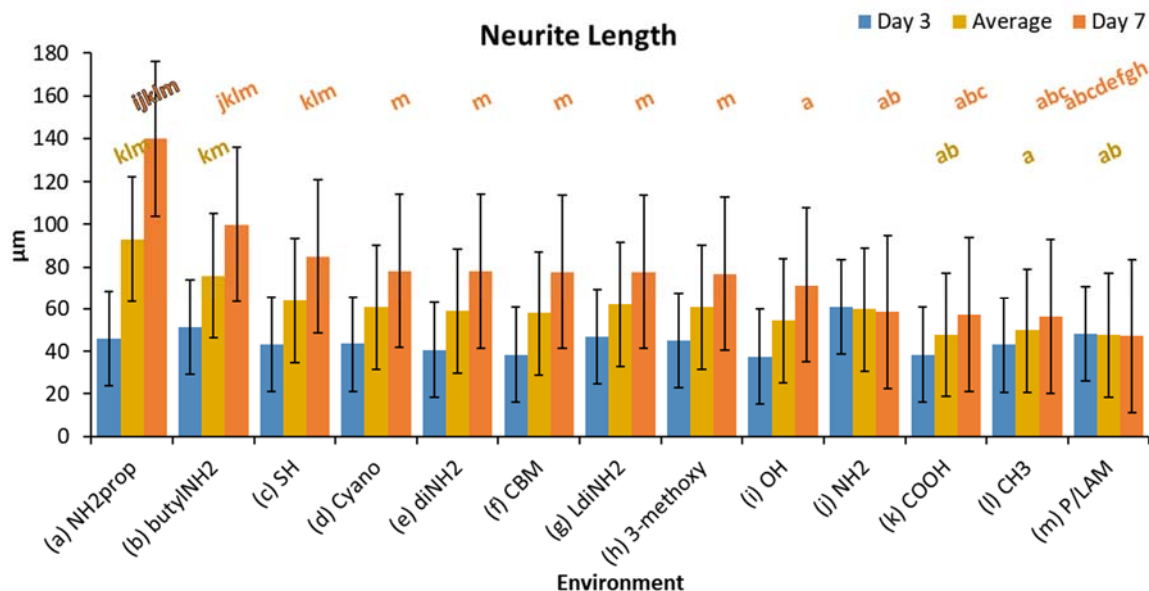


Figure 4.20: Neurite length graphs. These results show the median neurite length of environments used in experiments. *y* axis is the length in  $\mu\text{m}$  and *x* axis are the environments sorted by their longest to shortest processes from the later time point (day 7). The error bars indicate the median absolute deviation.

Graphs in Figure 4.20 show the neurite length on the *y* axis and the environments used in the study on the *x* axis sorted by longest to shortest neurites of the later time point (day 7). Day 3 graph (top left) shows amine (NH<sub>2</sub>) environments having the longest neurites. The lowest performer from the group are carboxyl (COOH) and hydroxyl (OH) environments both considered hydrophilic and acidic (pKa 4.5) compared to the rest. Biological (P/LAM) environments have medium neurite length.

At the later time point (day 7) graph (middle top), thiol (SH) is the best performer with longest neurites. Neuron density and proportion for this environment is mediocre compared to other synthetic environments (Figure 4.14).

The graph above shows neurite length on the two time-points and their average. For the new environments, day 3 performance is in this order *butylNH2* > *NH2prop* > *CBM* and day 7 performance is *NH2prop* > *butylNH2* > *CBM*. CBM sits between the two (new) amine environments on average. The only significant difference is between day 7 propamine and the rest of the day 3 scores.

#### 4.2.2.6 Type I astrocyte area

Astrocyte spreading is related with fibre length as astrocytes extend protrusions to interact with other cells and with the surface for migration and attachment (93). Astrocytes interact with themselves, other glial cells and neurons (194). In biological environments with laminin, astrocytes spread more and migrate towards more permissive ECM regions (360). In another study (357), astrocyte shape was found to change from stellate to spread when serum was absent in the culture. The spread means forming stress fibres and focal adhesions (due to Rho activation) because astrocytes are establishing and stabilising altered cytoarchitecture. The authors believe the shape of astrocytes modulates their interaction with neurons *in vivo*. The ideal environment will minimise both cell parameters or at least match laminin's performance. Laminin is set as threshold as enhanced astrocyte migration was observed (360).

Measurements were taken for 100 astrocytes per surface for type I astrocyte area from clearly labelled cells (GFAP) with the entire cell body or fibre length visible. Fibres were measured regardless of them being "connected" to other cells. The body of type I astrocytes is the red material surrounding the blue blob being the cell nuclei. Astrocyte

fibres for both cell types are the fine processes extended outwards from the cell body. The smaller the cell area and fibre length the better.

Since type I astrocyte area and astrocyte fibre length is related with forming stress fibres (357), the ideal environment will minimise type I astrocyte area or at least match laminin's performance. Laminin is set as threshold as enhanced astrocyte migration was observed (360). Below are images of type I astrocyte spreading and astrocyte fibre length:

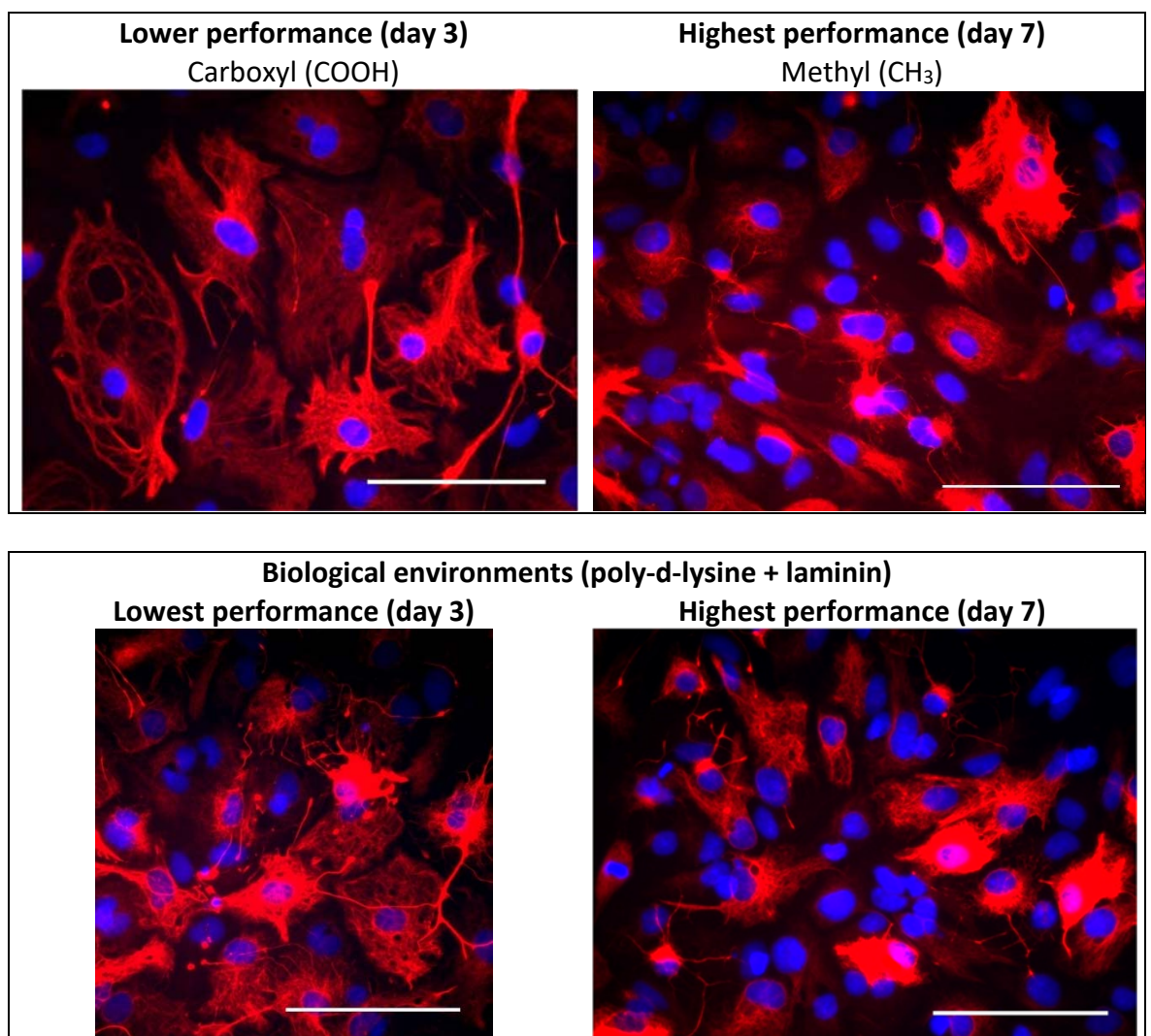
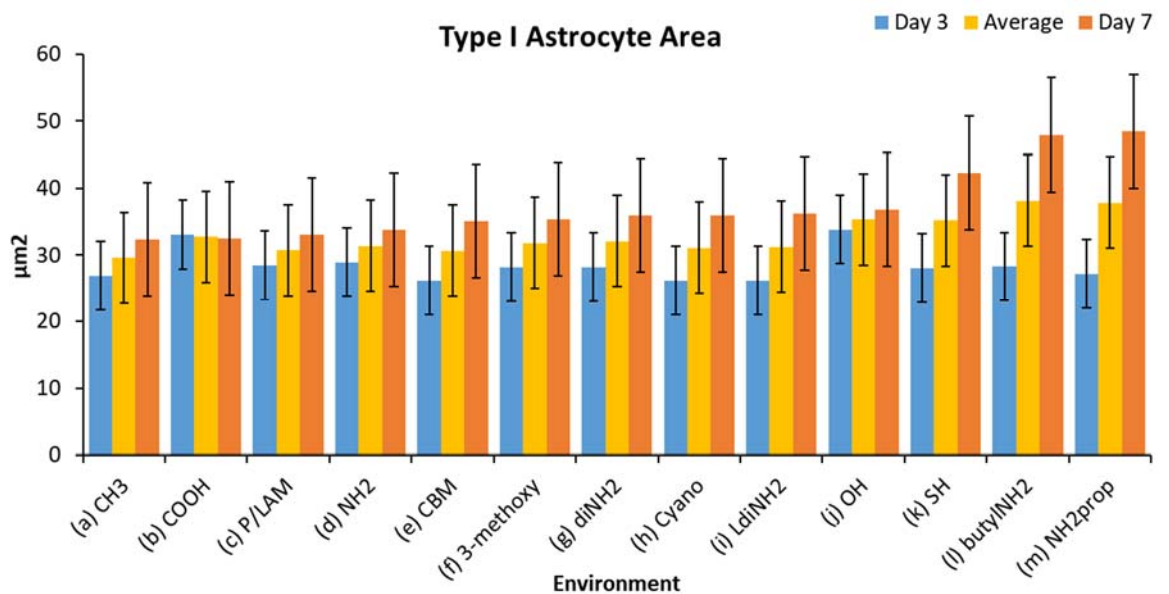


Figure 4.21: Type I astrocyte area and fibre length images from synthetic (top row) and biological (bottom row) environments. Left columns shows the largest cell areas and longest fibres ( $\mu\text{m}^2/\mu\text{m}$ ) and the right one the smallest/shortest. Blue dots indicate the presence of cells (DAPI) and red material indicates astrocytes (GFAP). Scale bar (bottom right) is 100  $\mu\text{m}$ .



Type I astrocyte area is the cell body in red, surrounding the cell nuclei (blue spots). Astrocyte fibres are the fine processes outwards from the astrocyte body. Measurements were taken for 100 astrocytes per surface from clearly labelled cells (GFAP) with the entire cell body or fibre length visible (41) regardless of them being “connected” to other cells. From the images above (Figure 4.21), the top row shows one of the lowest performer, carboxyl (COOH), with largest astrocyte area and longest fibres recorded in the early time point (day 3). In the top right image shows one of the highest performer, methyl (CH<sub>3</sub>) with the smallest astrocyte areas and shortest fibres recorded in the latter time point (day 7). The biological environments (P/LAM), scored similarly for astrocyte area in both time points but the fibre length was slightly longer in the later time point.

From the new environments, butylamine is a low performer with the highest type I astrocyte area on day 3 and carbomethoxy is the high performer with the lowest cell area on day 7. The differences between the cell performance values across environments is insignificant and this clearer in the graphs below:



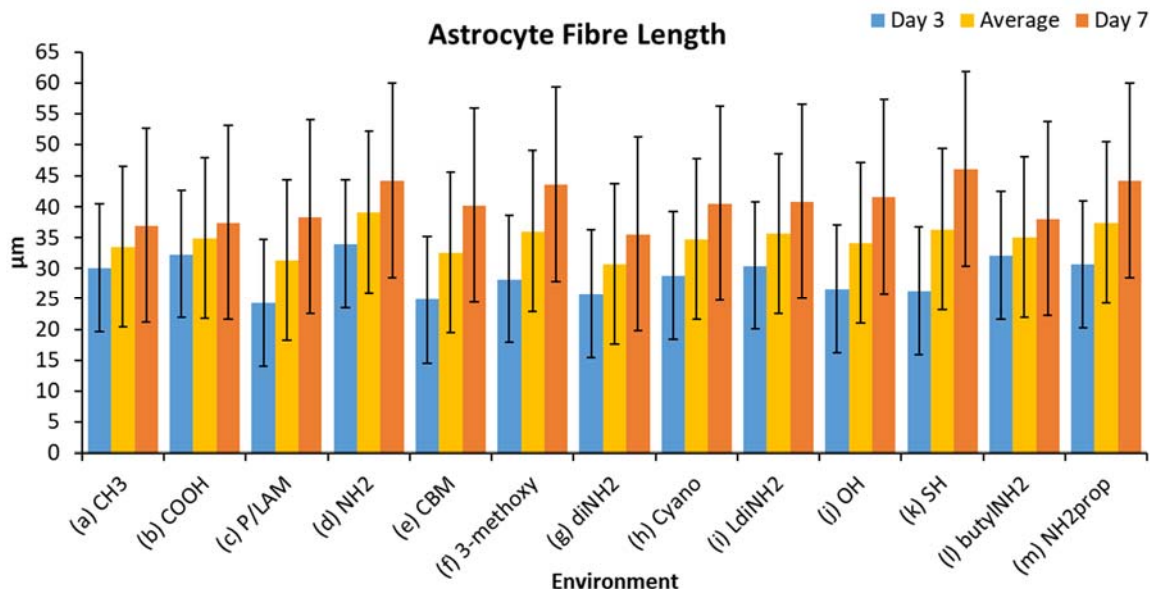


Figure 4.22: Type I astrocyte area graphs. These results show the median type I astrocyte area (top row) of environments used in experiments.  $y$  axis for astrocyte area is the area in  $\mu\text{m}^2$  and for the  $x$  axis, the environment sorted by smallest to largest cell areas from the later time point (day 7). The error bars indicate the median absolute deviation.

Graphs in the top of Figure 4.22 show the type I astrocyte area on the  $y$  axis and the environments used in the study on the  $x$  axis sorted by smallest to largest areas from the later time point (day 7). Day 3 results shows methyl ( $\text{CH}_3$ ), cyano and long diamine ( $\text{l-diNH}_2$ ) to be the best performers with smallest type I astrocyte area. The lowest performer from the group are carboxyl ( $\text{COOH}$ ) and hydroxyl ( $\text{OH}$ ) environments both considered hydrophilic and acidic ( $\text{pKa } 4.5$ ) compared to the rest. Biological ( $\text{P/LAM}$ ) environments have mediocre type I astrocyte area.

At the later time point (day 7), the best performers with the smallest type I astrocyte area are methyl ( $\text{CH}_3$ ) and carboxyl ( $\text{COOH}$ ) environments. The latter environment changing from being one of lowest performer to be the second highest is unexpected, but this is because cell area increased in other environments. It is speculated, carboxyl's molecular complexity has kept cell area to the same level as in the previous time point. Thiol ( $\text{SH}$ ) is the lowest performer with largest astrocyte area compared to the rest.

Overall thiol and hydroxyl environments are the lowest performers with largest type I astrocyte area on average for both time points. Methyl (CH<sub>3</sub>) environments perform the best also found in correlation tests. There is a –correlation with logP and type I astrocyte area at day 3 ( $r = -0.77$ ) meaning as lipophilicity increases, type I astrocyte area decreases. This cell parameter is similar throughout the time points and environments. In the later time point, some differences are observed. Longer duration in cell culture may show the environment's effect clearer for this cell parameter.

### Astrocyte Fibre Length

Bottom graph in Figure 4.22 show the astrocyte fibre length on the  $y$  axis and the environments used in the study on the  $x$  axis sorted by shortest to longest fibres from the later time point (day 7). Day 3 graph (top left) shows diamine (diNH<sub>2</sub>), thiol (SH) on par with biological environments (P/LAM) offering the best performance with shortest astrocyte fibres. The lowest performers are carboxyl (COOH) and amine (NH<sub>2</sub>) environments and for the latter, this was not expected.

In the later time point (day 7) graph, the best performers with the shortest astrocyte fibres are diamine, methyl (CH<sub>3</sub>) and carboxyl (COOH) environments. The latter environment changing from being one of lowest performers to be the second highest is unexpected, but this follows the same trend as in type I astrocyte area. Thiol (SH) and amine are the lowest performers with longest astrocyte fibres compared to the rest.

Overall diamine and methyl are the highest performers with shortest astrocyte fibres on average and amine environments perform the lowest. This was also found with a +correlation between logP and astrocyte fibre length at day 7 ( $r = 0.49$ ) meaning as lipophilicity increases so does fibre length.

Instead of interpreting the above experimental results like in the previous chapter here, we will model cell performance using machine learning techniques.

### 4.2.3 Computational cell models

Computationally models of cell responses allows to perform cell culture experiments *in silico*, thereby accelerating the process of finding better artificial environments for neural stem cell/progenitor cell culture *in vitro*. Better environments are expected to allow cells to behave similarly as they do in a biological environment.

#### 4.2.3.1 Linear regression

Linear regression is the first standard in modelling for regression problems. Table 4.1 shows the estimated coefficients for the cell cluster area model (day 7):

Table 4.1: Cell cluster area model summary. This is linear regression where the coefficients were estimated using least-squares method.

Chemical variable	Coefficient	Standard error	$P >  t $
Partition coefficient - LogP1	-5104.38	1592.10	0.00
LogP2	14289.27	4192.32	0.00
LogP3	-8528.60	2385.15	0.00
LogP4	-401.15	1583.15	0.80
LogP5	-2950.84	1991.24	0.14
Molecular mass	46.31	38.64	0.23
Molecular volume	-68.92	54.84	0.21
pKa	184.59	68.25	0.01

The  $R^2 = 0.67$  for the above model. The  $R^2$  is the proportion of the variance in the response variable, predictable from the independent variable(s). The closer this metric is to 1 the better the model fit. The mean absolute error (MAE) for this model is  $1892 \mu\text{m}^2$ .

Four variables were detected with large standard errors and failed significance tests. These are LogP4, LogP5, molecular mass, and molecular volume. This warrants investigation for collinearity.

#### 4.2.3.1.1 Collinearity

Collinearity is where two or more predictors are correlated with each other. When predictors are highly correlated, model interpretability is difficult as subtle changes in the data will provide very different regression coefficients. The simplest method to detect multi-collinearity is to examine the correlation coefficient between each pair of the predictors. Pair-wise correlation here may be sufficient, but not a necessary condition for multi-collinearity. Below is a correlation matrix with the predictor variables:

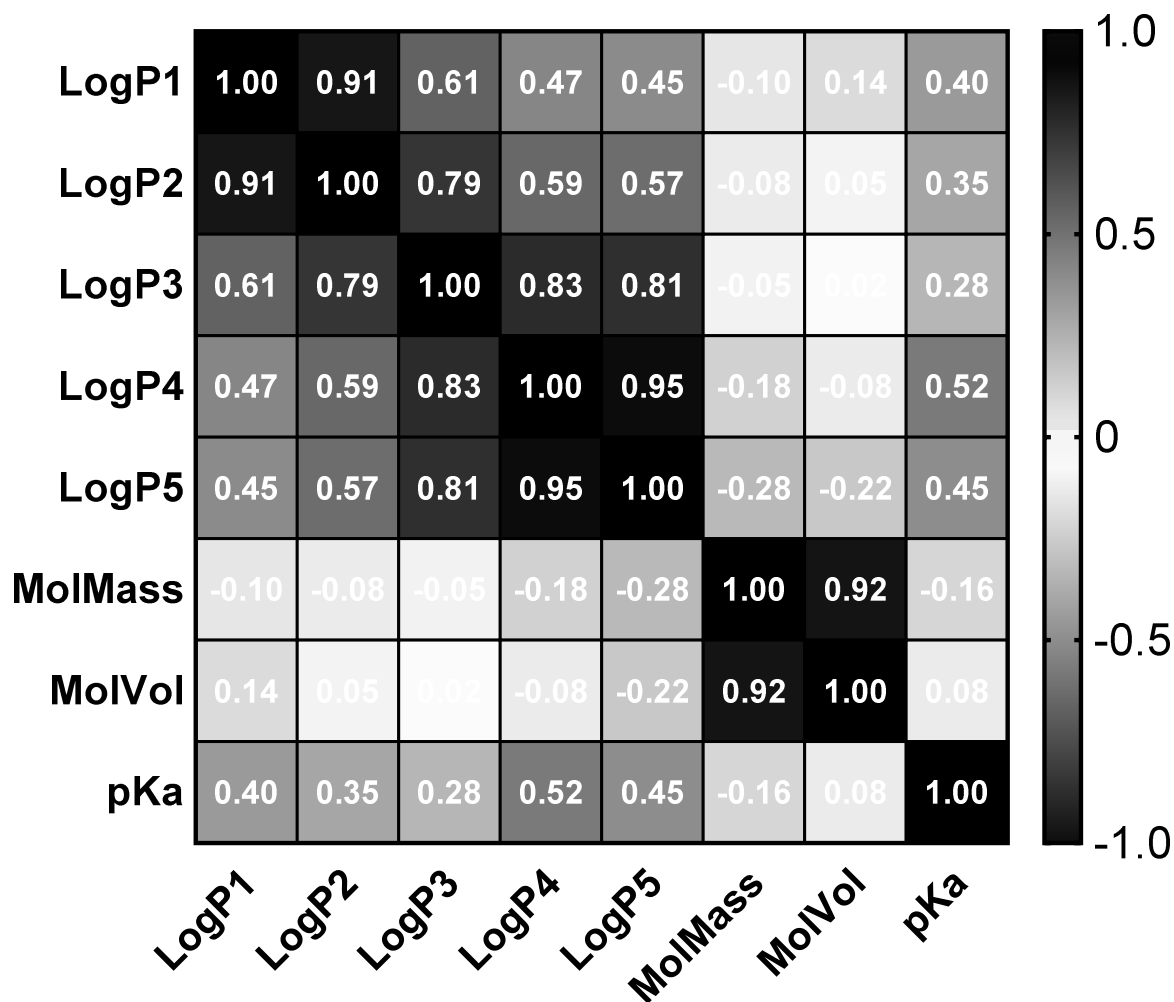


Figure 4.23: Correlation heatmap of chemical input parameters. The darker the cell the stronger the correlation positive (+) or negative (-). The diagonal set of cells with perfect (+) correlation dividing the correlation matrix from the top left to the bottom right corner can be ignored. These are correlations with parameters themselves. Very weak correlations have a white background and their correlation value is not important to interpret in this example.

As suspected, there is high collinearity between the logP group and molecular mass and volume. The correlations for some pairs of the predictors are strong. The next step is to assess the severity of multicollinearity.

#### 4.2.3.1.2 Variance inflation factor

VIF quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance (standard deviation<sup>2</sup> of the estimate) of an estimated regression coefficient is increased because of collinearity.

Below is a table with the VIF results modelling cell cluster area using linear regression:

Table 4.2: Variance inflation factor for linear regression (least squares). The variance inflation factor value indicates the magnitude of multi-collinearity. VIF also indicates the inflation of coefficient standard error compared to if they were uncorrelated (rightmost column).

<b>Chemical variable</b>	<b>VIF</b>	<b>Coefficient SE increase</b>
Partition coefficient - LogP1	7653333.27	2766.47
LogP2	53065913.60	7284.64
LogP3	17176625.96	4144.47
LogP4	7567463.88	2750.90
LogP5	11971667.26	3460.01
Molecular mass	2294700969.33	47903.04
Molecular volume	2441452450.40	49411.06
pKa	16.25	4.03

The table above indicates the two groups logP and molecular mass/volume as the cause of multi-collinearity, also found in correlation tests. This makes sense as the logP values are derived for the top 6 constituents of the same molecule and the molecular volume is related with molecular mass.

When dealing with multicollinearity, the possible avenues here include:

1. The obvious is additional data, if possible, to find which inputs are more important
2. Leave the model as is despite the multicollinearity. Multicollinearity does not affect the efficiency of extrapolation to new data assuming the same multicollinearity pattern in the new data as in the training data will be present.
3. Drop one or more collinear variables. Doing this may produce a model with significant coefficients but information may be lost. Removing such variable(s) adds bias in coefficient estimates of remaining predictors that were correlated with the dropped variable(s).
4. Use other methods that are affected less from the effect of multicollinearity e.g. random forest

Without removing any predictors  $R^2 = 0.67$ , the mean absolute error (MAE) is  $1892 \mu\text{m}^2$ . “Solving” the multicollinearity problem by removing collinear variables leaves LogP3 and pKa only and this model returns an  $R^2$  of 0.41 and mean absolute error of  $2711 \mu\text{m}^2$ . This effectively worsens the model as it now makes  $819 \mu\text{m}^2$  additional error on average and the model does not fit as well as before. Since we are interested in the predictions more than model interpretation, we chose combinations of the above solutions to multicollinearity. Solution 1, additional data, is not possible given time and cost limitations in performing cell culture experiments.

#### 4.2.3.2 Linear-regression alternatives

The following part of this section is about modelling cell responses using alternative methods. Choosing models in practice involves an iterative method to tune the hyper-parameters of learning algorithms to lower a measure of prediction error.

Selected feature selection methods and algorithms that learn single or ensemble models were investigated. The table below shows these with the number of user parameters (hyper-parameters) tested:

Table 4.3: Machine learning algorithms and user-parameters discovered in this project.

Type	Learning algorithms and user-parameters explored	
Function	Support Vector Regression (243,245)	3
Trees	One-level decision tree (257,258)	N/A
	Decision tree (235)	2
Rules	Model tree (249–251)	2
Instance based	k-nearest neighbours (240)	2
Ensemble	Ensemble of decision trees (bagging) (235)	3
<b>Support Vector Regression Kernels</b>		
	Pearson Universal kernel (247)	2



<b>Meta-methods</b>	
Locally Weighted learning (LWL) (236,237)	2
Gradient boosting with select base classifier (254)	2
Randomised ensemble with select base classifier (248)	1

<b>Feature evaluation and selection</b>	
Correlation feature subset evaluation (232)	1
Greedy search (234)	3

The complete list of algorithms and user-parameter value ranges tested are shown in 8.3 in appendices.

#### 4.2.3.3 Cross-validation model performance

Sets of learning algorithms, weighting schemes, ensemble methods and feature selection were assessed. The sets with the best model performance (defined as the lowest mean absolute error (MAE)) were selected from 10-fold cross-validation. Below is a table with the final models, their configurations, and below that follows a table with the prediction error:

Table 4.4: Machine-learning algorithms used in this work. For feature selection there must be an evaluator and a searcher typically correlation subset evaluator and greedy search. D3 and D7 stand for day 3 and day 7 time points.

<b>10-fold cross-validation model performance</b>			
Target	Feature selection	Meta-methods	Classifier
Cell cluster area	N/A	N/A	Decision tree (bagging) (235,238), num features=10, min examples = 18, allow unclassified examples
Neuron proportion	Correlation subset evaluator (232)	Locally weighted learning (236,237), weighted average, Euclidean distance	Ensemble of decision trees (235), 9 trees, min features=4
		Random feature selection (239),	k-Nearest Neighbours,

Type I astrocyte proportion		features=8, iterations=18	10 neighbours, distance= $\frac{1}{Euclidean}$
Type II astrocyte proportion	Greedy backwards search(234)	N/A	Support Vector Regression (243,245), <b>C = 0.52</b> , standardise data
			Puk (247), <b>O = 0.22, S = 2.98</b>
Proportion of unknown type cells	Correlation subset evaluator	N/A	Support Vector Regression, <b>C = 1.12</b> , normalise data
	Greedy forward search		Puk, <b>O = 0.91, S = 0.19</b>
Neurite length	N/A	Randomisable ensemble, 32 models averaged	Decision tree (bagging) (235,238), min features=10, min examples=9, holdout set=5, allow unclassified examples
Type I astrocyte area	N/A	N/A	Model tree (249) unpruned, min instances=2
Astrocyte fibre length	Correlation subset evaluator	Gradient boosting (254), 2 models	One-level decision tree (257,258)
	Greedy forward search		

The table consists of decision tree techniques, k-nearest neighbour, support vector regression and model tree. Filter feature selection methods such as correlation feature evaluation is used to remove collinear features before training models e.g. neuron proportion. This reduces the effect of multicollinearity by keeping predictors that correlate well with the response compared to others. Meta methods are used such as locally weighted learning, random feature selection, randomised ensemble of classifiers, and gradient boosting. Their purposes are to reduce the effect of outliers, prediction error

variance, and the effect of multicollinearity additionally. Below is a table with prediction error metrics of each model:

Table 4.5: Model performance from 10-fold cross-validation continued. D3 and D7 stand for day 3 and day 7. Model performance ratio is derived from the outcomes of all cross-validation iterations. A ratio closer to 0 means the closer the prediction is to the median of real outcome values and ratio of 1 means the prediction is outside of 1 standard deviation. Average real values and predictions are for all 10 chemistries for the specified cell response.

Target	Classifier	Model performance ratio		Average real values	Average predictions
Cell cluster area	Decision tree (bagging) (235)	D3	0.07	746.89 $\mu\text{m}^2$	765.22 $\mu\text{m}^2$
		D7	0.13	3650.7 $\mu\text{m}^2$	3274.69 $\mu\text{m}^2$
Neuron proportion	Ensemble decision trees (235)	D3	0.02	8.99 %	9.01 %
		D7	0.01	4.41 %	4.91 %
Type I astrocyte proportion	Randomised feature (239) k-Nearest Neighbours (240)	D3	0.25	89.84 %	90.55 %
		D7	0.15	94.60 %	94.24 %
Type II astrocyte proportion	Support Vector Regression (243,245)	D3	0.04	0.18 %	0.07 %
		D7	0.27	0.69 %	0.44 %
Proportion of unknown type cells	Support Vector Regression	D3	0.27	0.42 %	0.17 %
		D7	0.13	0.08 %	0.07 %
Neurite length	Randomisable ensemble of decision trees (bagging)	D3	0.02	48.09 $\mu\text{m}$	47.56 $\mu\text{m}$
		D7	0.02	77.20 $\mu\text{m}$	77.75 $\mu\text{m}$
Type I astrocyte area	Model tree (249)	D3	0.01	29.45 $\mu\text{m}^2$	29.54 $\mu\text{m}^2$
		D7	0.01	36.29 $\mu\text{m}^2$	36.13 $\mu\text{m}^2$
Astrocyte fibre length	Gradient boosted (254) decision trees (257,258)	D3	0.15	29.7 $\mu\text{m}$	30.98 $\mu\text{m}$
		D7	0.09	41.68 $\mu\text{m}$	40.42 $\mu\text{m}$

Each method used is detailed below along with granular cross-validation model performance graphs. Where possible, visual representations of the models are presented at the bottom of each section or appendices. The order of the algorithms described is sequential.

#### 4.2.3.4 Cell cluster area

Cell cluster area is related with cell spheres (neurospheres) spreading early after seeding them on modified surfaces, and with cell proliferation especially in the later time point (day 7). The effects in play here are both chemical and biological. Maximising the cell cluster area/neurosphere spreading is desirable as this increases cell differentiation potential.

Modelling cell cluster area was achieved with Brieman's decision tree algorithm (235) using a subset of features chosen at random (239). Decision trees are widely used in computational biology due to their accuracy and ease of interpretation. They are used in gene expression and clinical data (361) and additionally, assigning protein function and predicting splice (protein snapping) sites (362).

For cell cluster area, the maximum tree depth was set to  $depth = 6$  (tree levels). In the regression case, the mean is estimated from one part of the holdout set used (testing) and the remaining parts are used to grow the tree (train). This holdout set was set to  $N = 3$  parts. Some trees may be unable to provide an outcome referred to as unclassified instances. This was allowed with the ( $U$ ) switch. This can reduce variance in the final answer as some trees will be trained with a smaller set of data (bagging) and may not be able to provide a good answer. Below is a representation of the decision tree model for cell cluster area:

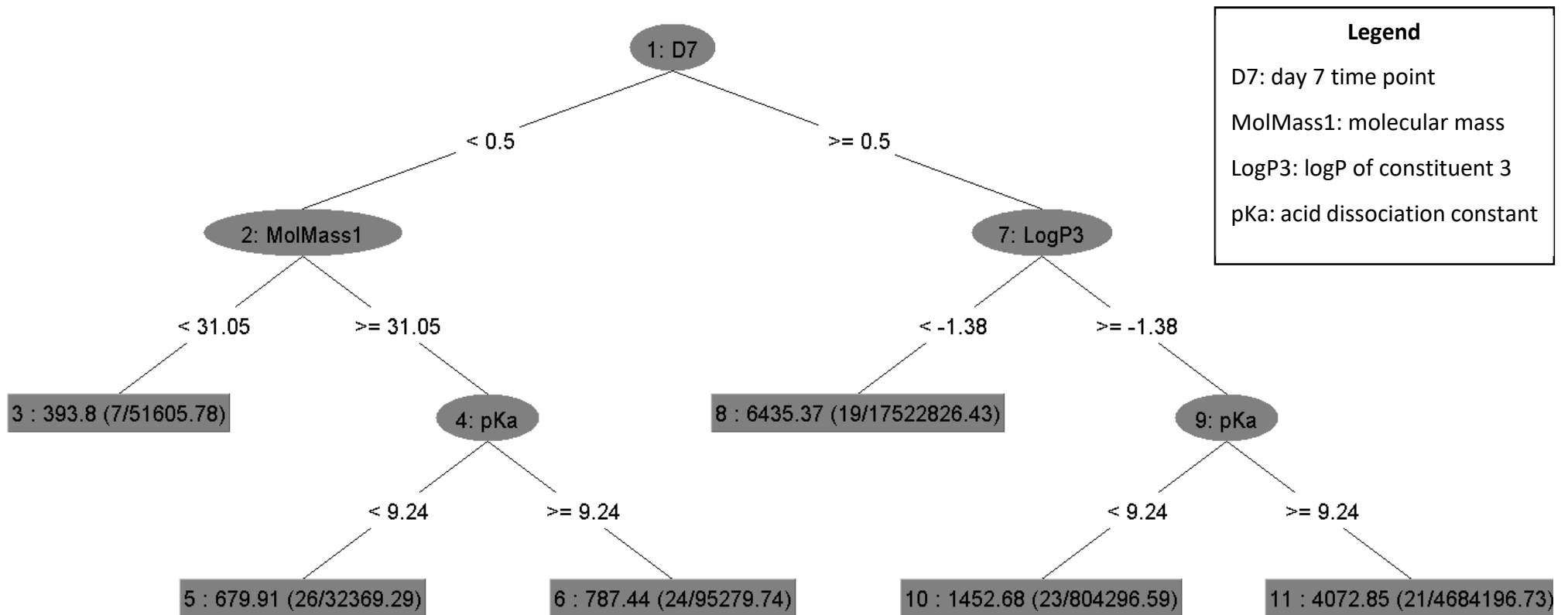
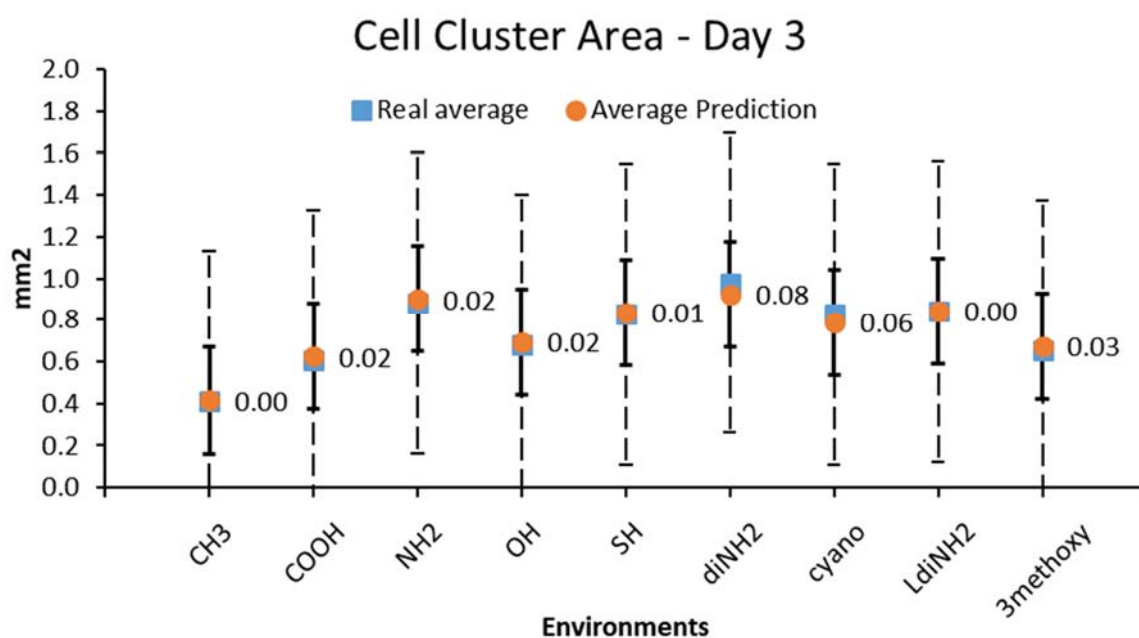


Figure 4.24: Cell cluster area model. This is a decision tree made with 4 input variables (D7, MolMass1, pKa, logP3). Nodes (in ovals) are input parameters in question for a logic test. The rectangles are the leaves and these are the possible outcomes for cell cluster area for a particular tree. The figures in parenthesis next to the value of each leaf represent the: (number of instances that reached / mean squared error in  $\mu\text{m}^2$ ).

From the figure above, the time point variable (i.e. D7) is the most important and after that follows molecular mass and the logP (lipophilicity) constituent 3 in the self-assembly molecule. The latter was found to be a good predictor in both current ( $r = -0.58$ ) and previous work (41) ( $r = -0.67$ ). This means as the surface lipophilicity increases, cell cluster area decreases. The final predictor in the model is the acid dissociation constant (pKa) that also has a strong +correlation with cell cluster area ( $r = 0.57$ ) meaning as the pKa value increases (less acidic), so does cell cluster area.

Below are graphs of cross-validation model performance where training data were split to the number of environments used in experiments (10 groups). Cell data from one environment were used for model testing. This was performed 10 times, and in each iteration, the training and test sets are different. This is to maximise the use of training data without being subject to data leak introduced by cross-validation.



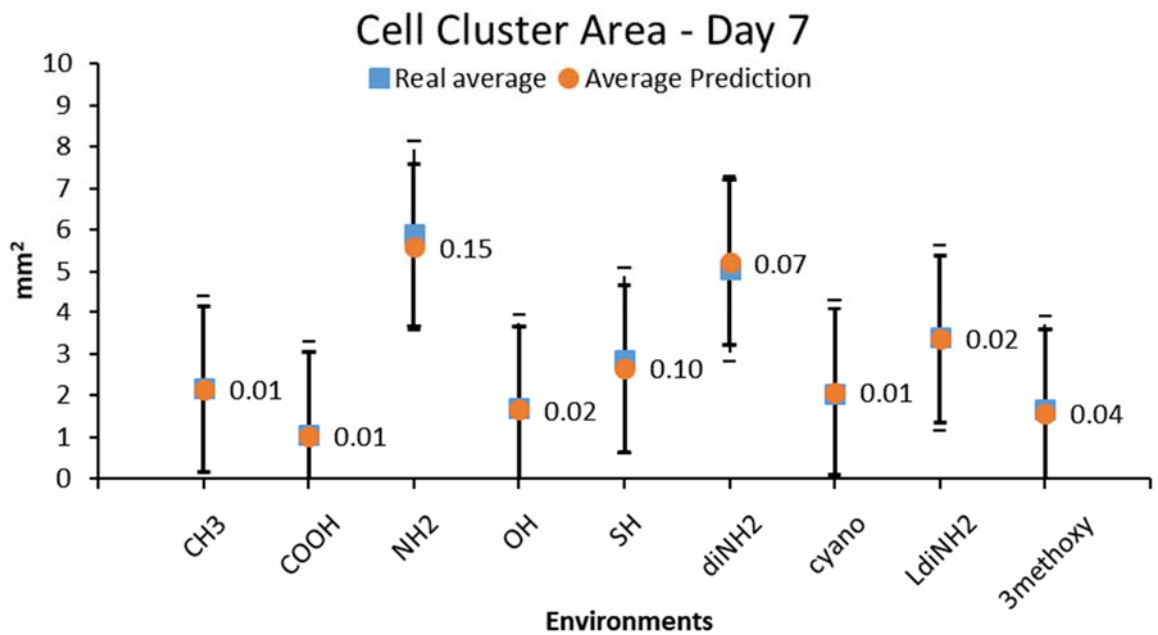


Figure 4.25: Cell cluster area day 3 and day 7 model performance from 10-fold cross validation.  $y$  axis represents the area in  $mm^2$  and in the  $x$  axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 0.02 \text{ mm}^2$  and  $MPR = 0.03$ . For the later time point (day 7),  $MAE = 0.11 \text{ mm}^2$  and  $MPR = 0.05$ . In other words, the model fit for both time points is remarkably good considering unacceptable model performance ratio 1 and the best possible ratio is 0. Even on similar chemistries such as diamine ( $diNH_2$ ) and long diamine ( $LdiNH_2$ ), the model predicts these well on both time points. Decomposing the prediction error gives bias (average error) for day 3 predictions as low as  $-1.92 \mu\text{m}^2$  and the variance (prediction standard deviation) to  $-0.25 \text{ mm}^2$ . For the later time point, the bias is  $66 \mu\text{m}$  and the variance is  $2 \text{ mm}^2$ . The closer these values are to zero the better but since these are inverse, we are after a trade-off and this is found for both time points.

#### 4.2.3.5 Neuron proportion

Successful cellular therapies to regenerate nervous tissue depend partly on the amount of neural cells delivered. Neuronal network allows function such as voluntary bodily

movement. Controlling the density and proportion of transplant relevant cell populations is a key element in developing and scaling up cell-based therapy. Neuron proportion tells us about differentiation (day 3) and proliferation (day 7).

Modelling neuron cell proportion was achieved with locally weighted learning (236,237) and ensemble of decision trees (bagging) (235). This classifier is a popular choice in genomic data analysis (363), bioinformatics (364) and life sciences (365) because models produced have high prediction accuracy and provide information on feature importance. Importance here is correlation and interactions among other features.

Below in Table 4.6, with the ranked inputs from correlation subset evaluator and greedy backwards search. After that follows a snippet of the neuron proportion model:

Table 4.6: Neuron proportion feature selection and evaluation. Backwards greedy search: started with all features then reduce one a time until there is no improvement in the merit score. The merit scores is the goodness of the remaining features in the subset after removing the features in the left.

<b>Features</b>	<b>Merit score</b>
Molecular volume	0.12
Molecular mass	0.12
Partition coefficient (logP) – level 3	0.12
Partition coefficient – level 5	0.12
Partition coefficient – level 2	0.12
Partition coefficient – level 4	0.12
Partition coefficient – level 1	0.23
Acidity measure (pKa)	0.47
Day 7	0.47
Day 3	0



RandomTree

=====

LogP4 < -0.21

```
| MolMass1 < 57.58
| | D3 < 0.5 : 4.16 (9/0.39)
| | D3 >= 0.5 : 8.06 (9/2.16)
| MolMass1 >= 57.58
| | D7 < 0.5
| | | LogP1 < -0.14 : 12.6 (10/18.42)
| | | LogP1 >= -0.14 : 7.89 (10/7.01)
| | D7 >= 0.5
| | | LogP4 < -0.37
| | | | pKa < 8.57 : 9.36 (19/25)
| | | | pKa >= 8.57 : 8.51 (10/3.17)
| | | LogP4 >= -0.37 : 6.33 (8/5.47)
```

LogP4 >= -0.21

```
| D3 < 0.5
| | LogP5 < 1.27
| | | LogP5 < 0.03
| | | | LogP2 < -0.11 : 2.13 (9/0.36)
| | | | LogP2 >= -0.11
| | | | | MolVol1 < 106.67
| | | | | | LogP4 < -0.11 : 3.07 (12/0.38)
| | | | | | LogP4 >= -0.11 : 2.9 (5/0.92)
| | | | | MolVol1 >= 106.67 : 2.56 (14/0.5)
| | | | LogP5 >= 0.03 : 4.12 (9/1.79)
| | | LogP5 >= 1.27 : 6.16 (11/0.88)
| D3 >= 0.5
| | pKa < 10.86
| | | LogP5 < 0.03
| | | | LogP4 < -0.11
| | | | | LogP3 < -0.82 : 7.19 (9/7)
| | | | | LogP3 >= -0.82
| | | | | | MolMass1 < 80.16 : 8.06 (6/0.44)
| | | | | | MolMass1 >= 80.16 : 7.9 (11/4.99)
| | | | | LogP4 >= -0.11 : 9.88 (6/4.6)
| | | | LogP5 >= 0.03 : 5.47 (8/4.97)
| | | pKa >= 10.86 : 15.99 (5/6.74)
```

Size of the tree : 37

RandomTree

=====

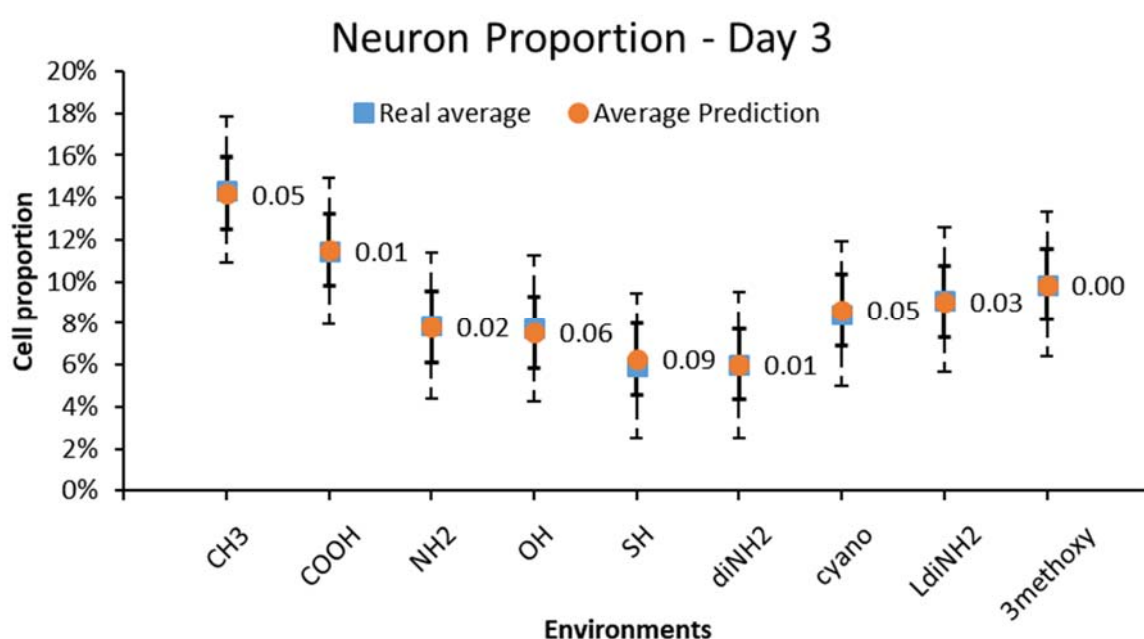
MolMass1 < 103.64

```
| D7 < 0.5
| | LogP4 < 1.54
| | | MolVol1 < 104.41
| | | | MolMass1 < 85.64
```

Snippet 1: Neuron Proportion model. Double click to expand full forest made of 9 decision trees.

From the snippet above, the top attributes appearing are LogP4, Molecular mass, time point variables, and pKa. In previous work, LogP4 was found to have a  $-$ correlation with neuron density ( $r = -0.48$ ) and the pKa with a  $+$ correlation ( $r = 0.38$ ) but the latter is not significant. From this work, we found  $-$ correlations with pKa and neuron density and proportion ( $r = -0.68$ ,  $r = -0.52$ ) agreeing with previous work (41). The logP correlations suggest that as surface lipophilicity increases, neuron density decreases. Normally, lipophilic surface means higher cell density but for neurons, but the rules are different. These cells are believed to be on top of an astrocyte carpet in *in vitro* 2D cultures (270) and surface lipophilicity may not affect their cell density as much. Methyl ( $\text{CH}_3$ ) environments could be the exception to the rule. The pKa correlation suggests as the surface pKa increases so does neuron density. Molecular mass and volume are next in importance, both found to have  $-$ correlations with neuron density  $r = -0.47$  and  $r = -0.51$  respectively. This means as the molecular mass and volume increase, neuron density decreases.

Below are graphs of model performance from cross-validation for each time point:



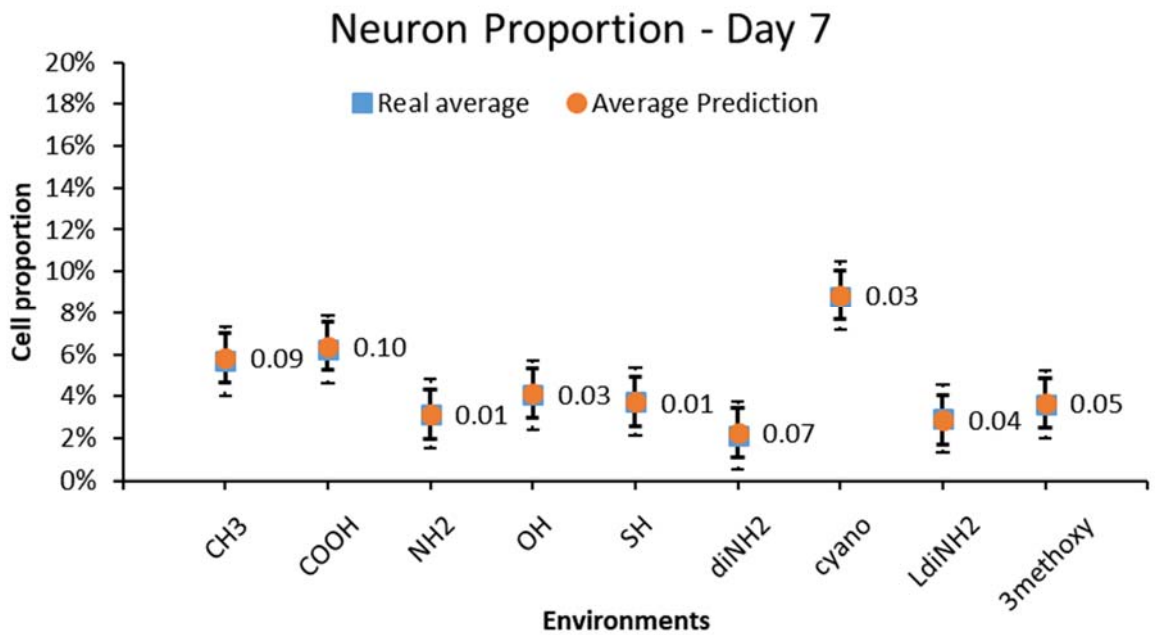


Figure 4.26: Neuron proportion day 3 and day 7 model performance from 10-fold cross validation.  $y$  axis represents cell proportion as a percentage and in the  $x$  axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 0.13 \%$  and  $MPR = 0.04$ . For the later time point (day 7),  $MAE = 0.08 \%$  and  $MPR = 0.05$ . In other words, the model fit for both time points is remarkably good considering unacceptable model performance ratio 1 and the best possible ratio is 0. Decomposing the prediction error gives bias (average error) for day 3 predictions as low as 2.4 % and the variance (prediction standard deviation) is 1.7 %. For the later time point, the bias is 1.22 % and the variance is 1.17 %. The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. This model owes its low variance to the numerous decision trees used.

#### 4.2.3.6 Type I astrocyte proportion

Astrocytes are robust glial cells that play several roles in the central nervous system. There are two types of astrocytes, where type I has fibroblast-like morphology and type II has spindle-like morphology. They manage chemical signals (neurotransmitters) exchanged by neurons, and strengthen neuron connections (synapses) (356) called long-term potentiation (275) among other functions. We are after lower astrocyte density for all cell types because higher cell density means smaller extracellular volume and amount suggesting cells use the resources in the vicinity quicker (272,273). Cell proportion tells us about differentiation (day 3) and proliferation (day 7). Generally *in vitro*, astrocyte type I cells dominate cultures compared to neurons but the degree of dominance can inform on stress (358,359). Extrapolating from this, lower type I astrocyte proportion is desirable. On the other hand, astrocytes type II is rare so increasing their proportion is preferred.

Modelling type I astrocyte cell proportion involved random feature selection and k-nearest neighbours. Instance-based methods have been used to classify DNA microarray data with remarkable model performance (366), and to evaluate biological ontologies (formal naming and definitions) (367). Random feature selection method is used (239) to select a uniform number of features  $n$  to train classifiers from the full set  $N$ . In a situation where discriminative information is spread across the features, will result to reduced correlation between predictors.

Correlation based feature selection with backwards greedy search was used to select and evaluate features to predict type I astrocyte proportion. There is no visual representation of the resulting model. Instead, sensitivity analysis of the model inputs and their effect on the cell outcome is explored in section 4.2.4. Below is a table with the ranked inputs from correlation subset evaluator and greedy backwards search:

Table 4.7: Type I astrocyte proportion feature selection and evaluation. Backwards greedy search: started with all features then reduce one a time until there is no improvement in the merit score. The merit scores is the goodness of the remaining features in the subset after removing the features in the left.

Features	Merit score
Molecular volume	0.17
Partition coefficient (logP) – level 3	0.17
Partition coefficient – level 5	0.17
Molecular mass	0.17
Partition coefficient – level 2	0.17
Partition coefficient – level 4	0.17
Partition coefficient – level 1	0.19
Acidity measure (pKa)	0.38
Day 7	0.38
Day 3	0

From this work, +correlations were found with type I astrocyte density and logP ( $r = 0.61$ ). From previous work (41), very similar outcomes are found  $r = 0.79$ . This means cell density increases as the lipophilicity increases on the culture surface. These correlations are expected as astrocytes are thought to be closer to the culture surface compared to neurons (270). LogP3 also has a –correlation with type I astrocyte proportion ( $r = -0.48$ ) meaning as surface lipophilicity increases, cell proportion decreases. Molecular volume and pKa have –correlations with astrocyte density with the new data ( $r = -0.52$  and  $r = -0.71$ ) and previous ( $r = -0.62$  and  $r = -0.61$ ). This means that as the molecular volume and pKa increase individually, type I astrocyte density decreases in both situations. There are correlations with the remaining predictors, but these are not significant.

Below are graphs of model performance from cross-validation for each time point:

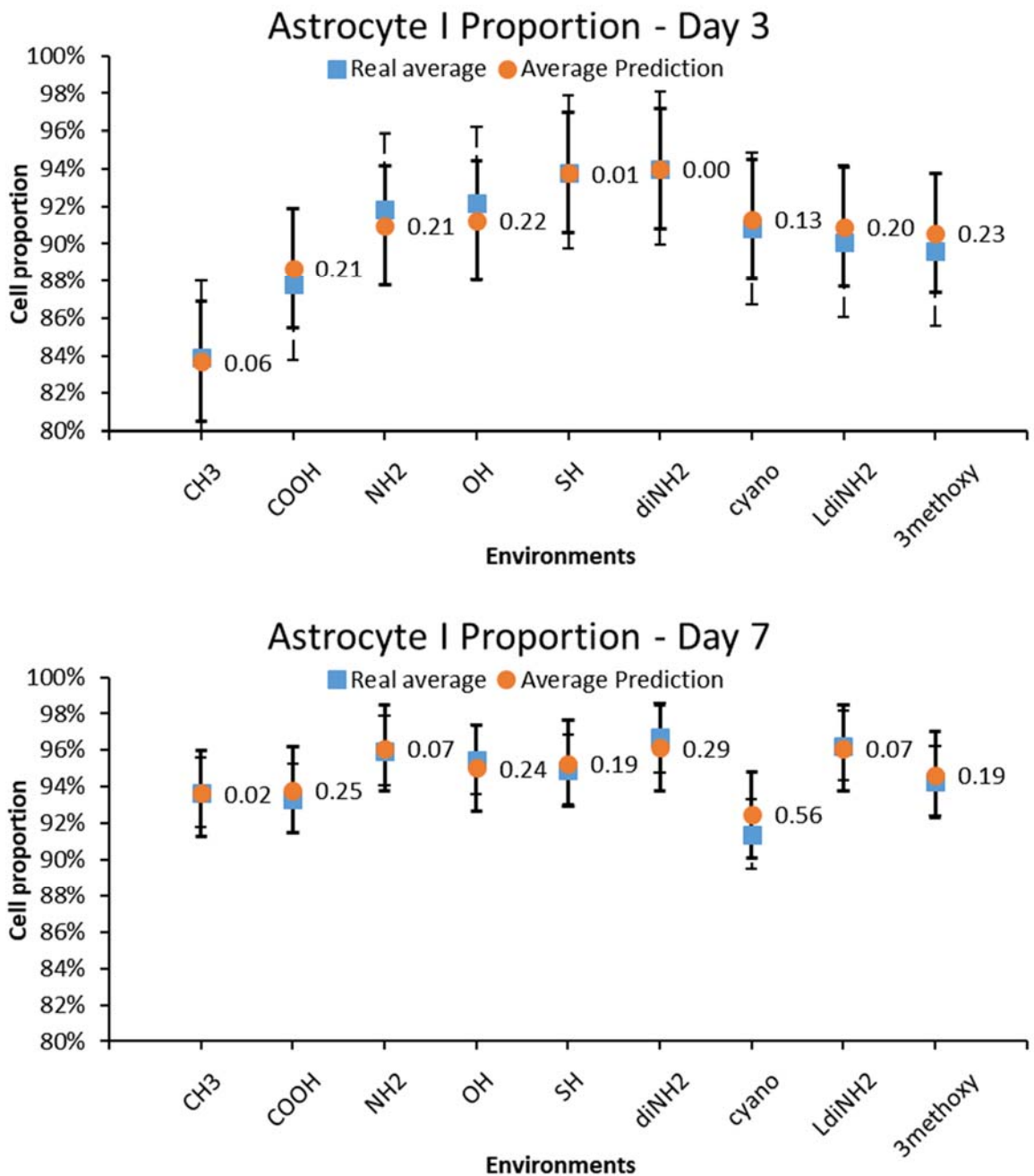


Figure 4.27: Type I astrocyte proportion day 3 and day 7 model performance from 10-fold cross validation.  $y$  axis represents cell proportion as a percentage and in the  $x$  axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 0.58\%$  and  $MPR = 0.14$ . For the later time point (day 7),  $MAE = 0.4\%$  and  $MPR = 0.21$ . In other words, the model fit for both time points is good considering unacceptable model performance ratio 1 and the best possible ratio is 0. Decomposing the prediction error gives bias (average error) for day 3 predictions as low as

0.11 % and the variance (prediction standard deviation) is 3.19 %. For the later time point, the bias is 0.26 % and the variance is 3.86 %. The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. The variance of predictions for the latter time point is sometimes outside the standard deviation of real values but this can be reduced with additional data.

#### 4.2.3.7 Type II astrocyte proportion

Astrocytes are robust glial cells that play several roles in the central nervous system. They manage chemical signals (neurotransmitters) exchanged by neurons, strengthen neuron connections (synapses) (356) called long-term potentiation (275) among other functions. Astrocytes were discussed in the previous section to this one. In short, astrocytes type II are rare *in vitro* so increasing their proportion is desirable.

Modelling type II astrocyte cell proportion was achieved with support vector regression (SVR). SVR has been successfully used to model biological data in bioinformatics such as protein function prediction and gene expression among others (368). It can deal with biological variation and generalise well on new data.

Unfortunately, SVR is a “black box” algorithm and the resulting model is difficult to interpret. This is because of the kernel trick transforming data to a higher dimension before fitting a linear model. This means the original values of support vectors are not shown. In a later chapter, the output of this model will be investigated by tuning one input at a time for their effect on cell performance. This is termed as sensitivity analysis (369) and this can be found in section 4.2.4. The features chosen to pass for predicting type II astrocyte

proportion were selected with correlation-based feature selection and backwards greedy search:

Table 4.8: Type II astrocyte proportion feature selection and evaluation. Backwards greedy search: started with all features then reduce one a time until there is no improvement in the merit score. The merit scores is the goodness of the remaining features in the subset after removing the features in the left.

Features	Merit score
Molecular mass	0.27
Partition coefficient (logP) – level 1	0.27
Molecular volume	0.27
Partition coefficient – level 2	0.27
Partition coefficient – level 4	0.27
Partition coefficient – level 3	0.27
Acidity measure (pKa)	0.27
Partition coefficient – level 5	0.40
Day 7	0.40
Day 3	0

From this work, significant correlations were found between the logP (lipophilicity), type II astrocyte density and proportion. LogP has +correlations with cell density and proportion ( $r = 0.54$  and  $r = 0.51$ ) meaning as the logP value increases so does the density and proportion of astrocytes. From one hand, we want to maximise the proportion of type II astrocytes but also minimise cell density. Perhaps achieving this is with a paracrine effect from high cell density. Surface acidity measure (pKa) was found to +correlate with type II astrocyte proportion ( $r = 0.44$ ) although not significant. In other words, this means as the pKa value increases (less acidic) so does the proportion of astrocytes.

Below are graphs of model performance from cross-validation for each time point:



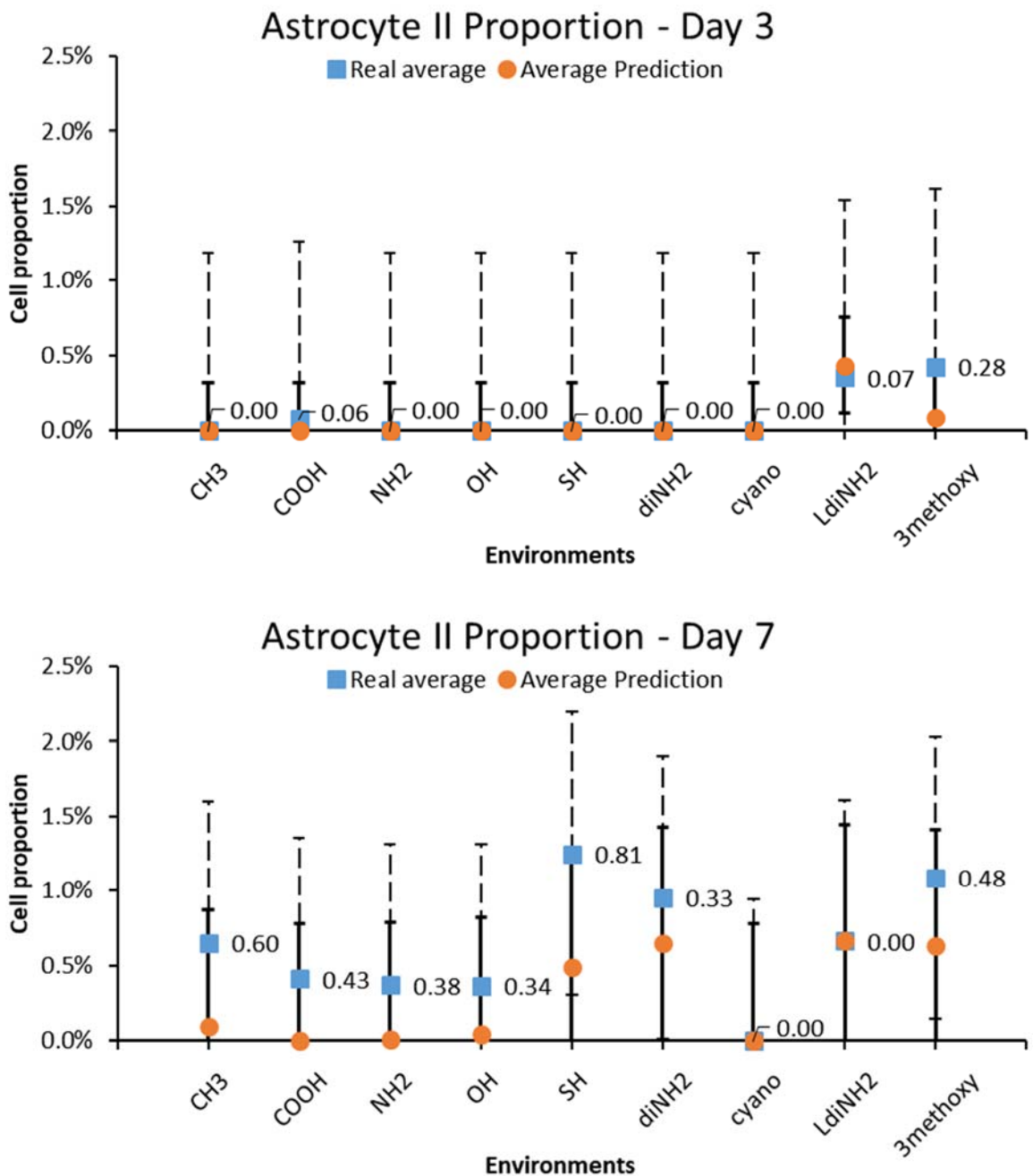


Figure 4.28: Type II astrocyte proportion day 3 and day 7 model performance from 10-fold cross validation. *y* axis represents cell proportion as a percentage and in the *x* axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 0.05\%$  and  $MPR = 0.05$ . For the later time point (day 7),  $MAE = 0.35\%$  and  $MPR = 0.37$ . In other words, the model fit for the early time point is excellent and for the later time point the fit is good. Decomposing the prediction error gives bias (average error) for day 3 predictions as low as  $-0.04\%$  and the variance (prediction

standard deviation) is 0.31 %. For the later time point, the bias is -0.35 % and the variance is 0.77 %. The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. Astrocytes type II are rare in cultures and the models appear to fit well with low bias and variance but there is room for improvement with additional data for this cell parameter.

#### 4.2.3.8 Proportion of unknown type cells

Unknown type cells are cells that did not test positive for the markers (tags) used in experiments. In other words, these cells are unidentified of type but we know they are present as their nuclei tested positive (DAPI) and they are visible in cell images. These cells could be neural stem cells/progenitors, oligodendrocytes, ependymal cells or microglia. In the worst-case scenario, unknown type cells are assumed as neural stem cells/progenitors making copies of themselves therefore minimising their proportion is desirable. This is because undifferentiated cells cannot enter a patient's brain in a transplant therapy (91,282).

The same variant of support vector regression (246) was used to model unknown type cell proportion with the universal Puk kernel both described in the previous section (4.2.3.7). Support vector regression (SVR) regularisation constant was set to  $C = 1.12$ . This determines the trade-off between the model complexity and the amount up to which deviations larger than  $\varepsilon$  are accepted. As previously, the insensitive loss function was set to  $\varepsilon = 0.001$ . This time, each attribute was standardised to have zero mean and unit variance with  $x' = \frac{x - \bar{x}}{\sigma}$  (where  $\sigma$  is the sample standard deviation). Puk kernel's omega parameter was set to  $O = 0.91$  and the sigma was set to  $S = 0.19$ .

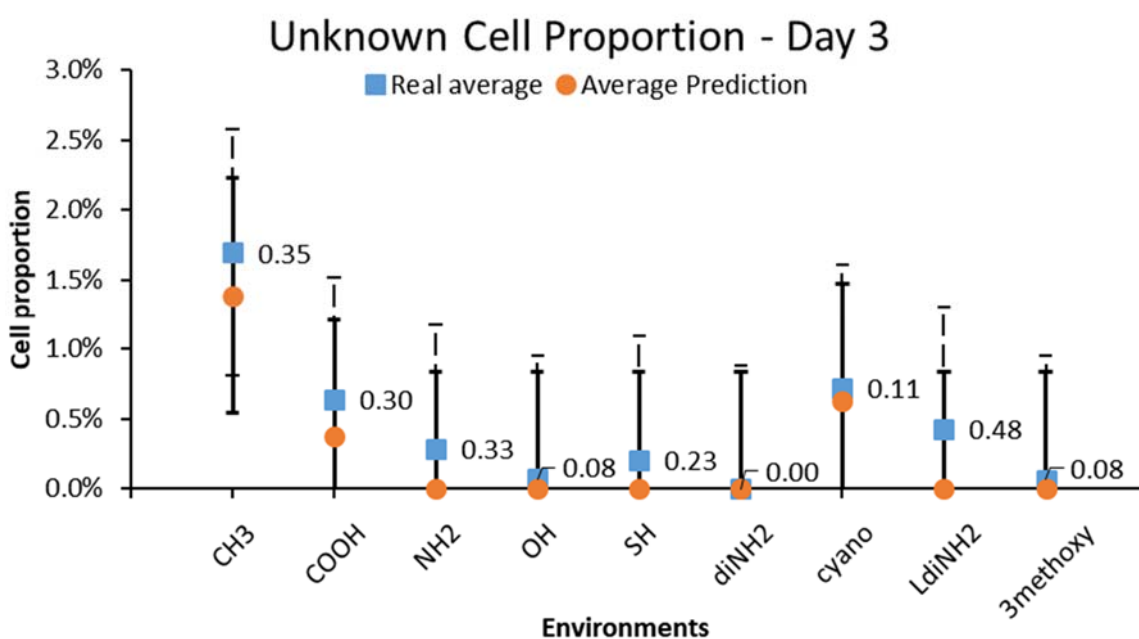
As previously, the output cannot be interpreted easily. In section 4.2.4, sensitivity analysis (369) is performed to investigate each feature for its effect with the target class. Features used to predict this cell parameter are selected using correlation and forward greedy search. The features include:

Table 4.9: Proportion of unknown type cells feature selection and evaluation. Forwards greedy search: started with no features and added one at a time until there is no improvement in modelling accuracy. The merit score is the goodness of the subset after adding the corresponding feature in the left.

Features	Merit score
Acidity measure (pKa)	0.15
Day 3	
Day 7	

From this work, the acid dissociation constant has a strong +correlation with unknown type cell proportion ( $r = 0.75$ ). This means as surface acidity decreases, unknown type cell proportion increases with it. The pKa and the day indicators are the best predictors from the group for unknown type cell proportion.

Below are graphs of model performance from cross-validation for each time point:



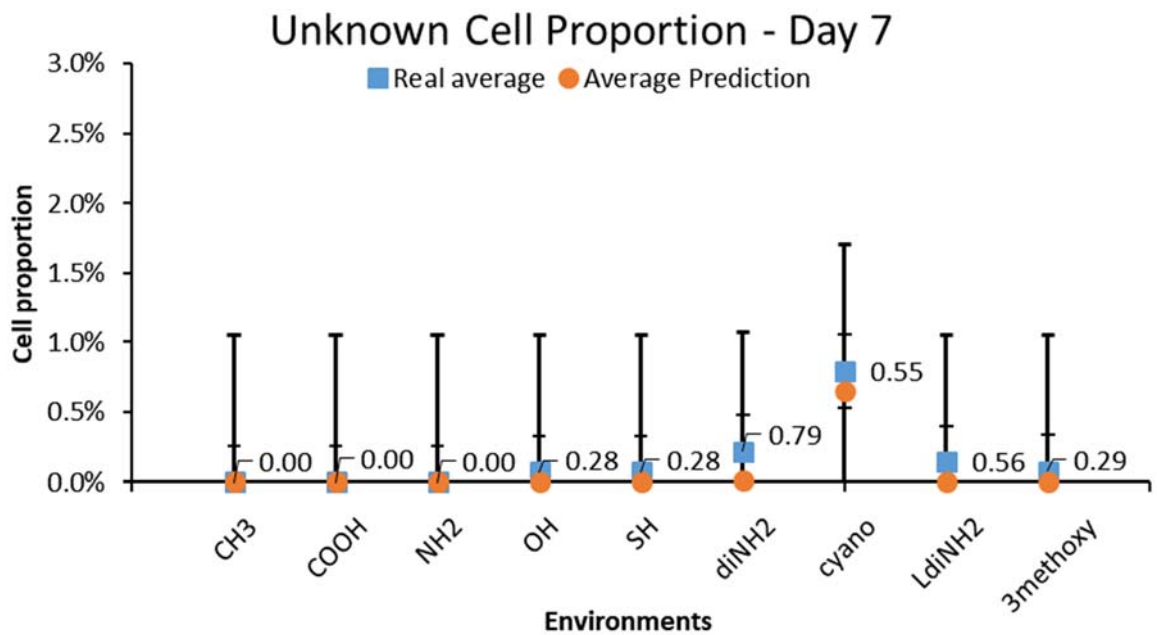


Figure 4.29: Proportion of unknown type cells day 3 and day 7 model performance from 10-fold cross validation.  $y$  axis represents cell proportion as a percentage and in the  $x$  axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 0.19\%$  and  $MPR = 0.22$ . For the later time point (day 7),  $MAE = 0.08\%$  and  $MPR = 0.31$ . In other words, the model fit for both time points is good. Decomposing the prediction error gives bias (average error) for day 3 predictions at  $-0.19\%$  and the variance (prediction standard deviation) is  $0.83\%$ . For the later time point, the bias is  $-0.25\%$  and the variance is  $1.05\%$ . The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. Prediction variance for the later time point is larger than the standard deviation of real values. As with astrocytes type II, unknown type cells are few in numbers and the models appear to fit well with low bias but prediction variance can be decreased with additional data for this cell parameter.

#### 4.2.3.9 Neurite length

Functionary nerve tissues consists of neural projections (neurites or axons) to communicate with neighbouring cells using electrical conduction across large sections of tissue. Neurite length is a good indicator of this in artificial environments (*in vitro*). One aim of neuro-regenerative biomaterials is to grow and guide neurons to specific injury areas and re-wire compromised neural circuit and restore function. Increasing neurite length is desirable in order to connect to neighbouring cells and communicate across large sections of tissue.

Modelling neurite length is achieved here with randomisable ensemble of decision trees. Each tree was created from a subset of features selected at random (239) but, unlike bagging (random forest), the number of subsets (holdout sets) was not equal to the number of trees. In the regression case, the mean is estimated from one part of the holdout set used (testing) and the remaining parts are used to grow the tree (train). The benefit of this approach is that we can use deeper trees and still reduce the variance in the final answer. On top of that, this approach taps into the discriminative information spread across the features resulting in reduced correlation between estimators. This led to small improvements in predictive performance over Random Forest by reducing the mean absolute error in 10-fold cross-validation.

The random ensemble iteration was set to 32 ( $I = 32$ ) for the equivalent number of decision trees. The decision tree algorithm was configured to choose from all features available ( $K = 10$ ). Controlling overfitting was achieved by limiting the number of instances reaching a leaf (weight) was set to ( $M = 7$ ). To get the specialised trees, the maximum depth of the trees to unlimited ( $depth = 0$ ). In the regression case, the mean is estimated from one part of the holdout set used (testing) and the remaining parts are used to grow the tree (train). This holdout set was set to  $N = 5$  parts. As previously mentioned,

some trees may be unable to provide an outcome referred to as unclassified instances. This was allowed with the (*U*) switch. This can reduce variance in the final answer as some trees will be trained with a smaller set of data and may not be able to provide a good answer. Below is a visual model representation:

## RandomTree

=====

D7 < 0.5

```
| LogP2 < 0.78
| | MolMass1 < 405058.55
| | | LogP4 < -0.37
| | | | LogP1 < 0.21 : 0 (0/0)
| | | | LogP1 >= 0.21 : 43.06 (8/66.08)
| | | | LogP4 >= -0.37
| | | | LogP1 < -1.02 : 32.54 (8/73.52)
| | | | LogP1 >= -1.02 : 43.84 (7/17.41)
| | MolMass1 >= 405058.55 : 68.23 (8/142.59)
| LogP2 >= 0.78
| | LogP1 < 1.2 : 54.27 (7/166.59)
| | LogP1 >= 1.2
| | | MolMass1 < 58.62 : 0 (0/0)
| | | MolMass1 >= 58.62
| | | | LogP5 < -0.69 : 38.1 (7/230.35)
| | | | LogP5 >= -0.69 : 41.6 (13/118.54)
```

D7 >= 0.5

```
| LogP1 < -1.07 : 52.84 (12/127.52)
| LogP1 >= -1.07
| | MolVol1 < 93.46
| | | LogP2 < 1.38
| | | | LogP3 < -0.16 : 70.54 (8/109.53)
| | | | LogP3 >= -0.16 : 64.75 (6/300.48)
| | | | LogP2 >= 1.38 : 46.36 (8/90.81)
| | MolVol1 >= 93.46
| | | MolVol1 < 104.41
| | | | pKa < 9.91 : 78.87 (7/425.58)
| | | | pKa >= 9.91 : 98.53 (12/212.72)
| | | MolVol1 >= 104.41
| | | | LogP2 < 0.96 : 85.8 (7/451.29)
| | | | LogP2 >= 0.96 : 62.75 (8/2411.52)
```

Size of the tree : 33

## RandomTree

=====

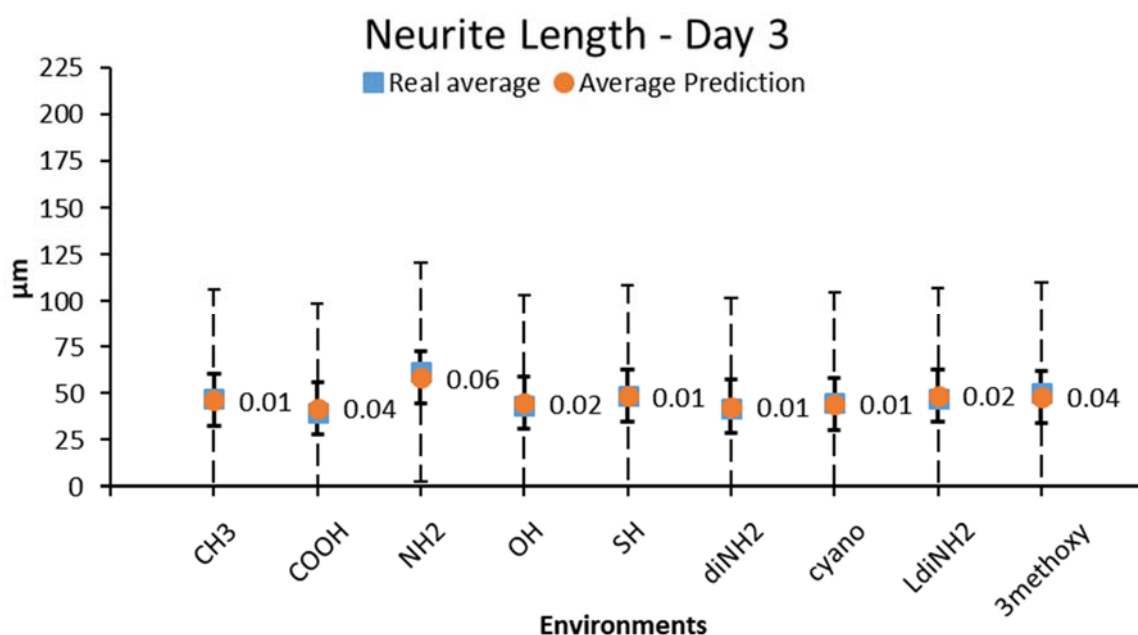
D7 < 0.5

```
| LogP2 < 0.78
| | LogP4 < -0.37
| | | LogP3 < -568 : 59.43 (4/120.58)
| | | LogP3 >= -568
| | | | LogP3 < -0.32 : 36.76 (8/178.35)
| | | | LogP3 >= -0.32 : 45.26 (8/66.77)
| | LogP4 >= -0.37
| | | LogP2 < -0.75 : 48.84 (7/33.34)
```

Snippet 2: Neurite length model. Double click to expand full set of trees made of 32 randomised decision trees.

The top variables appearing in the decision trees are assumed to be the more important ones. These include the time point indicator, logP (lipophilicity), molecular mass and volume of the untethered surface chemistry, and surface acidity (pKa). From this work, the logP has a +correlation but not significant ( $r = 0.37$ ) and from previous work (41) the -correlation is a strong one ( $r = -0.81$ ). The difference in the relationship between the two studies is attributed to different sampling methodology. Previous work measured mature neuron axon length whereas in this work, neurites were measured, that is all protrusions from neurons including “immature” ones. From this and previous work, molecular mass and volume were both found to +correlate with neurite length ( $r = 0.41$  and  $r = 0.46$ ) although these are not significant. This means as molecular mass and volume increase, neurite length increases. Lastly, the pKa +correlates with neurite length ( $r = 0.51$ ) from this work and from the previous work as well ( $r = 0.48$ ) although the latter is not significant. This means as surface acidity decreases, neurite length increases.

Below are graphs of model performance from cross-validation for each time point:





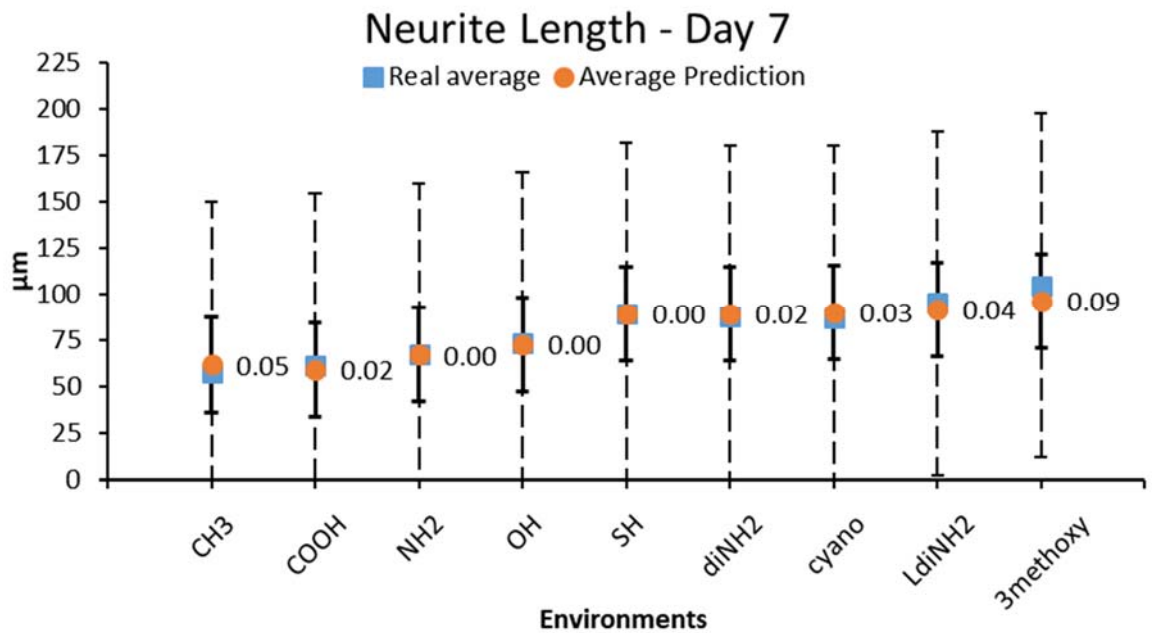


Figure 4.30: Neurite length day 3 and day 7 model performance from 10-fold cross validation.  $y$  axis is cell projection length in  $\mu\text{m}$  and in the  $x$  axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 1.43 \mu\text{m}$  and  $MPR = 0.02$ . For the later time point (day 7),  $MAE = 2.64 \mu\text{m}$  and  $MPR = 0.03$ . In other words, the model fit for both time points is excellent. Decomposing the prediction error gives bias (average error) for day 3 predictions at  $-0.23 \mu\text{m}$  and the variance (prediction standard deviation) is  $14.02 \mu\text{m}$ . For the later time point, the bias is  $-0.6 \mu\text{m}$  and the variance is  $25.41 \mu\text{m}$ . The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. Prediction variance for both time points is small due to the numerous decision trees used.

#### 4.2.3.10 Type I astrocyte area

Astrocyte spreading is related with fibre length as astrocytes extend protrusions to interact with other cells and with the surface for migration and attachment (93). Astrocytes interact with themselves, other glial cells and neurons (194). Astrocyte spreading means forming

stress fibres and focal adhesions (due to Rho activation) because astrocytes are establishing and stabilising altered cytoarchitecture (357). Minimising both type I astrocyte area and fibre length is preferred, and laminin's performance sets the upper boundary.

Modelling type I astrocyte area was achieved with model tree algorithm. Model trees have been used in toxicological and epidemiological studies due to their flexibility and power, and the derived models have informed experts from both fields (370). They are also used as a tool to classify new proteins to structural families (371).

The  $N$  switch disables pruning the trees generated and the parameter determining the minimum number of instances to create a leaf node was set to  $M = 16$ . Below is Snippet 3 with the output of model for type I astrocyte area consisting of 20 rules:

Rule: 1  
 IF  
     D3=0 <= 0.5  
     LogP1 <= -0.67  
     LogP1 <= -568.355  
 THEN  
  
 A1A =  
     0.0019 \* LogP1  
     - 0.0288 \* pKa  
     + 0.9811 \* D3=0  
     + 31.5456 [9/55.953%]

Rule: 2  
 IF  
     D3=0 <= 0.5  
     LogP1 <= -0.67  
     LogP1 <= -1.075  
 THEN  
  
 A1A =  
     -0.5293 \* LogP1  
     - 0.0257 \* MolVol1  
     - 0.0415 \* pKa  
     + 1.1231 \* D3=0  
     + 34.9074 [9/63.713%]

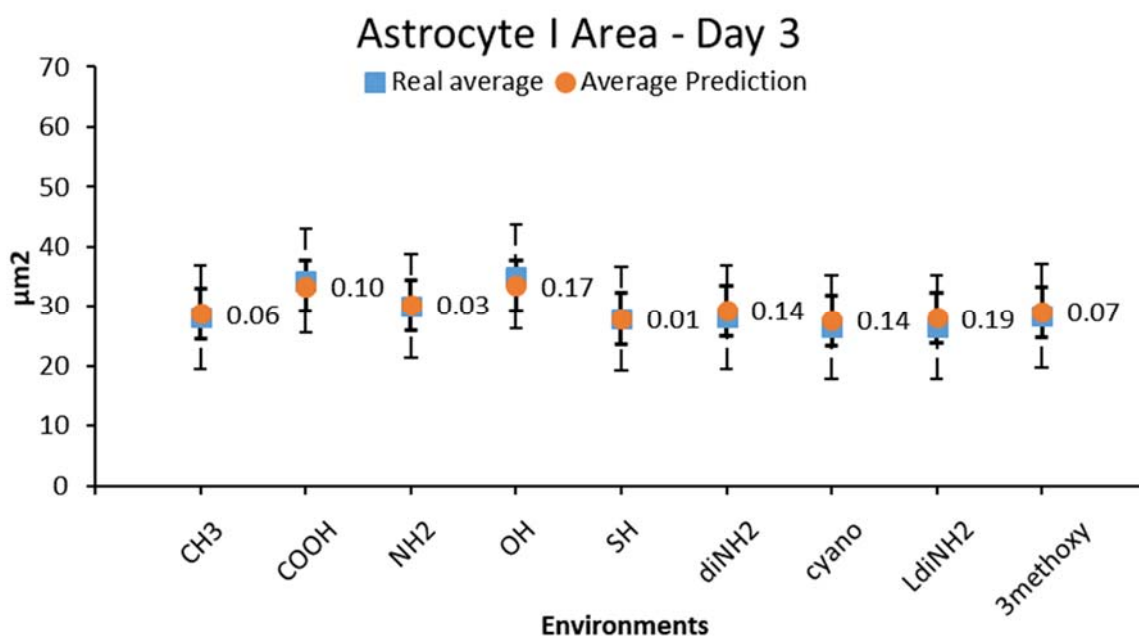
Rule: 3  
 IF  
     D3=0 <= 0.5  
     MolMass1 <= 52.59  
     LogP1 <= 0.55  
 THEN  
  
 A1A =  
     -1.2061 \* LogP1  
     - 0.0216 \* MolMass1  
     + 1.3428 \* D3=0  
     + 33.1024 [9/64.828%]

Rule: 4  
 IF  
     D3=0 <= 0.5  
     pKa <= 10.655  
     LogP5 <= -0.455  
     MolVol1 <= 106.675  
     LogP1 <= 1.22  
 THEN  
  
 A1A =  
     -0.7741 \* LogP1

Snippet 3: Type I astrocyte Area (A1A) model from a model tree classifier (M5Rules). The model instances then performs logic tests on its features with IF clauses. If all conditions are true the decision goes to a leaf to estimate A1A. The estimation is either a classification or a small linear model. In the latter, the values before features are the coefficients. The values in braces after the outcome: [number of instances the rule applies for (coverage) / and the percentage root mean squared error for instances that reach these leaves].

The logP (lipophilicity) of the untethered surface chemistries appears important, as they are included in both rules and linear models at leafs. From this work, logP has a –correlation with type I astrocyte area ( $r = -0.77$ ). This means as surface lipophilicity increases, type I astrocyte area decreases. Molecular mass and volume from this work have –correlations ( $r = -0.15$  and  $r = -0.42$ ) with type I astrocyte area but both are not significant. The surface acidity measure (pKa) correlates negatively in the early time point ( $r = -0.40$ ) and positively in the latter time point ( $r = 0.38$ ) but both are not significant. For the former, this means as the surface acidity decreases, type I astrocyte area decreases as well. For the latter correlation, the inverse is happening. As surface acidity decreases, type I astrocyte area increases.

Below are graphs of model performance from cross-validation for each time point:



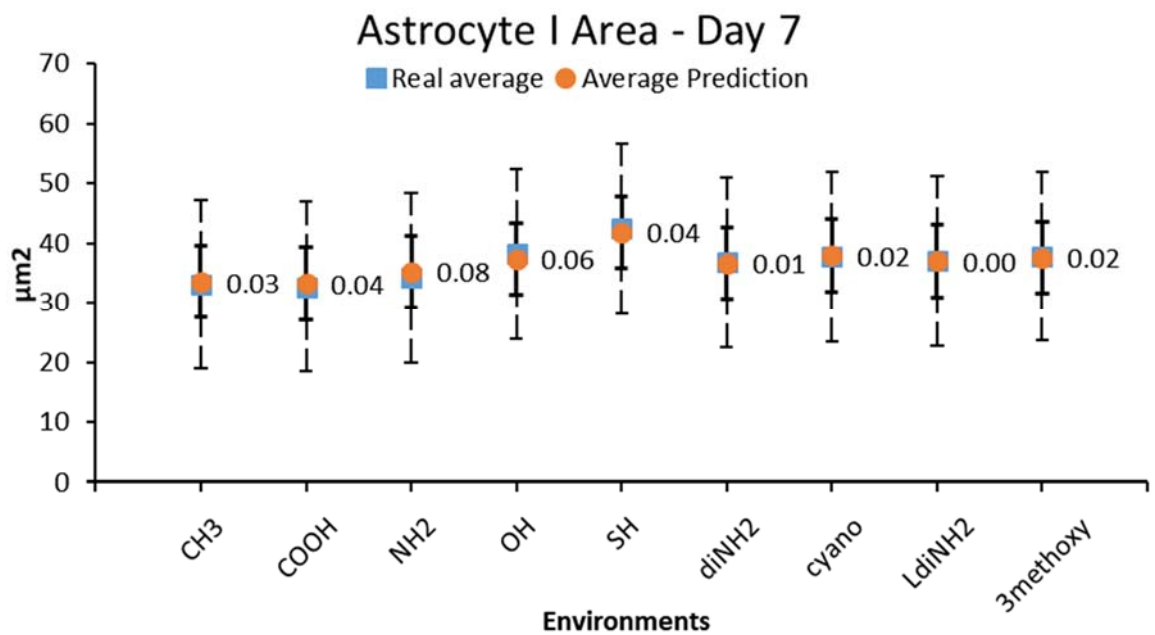


Figure 4.31: Type I astrocyte area day 3 and day 7 model performance from 10-fold cross validation.  $y$  axis represents the area in  $\mu\text{m}^2$  and in the  $x$  axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 0.87 \mu\text{m}^2$  and  $MPR = 0.1$ . For the later time point (day 7),  $MAE = 0.45 \mu\text{m}^2$  and  $MPR = 0.03$ . In other words, the model fit for both time points is excellent. Decomposing the prediction error gives bias (average error) for day 3 predictions at  $-0.36 \mu\text{m}^2$  and the variance (prediction standard deviation) is  $4.24 \mu\text{m}^2$ . For the later time point, the bias is  $0.05 \mu\text{m}^2$  and the variance is  $6.04 \mu\text{m}^2$ . The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. Prediction variance for both time points is small due to the design of model tree learning. The “best” rules are selected from constructed trees where these reduce the standard deviation of the outcome (estimate).

#### 4.2.3.11 Astrocyte fibre length

Astrocyte spreading is related to fibre length as astrocytes extend protrusions to interact with other cells and with the surface for migration and attachment (93). Astrocytes interact

with themselves, other glial cells and neurons (194). Minimising both type I astrocyte area and fibre length is preferred, and laminin's performance sets the upper boundary.

Modelling astrocyte fibre length was achieved with gradient boosting and one-level decision tree. Stochastic gradient boosting (254) method enhances the performance of 'base' classifiers. It is a method to increase model complexity and improve its fit by combining models learnt from base learners. It starts with a simple predictor such as the mean. Subsequent models from each iteration build the model stage-wise on a subsample of data, drawn at random (without replacement) to reduce computation time and add randomness. Randomness reduces prediction variance and therefore overfitting. The residuals left from the previous iteration are modelled again. Overall prediction is given by the sum of the outputs of the collection of models.

Each iteration of gradient boosting fits a model to residuals left by the classifier from the previous iteration. This parameter was set to  $I = 2$  for two 1-rule models to be fit. 1-rule does not have any hyper-parameters to tune. Features for this learning scheme were selected using correlation and backwards greedy search. Below is a table with the ranked features:

Table 4.10: Astrocyte fibre length feature selection and evaluation. Forwards greedy search: started with no features and added one at a time until there is no improvement in modelling accuracy. The merit score is the goodness of the subset after adding the corresponding feature in the left.

<b>Features</b>	<b>Merit score</b>
Day 3	0.56
Day 7	0.56
Acidity measure (pKa)	0.22
Partition coefficient (logP) – level 1	0.08
Partition coefficient – level 2	0.07
Partition coefficient – level 3	0.07
Partition coefficient – level 4	0.07
Partition coefficient – level 5	0.07

Molecular mass	0.07
Molecular volume	0.07

From the table above, the time points are found as important variables, but these are categorical and cannot correlate with a numerical response. Previous work (41) found acid dissociation constant (pKa) to +correlate with astrocyte fibre length ( $r = 0.35$ ) but this is not significant. The logP (lipophilicity) of untethered surface chemistries appears as a group and with good reason. From this and previous work (41), logP +correlates with astrocyte fibre length ( $r = 0.49$  and  $r = 0.79$ ) both significant. This means as surface lipophilicity increases so does astrocyte fibre length. Molecular mass and volume were both found to - correlate with the cell parameter ( $r = 0.66$ ) in previous work. Below is a visual representation of the model:

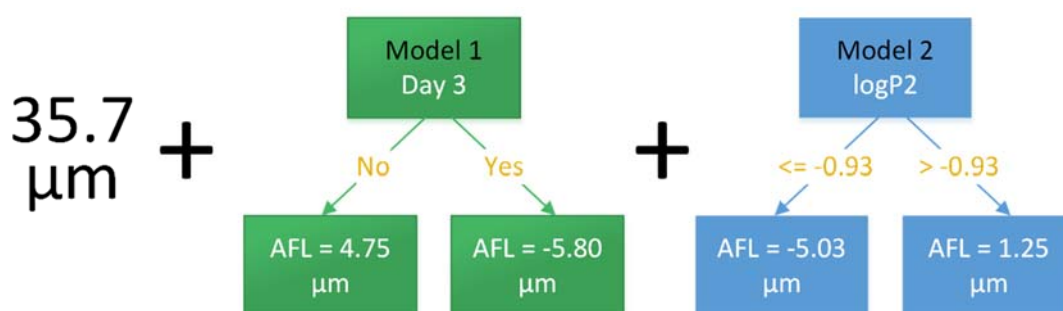


Figure 4.32: Astrocyte fibre length model. The prediction starts with 35.7  $\mu\text{m}$  then goes through through 2 shallow trees (stump) where the outcome of both are added together for the final answer. This method is called gradient boosting and  $n$  models are fit on residuals from previous predictions. The number of models fitted is determined by the gradient boosting learning rate (shrinkage) (255).

From the model representation above, the day indicator and logP2 were selected. logP (lipophilicity) of untethered surface chemistries was found in this and previous work (41) to +correlate with astrocyte fibre length ( $r = 0.49$  and  $r = 0.79$ ) both significant. This means as surface lipophilicity increases so does astrocyte fibre length.

Below are graphs of model performance from cross-validation for each time point:

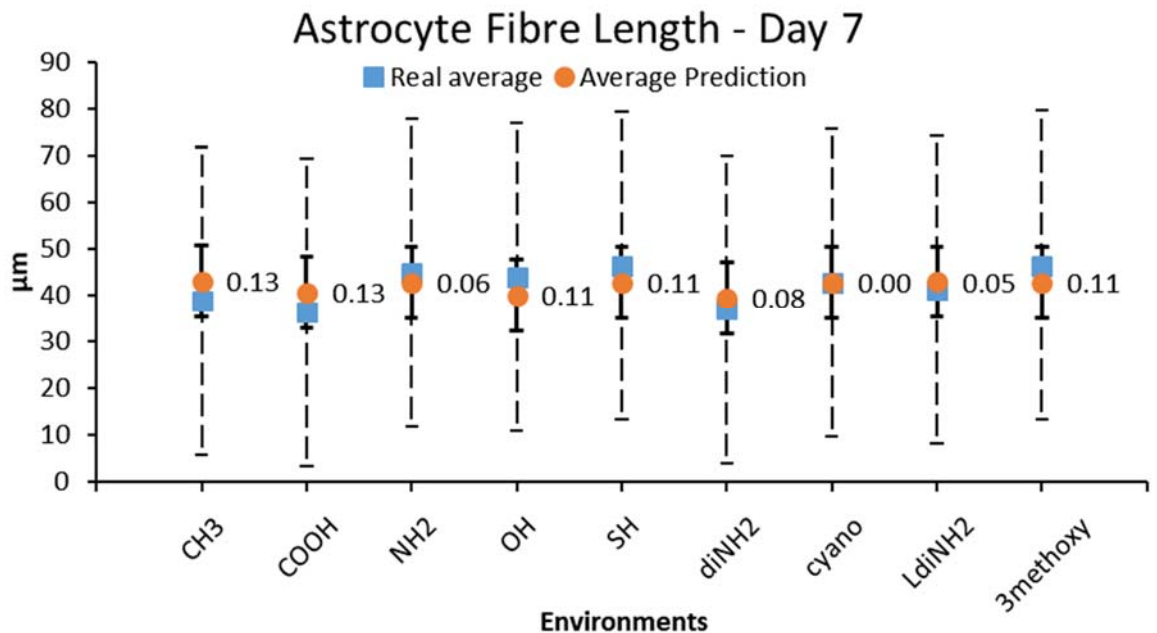
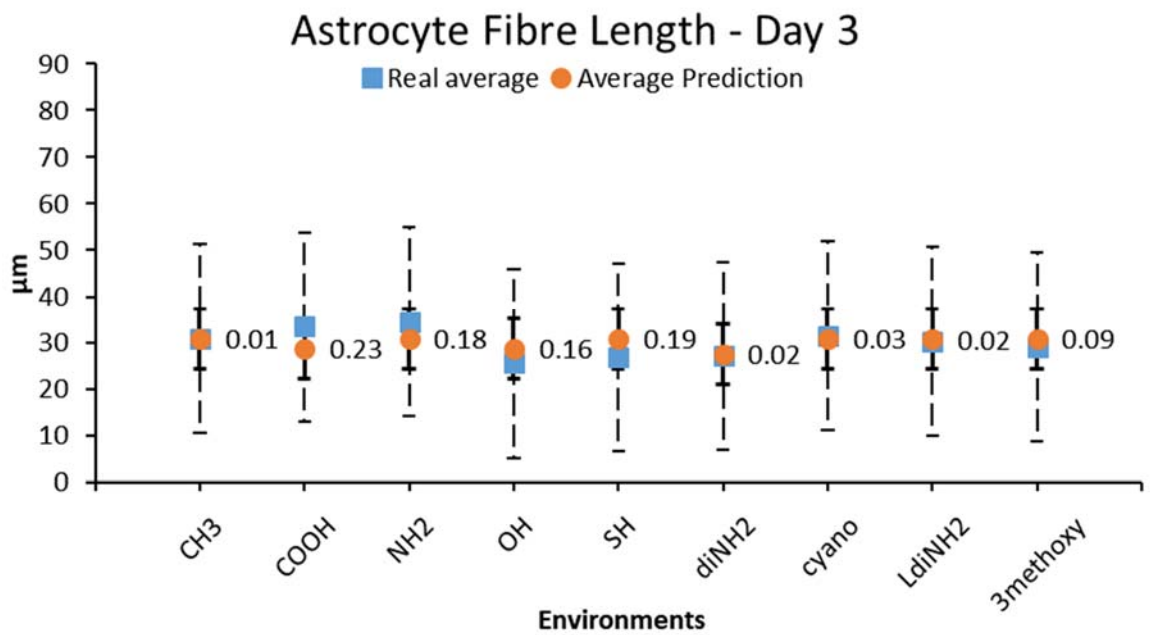


Figure 4.33: Astrocyte fibre length day 3 and day 7 model performance from 10-fold cross validation. y axis is cell projection length in  $\mu\text{m}$  and in the x axis are the cell culture environments used in experiments. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

On average, the mean absolute error (MAE) and model performance ratio (MPR) for the early time point is  $MAE = 2.11 \mu\text{m}$  and  $MPR = 0.1$ . For the later time point (day 7),  $MAE = 2.91 \mu\text{m}$  and  $MPR = 0.09$ . In other words, the model fit for both time points is excellent. Decomposing the prediction error gives bias (average error) for day 3 predictions at  $0.06 \mu\text{m}$  and the variance (prediction standard deviation) is  $6.45 \mu\text{m}$ . For the later time



point, the bias is  $-0.05 \mu\text{m}$  and the variance is  $7.71 \mu\text{m}$ . The closer these values are to zero the better but since the two sources of error are inversely related, we are after a trade-off that minimises the mean absolute error best. Even though shallow trees that are prone to high bias are used, gradient boosting corrects this by combining multiple models. This leads to prediction bias is close to 0 for both time points. Gradient boosting starts with a basic prediction then “fixes” it along the way by fitting another model on the residuals left from the previous iteration. Variance is low for the same reason, gradient boosting, as 3 models are used (average + model 1 + model 2) as shown in Figure 4.32. The next section following is sensitivity analysis to unveil the important chemical inputs the models use for prediction.

#### 4.2.4 Sensitivity analysis

This section is for investigating the models for the effect of individual chemical parameters on cell estimates. This is necessary to expose which chemical inputs matter the most in models where their inner workings are not easily interpreted. The modelling results may not reflect the real effect because here, we are exploring the inner workings of computational models.

A practical and common approach used for sensitivity analysis is where one-factor-at-a-time is changed to see what effect it produces on the output (372–374). The idea here is to tune one input variable while keeping others fixed typically around the centre of their value space. This is repeated for each of the input of interest. Sensitivity is then measured by observing changes in the output. Any change observed in the output will unambiguously be due to the single variable changed. A limitation of this approach is that it cannot detect the presence of interactions between input variables such as the ones found in an earlier

section (4.2.3.1.1). In the case where multi-collinearity is present then the input's effect may be like the ones it correlates strongly.

The process is as follows: all but one inputs are fixed at their baseline value. The remaining input in question is varied between its minimum and the maximum value. The cell models are used for cell estimates and the results are collected. These are cleaned for erroneous output and each is then tested with bivariate correlation against the input in question. Correlation shows the statistical relationship between variables. These relationships assume both dependence through in common use and linearity. Pearson's correlation has an advantage over using untransformed data to find correlation between variables but is also sensitive to outliers (203). The correlation between an input variable and cell variables will capture the effect one has on the other. Correlation significance tests tells us the upper and lower thresholds accepting correlations as significant. Significant here means the chances of accepting a false positive or false negative are within the threshold of choice (5% and 20% respectively). The stronger the correlation (close to 1 or -1), the smaller the correlation standard error and the more significant the correlation is.

Theoretical chemical designs were generated from user input and the predictive models provided the cell performance estimates. These were collected, and correlation tests were performed in pairs. With a sample size of at least  $n = 59$ , the graphs (Figure 4.34) show the correlation coefficient ( $r = 0.32$ ) comes with a risk of accepting false positives to 5% (left graph) and false negatives (right graph) to 20%. Graphs in Figure 4.35 establish thresholds accepting correlations as significant if they are  $\leq -0.21$  or  $\leq 0.21$ . Outside of these thresholds, the chance accepting a false positive (type I error) and false negatives (type II error) increase. Correlations between -0.2 and 0.2 are not significant and need

further evidence to support them. Below is correlation significance and after that correlation graphs for each chemical parameter against all cell parameters:

#### 4.2.4.1 Correlation significance

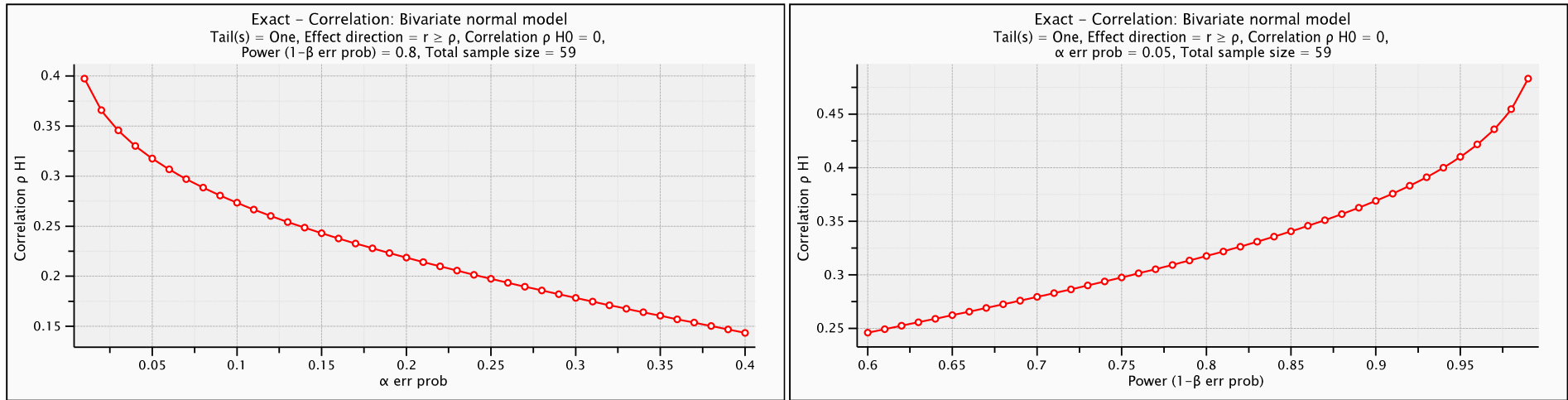


Figure 4.34: Correlation significance with a sample size of at least  $n = 59$ . Left: y axis is the correlation coefficient ( $H_1$ ) and x axis is the  $\alpha$  probability accepting false positives. Power was set at  $1 - \beta = 20\%$  chance accepting a false negative. Right: y axis is the correlation coefficient ( $H_1$ ) and x axis is the  $\beta$  probability accepting false negatives and.  $\alpha$  was set at 5% chance accepting a false positive.

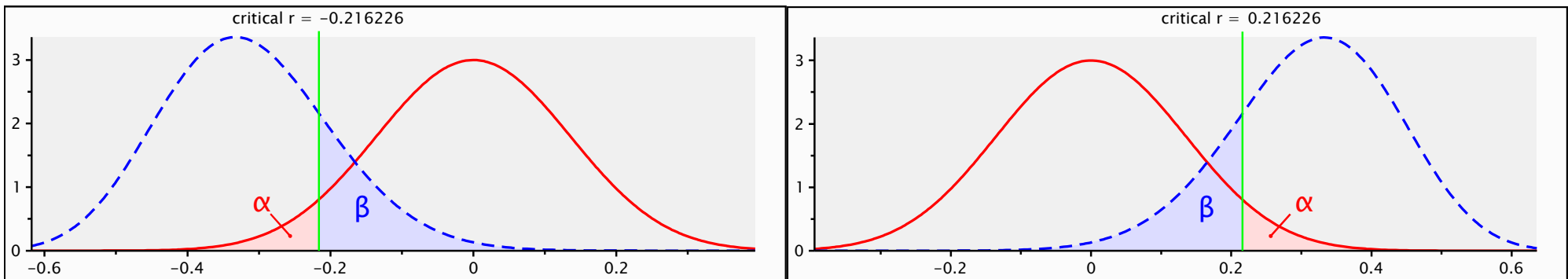


Figure 4.35: Critical correlation coefficient accepted as significant with a sample size of at least  $n = 59$ . y axis is the probability density for  $\alpha$  and  $\beta$  distributions and x axis is the correlation coefficient. Left graph shows the critical correlation coefficient for negative correlations and the right one for positive correlations. Correlations  $\geq -0.21$  or  $\leq 0.21$  are accepted as significant.

#### 4.2.4.2 Cell cluster area and neuron proportion

Cell cluster area is related with cell sphere (neurospheres) spreading early after seeding them on modified surfaces, and with cell proliferation especially in the later time point (day 7). The effects in play here are both chemical and biological. Maximising the cell cluster area/neurosphere spreading is desirable as this increases cell differentiation potential.

Below are correlation graphs between theoretical chemical designs and cell cluster area and neuron proportion estimates:

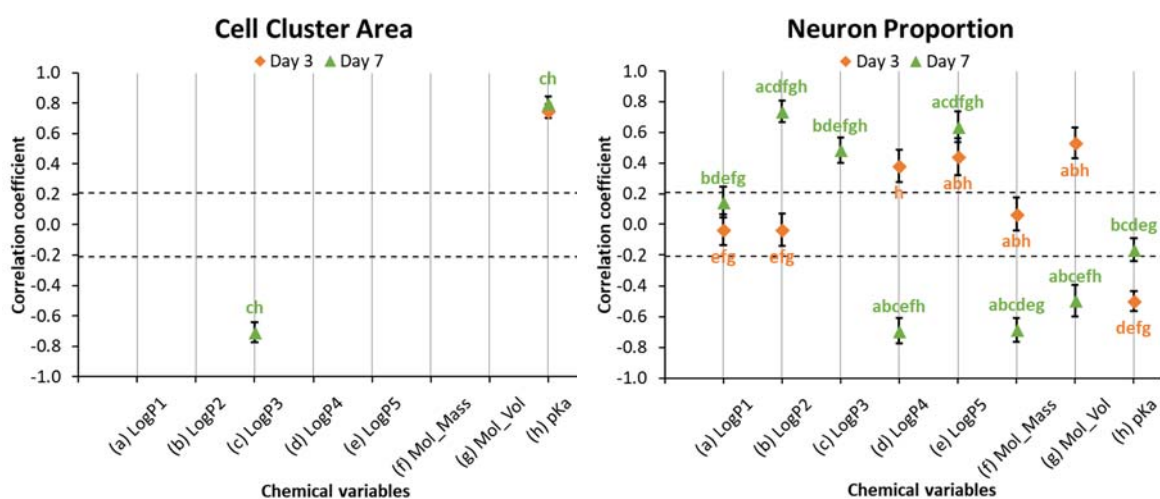


Figure 4.36: Sensitivity analysis of chemical inputs predicting cell cluster area and neuron proportion estimates. *x* axis are the chemical inputs and the *y* axis is the correlation coefficient. The dashed lines indicate the upper and lower critical correlation value considered significant with sample size of at least  $n = 59$ . Correlations between 0.20 and -0.20 are not significant. Error bars is the standard error of correlations. Abbreviations:  $\text{LogP}_n$  = lipophilicity measure of  $n$  constituent of the molecule,  $\text{mol\_mass}$  and  $\text{mol\_vol}$  stand for molecular mass and volume. Data labels indicate significant differences between data from the same time point.

From the figure above, the cell cluster area model is mostly affected from the logP (lipophilicity) of the 3<sup>rd</sup> constituent of the surface molecule and surface acidity (pKa). Experimentally, logP3 was found to be a good predictor in both current ( $r = -0.58$ ) and previous work (41) ( $r = -0.67$ ). This means as the surface lipophilicity increases, cell cluster area decreases. The pKa also has a strong +correlation with cell cluster area ( $r = 0.57$ ) meaning as the pKa value increases (less acidic), so does cell cluster area.

Neuron proportion is affected from logP2 to logP5, molecular mass, volume, and pKa. Experimentally, LogP4 was found in previous work to have a –correlation with neuron density ( $r = -0.48$ ) and the pKa with a +correlation ( $r = 0.38$ ) but the latter is not significant. From this work, we found –correlations with pKa and neuron density ( $r = -0.68$ ), neuron proportion ( $r = -0.52$ ) agreeing with previous findings (41). The logP correlations suggest that as surface lipophilicity increases, neuron density decreases. It is believed neurons are on top of an astrocyte carpet in *in vitro* 2D cultures (270). The pKa correlation suggests as the surface pKa increases so does neuron density. Molecular mass and volume were found to have –correlations with neuron density  $r = -0.47$  and  $r = -0.51$  respectively. This means as the molecular mass and volume increase, neuron density decreases. We interpret molecular volume and mass as chemistry complexity. After all, environments made with very complex molecules (laminin proteins) are used as the biological control in this project.

#### 4.2.4.3 Astrocyte type I and II proportion

There are two types of astrocytes, where type I has fibroblast-like morphology and type II has spindle-like morphology. Cell proportion tells us about differentiation (day 3) and proliferation (day 7). Below are correlation graphs between theoretical chemical designs and astrocyte type I and II proportion estimates:

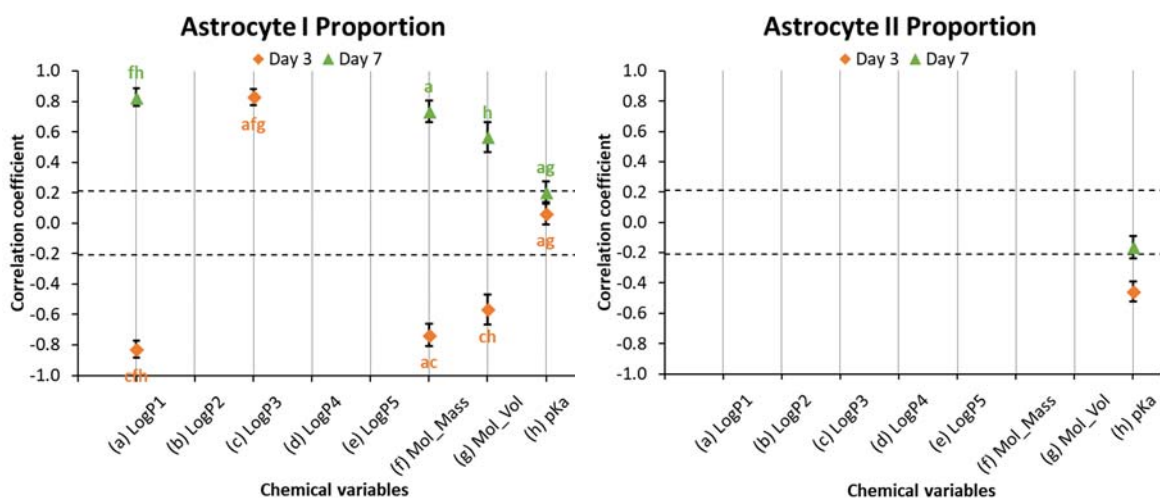


Figure 4.37: Sensitivity analysis of chemical inputs on astrocyte type I and II proportion estimates. The dashed lines indicate the upper and lower critical correlation value considered significant with sample size of at least  $n = 59$ . Data labels indicate significant difference between data from the same time point.

Type I astrocyte proportion model is affected by surface molecule logP1, logp3, molecular mass, volume and pKa. Cell proportion gives us the cell counts and standardising the latter with cell cluster area gives cell density per area ( $\text{mm}^2$ ). In other words, cell proportion is related with cell density. Experimentally from this and previous work, +correlations were found with type I astrocyte density and logP ( $r = 0.61$  and  $r = 0.79$  (41)). This means cell density increases as the lipophilicity increases on the culture surface. This adds to the hypothesis that astrocytes are closer to the culture surface compared to neurons (270). LogP3 also has a -correlation with type I astrocyte proportion ( $r = -0.48$ ) meaning as surface lipophilicity increases, cell proportion decreases. From this and previous work (41), molecular volume has -correlations with astrocyte density ( $r = -0.52$  and  $r = -0.71$  respectively). A similar relationship is found with pKa in this and previous work ( $r = -0.62$  and  $r = -0.61$  respectively). This means that as the molecular volume and pKa increase individually, type I astrocyte density decreases in both situations.

Type II astrocyte proportion model is affected mainly by surface acidity measure (pKa). Experimentally from this work, the pKa was found to +correlate with type II astrocyte

proportion ( $r = 0.44$ ) although not significant. In other words, this means as the pKa value increases (less acidic) so does the proportion of astrocytes.

#### 4.2.4.4 Proportion of unknown type cells and neurite length

Unknown type cells are cells that did not test positive for the markers (tags) used in experiments. In other words, these cells are unidentified of type but we know they are there as their nuclei tested positive (DAPI) and they are visible in cell images. These cells could be neural stem cells/progenitors, oligodendrocytes, ependymal cells or microglia. Worst-case scenario, unknown type cells are assumed as neural stem cells/progenitors therefore minimising their proportion is desirable. This is because progenitor cells can make copies of themselves and undifferentiated cells cannot enter a patient's brain in a transplant therapy (91,282).

Functional nerve tissues consist of neural projections (neurites or axons) to communicate with neighbouring cells using electrical conduction across large sections of tissue. Neurite length is a good indicator of this in artificial environments (*in vitro*). One aim of neuro-regenerative biomaterials is to grow and guide neurons to specific injury areas and re-wire compromised neural circuit to restore function. Increasing neurite length is desirable in order to connect to neighbouring cells and communicate across large sections of tissue. Below are correlation graphs between theoretical chemical designs, unknown type cell proportion and neurite length estimates:



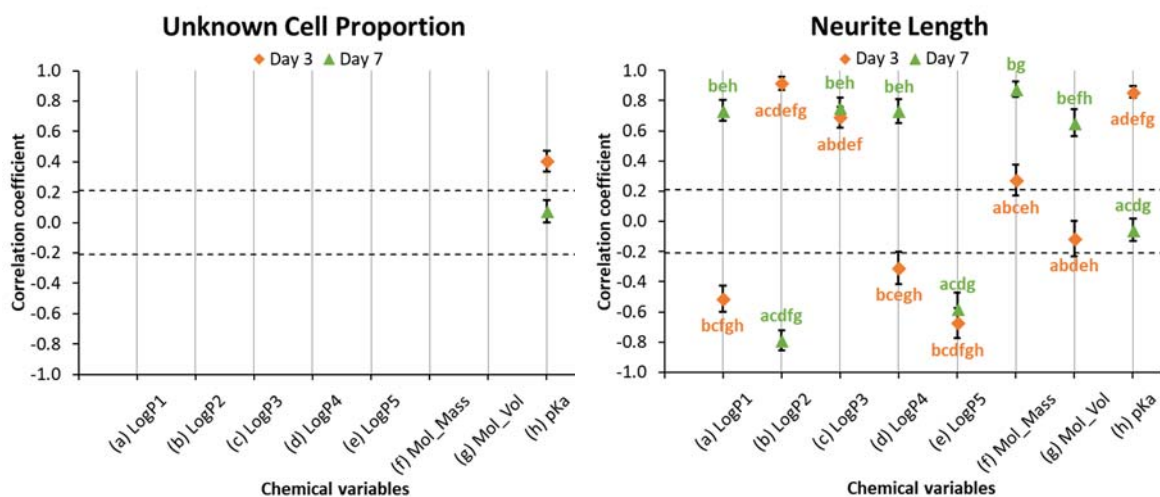


Figure 4.38: Sensitivity analysis of chemical inputs on unknown type cell proportion and neurite length. The dashed lines indicate the upper and lower critical correlation value considered significant with sample size of at least  $n = 59$ . Data labels indicate significant difference between data from the same time point.

Proportion of unknown type cells model is affected by the surface acidity measure (pKa). Experimentally, the pKa was found to +correlate with unknown type cell proportion ( $r = 0.75$ ). This means as surface acidity decreases, unknown type cell proportion increases with it.

Neurite length model is affected by all chemical inputs. Experimentally from this work, the logP was found to +correlate but not significant ( $r = 0.37$ ) and from previous work (41) the -correlation is a strong one ( $r = -0.81$ ). Both relationships were found in the model output. From this work, molecular mass and from previous work data, molecular volume were both found to +correlate with neurite length ( $r = 0.41$  and  $r = 0.46$ ) although none of these are significant. This means as molecular mass and volume increase, neurite length increases. Lastly, the pKa +correlates with neurite length from this and from the previous work ( $r = 0.51$  and  $r = 0.48$ ) although the latter is not significant. This means as surface acidity decreases, neurite length increases.

#### 4.2.4.5 Astrocyte area and fibre length

Astrocyte spreading is related with fibre length as astrocytes extend protrusions to interact with other cells and with the surface for migration and attachment (93). Astrocytes interact with themselves, other glial cells and neurons (194). Astrocyte spreading means forming stress fibres and focal adhesions (due to Rho activation) because astrocytes are establishing and stabilising altered cytoarchitecture (357). Minimising both type I astrocyte area and fibre length is preferred, and laminin's performance sets the upper boundary. Below are correlation graphs between theoretical chemical designs, type I astrocyte area and fibre length estimates:

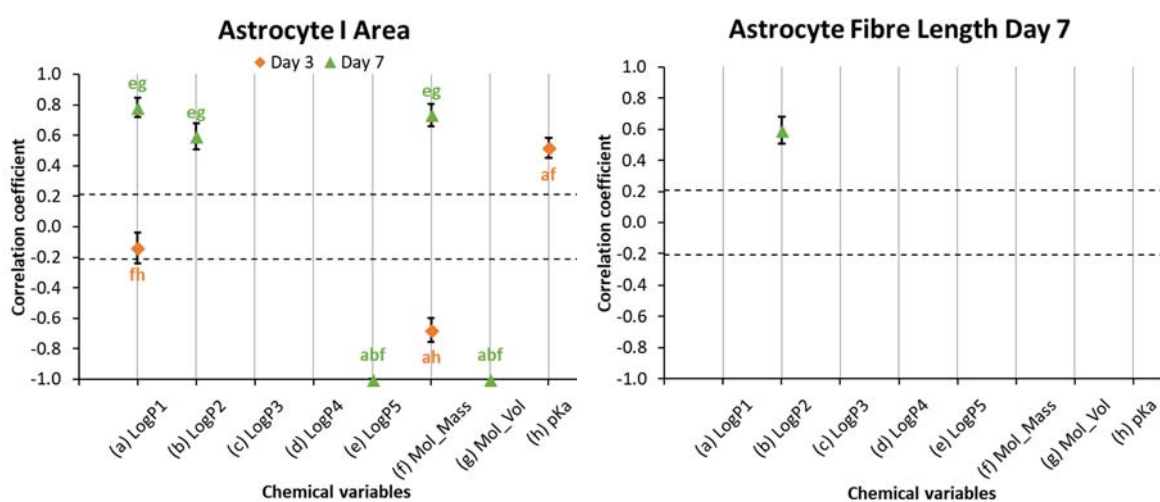


Figure 4.39: Sensitivity analysis of chemical inputs on type I astrocyte area and fibre length. The dashed lines indicate the upper and lower critical correlation value considered significant with sample size of at least  $n = 59$ . Data labels indicate significant difference between data from the same time point.

Type I astrocyte area model is affected by all chemical inputs except logP3 and logP4. Experimentally from this work, logP has a  $-$ correlation with type I astrocyte area ( $r = -0.77$ ). This means as surface lipophilicity increases, type I astrocyte area decreases. Molecular mass and volume from this work have  $-$ correlations ( $r = -0.15$  and  $r = -0.42$ ) with type I astrocyte area and both are not significant. The surface acidity measure (pKa) correlates negatively in the early time point ( $r = -0.40$ ) and positively in the latter time point ( $r = 0.38$ ) but both are not significant. For the former, this means as the surface

acidity decreases, type I astrocyte area decreases as well. For the latter correlation, the inverse is happening. As surface acidity decreases, type I astrocyte area increases.

Fibre length model is affected mostly from the logP2 chemical input. Experimentally from this and previous work (41), logP +correlates with astrocyte fibre length ( $r = 0.49$  and  $r = 0.79$ ) both significant. This means as surface lipophilicity increases so does astrocyte fibre length.

### 4.3 DISCUSSION

In the previous chapter, neural cell responses have been investigated on a range of substrates with defined chemical characteristics. Cells respond to their environment therefore biomaterial design is key in optimising cell culture for *in vitro* applications. Understanding cell-substrate interactions allows designing surfaces to influence cell differentiation and control their morphology. The application of this work can be, for example, to generate dopaminergic neurons lost during the progression of Parkinson's disease. Controlling stem cell differentiation to mature dopaminergic neurons is key to enhance regeneration of clinical therapies. Here, we added to the previous investigation of synthetic environments to control cell performance and match this with that of biological environments. Cell performance data used in this chapter are from E16 Sprague-Dawley rat cortex chosen to maximise the differentiation potential of neural stem cells and progenitors to cholinergic neurons (work with acetylcholine) that degenerate in Alzheimer's disease (375). This cell type is necessary for memory and learning (376).

### 4.3.1 Cell performance

This section details the morphological cell performance observed in cell images. Cell performance metrics allows profiling environments for their effect on tissue formation. Environments with defined surface chemistry were seeded with cell spheres (neurospheres) and at two time points (day 3 and 7), these were “fixed” in place. Fixed cells were tagged with fluorescent markers that selectively bind to target cell types. Below is a table with the optimisation goal for the cell parameters in this section:

Table 4.11: Cell parameter optimisation intent.

Cell parameter	Goal	Reason
Cell density (for all cell types)	Minimise	Increases chance for cell differentiation to neurons and glia
Neuron proportion	Maximise	Difficult to obtain, functional component of nervous system
Type I astrocyte proportion	Minimise	High proliferation ability and therefore increase density and paracrine signalling
Type II astrocyte proportion	Maximise	Rare in synthetic environments
Proportion of unknown type cells	Minimise	Lower risk of undifferentiated cells

#### 4.3.1.1 Cell cluster area and spreading

Methyl’s low performance on the early time point can be explained from the lipophilic nature of methyl (logP 1.82) where less cell migration is expected to minimise interaction with these environments. This means cell clusters do not merge as they do in other environments. There is a strong negative correlation with logP and cell cluster area at day 3 ( $r = -0.55$ ) meaning that as the logP value increases (lipophilicity), cell cluster area decreases. In addition, methyl environments are the most basic from the group with pKa value of 48. Environments containing nitrogen terminations are Cyano, l-diNH<sub>2</sub>, diNH<sub>2</sub>, NH<sub>2</sub>. The most lipophilic from this group is long diamine and the least lipophilic is diamine both exhibiting similar performance with other nitrogen containing environments as well. This

means the mechanism responsible for cell cluster spreading must lie in the head group as all of these environments have similar pKa values (9.7) also found in correlation tests ( $r = 0.55$ ). For the later time point, the low performance of acidic environments COOH, OH and 3-methoxy is expected. Their average pKa value is 4.5 being the lowest among all synthetic environments.

Carbomethoxy and butylamine perform very similarly on both time points and therefore on average. All environments in this project terminating with hydroxyl perform lower than carbomethoxy. The main difference is carbomethoxy termination is similar with the carboxyl (COOH) but with an extra carbon on the hydroxyl group (Figure 4.40). This is a clue that the hydroxyl termination does not help increase cell cluster area.

The main difference between propamine and amine is their amine group's position being on the 2<sup>nd</sup> from the top constituent on the former compared to the very top on the latter. In addition, butylamine shows the amine group's positional effect more clearly. The trend observed is the further up the amine group is the larger the clusters judging by amine and diamine. Diamine is still the top synthetic environment although it is similar with amine and propamine in terms of pKa value, 10.

Another trend observed is the logP values of the top 3 synthetic environments. It appears the lower the logP value the larger the cell cluster area (*diamine > amine > amineprop*). In previous work (264) (section 3.2.3.3) investigating amine, diamine, and triamine environment found that as amine content decreases, cell cluster area/neurosphere spreading increases. In relevant literature (377) however, a positive correlation was found, agreeing with the above finding of this work.

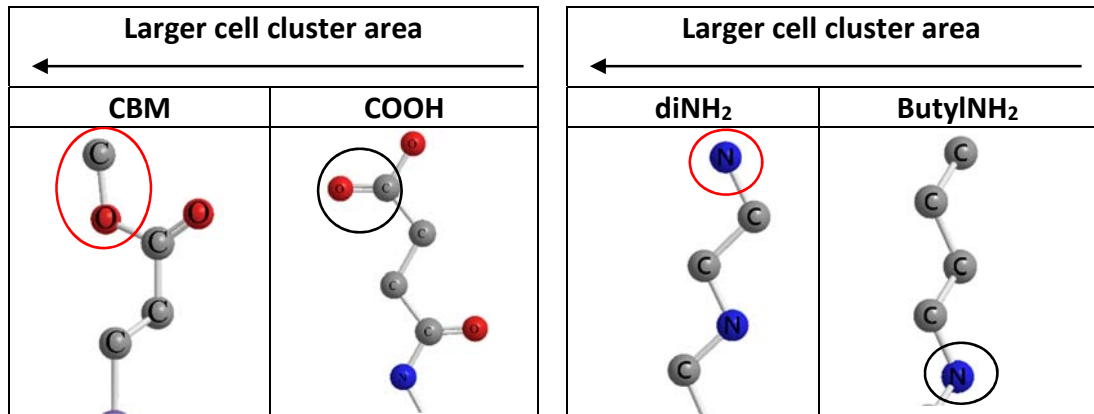


Figure 4.40: Chemical structures of carbomethoxy (CBM), carboxyl (COOH), butylamine (butylNH<sub>2</sub>) and diamine (diNH<sub>2</sub>). The constituents circled in red are for comparing the position of molecule constituents between chemistries in each group.

#### 4.3.1.2 Cell density and proportion

Cell density is a measure of how close the cells are to each other and cell proportion tells us how many of them are there compared to total cell counts. An imbalance in the proportion and migration of cells can have adverse effects for transplant recipients such as uncontrolled movement (overproduction of serotonin in the transplant) (282). Another effect is teratomas from progenitors or stem cells if they are present in the transplant tissue (91). Developing therapy grade tissue requires a benchmark environment and laminin was found to serve this purpose well (41,97).

Cell density and proportion measures were obtained for two time points – day 3 and day 7. These cell responses inform on cell differentiation on the early time point (day 3). At this stage, high density means cells reside inside the neurosphere because they are avoiding interacting with their environment. The other time point, day 7, is a good indicator of proliferation (101). Here, if cell density is similar but the cell cluster area is larger means cells are dividing.

#### 4.3.1.3 Neuron density and proportion

Cell density is calculated with the total cell type count standardised by the cell cluster area. This means, the more neurons found and the larger the cell cluster area the better as this minimises the cell density. This kind of cell behaviour is observed in biological environments (41,97,378).

As shown in the bottom graphs of Figure 4.14, diamine ( $\text{diNH}_2$ ) environments provide the lowest neuron density and the lowest neuron proportion with similar findings from previous work (264). Methyl ( $\text{CH}_3$ ) environments perform the least well, with highest neuron density, but they do come with high neuron proportion. The best balance between the 2 cell parameters is seen on amine ( $\text{NH}_2$ ) environments with low neuron density and good neuron proportion. Previous work (41) found very similar trends but the values obtained are different due to different sampling methodology in cell counting.

For the early time point, the good performers of cell density are amines ( $\text{NH}_2$ ,  $\text{diNH}_2$ ,  $\text{l-diNH}_2$ ), cyano, thiol ( $\text{SH}$ ) and hydroxyl ( $\text{OH}$ ) have similar and low neuron density but different neuron proportion. The interesting part is that with the exception of hydroxyl, all of the other environments in this group have similar acidity measures of 9.8 pKa. We hypothesise the difference in neuron proportion could be attributed to the lipophilicity of environments. Correlation tests revealed a significant +correlation ( $r = 0.54$ ) between  $\log P$  and neuron density at day 3. In addition, -correlations with pKa and neuron density, proportion ( $r = -0.68, r = -0.52$ ) are observed at day 7 meaning as the pKa values increases (less acidic), neuron density and proportion decrease. Amine and long diamine are more lipophilic compared to diamine, which supports the hypothesis. The exception is thiol being the most lipophilic but having the smallest proportion of neurons. A similar

trend is observed in the latter time point with these environments and thiol and methyl both pick up neuron proportion further supporting the hypothesis.

3-methoxy, COOH, and CH<sub>3</sub> environments offer higher cell proportion compared to CBM. 3-methoxy's termination is like carbomethoxy but the latter performs better in lowering neuron density. The difference between the two may not be significant although this hypothesis is supported from previous findings (section 3.3.2) where moderately hydrophilic/borderline hydrophobic surfaces reduce neuron density. It is suspected that CBM's double bonded oxygen on the carbon (circled in

Figure 4.41, A) is involved in lowering cell density, as this is the main difference with 3-methoxy. The answer could be in the termination's logP values. OH, 3-methoxy, and COOH have a logP of around -0.66. Methyl's (CH<sub>3</sub>) logP being the worst performer is 1.82. CBM on the other hand, has a logP value around -0.08 indicating this value is closer to the ideal for oxygen containing molecules and lower neuron density. A molecule having the carbomethoxy group has been used for neuroimaging to study dopamine reuptake in Alzheimer's disease patients (379). This could be the explanation for CBM performing better in cell density than other oxygen containing surface chemistries with similar cell cluster area.

The amines are diamine (diNH<sub>2</sub>), amine (NH<sub>2</sub>), propamine (NH<sub>2</sub>prop), butylamine (butylNH<sub>2</sub>), and aminohexyl (I-diNH<sub>2</sub>). These show the importance of the position of the amine group in the backbone self-assembly molecule used to change the surface chemistry of cell culture surfaces. The lower the amine group is found in the backbone of surface molecules, the higher the cell density is observed. On butylamine and aminohexyl (I-diNH<sub>2</sub>) environments, lower neuron proportion is observed (Figure 4.14 and



Figure 4.41, B). This effect could be related with the carbon content because the two environments have the highest count of carbon atoms in their backbones compared to the other amines. On day 7, diamine and butylNH<sub>2</sub> share something interesting. Both have low neuron proportion and an amine group around the 3<sup>rd</sup> and 4<sup>th</sup> constituent of the molecule. Diamine's top amine group is the reason lower cell density is observed on this environment.

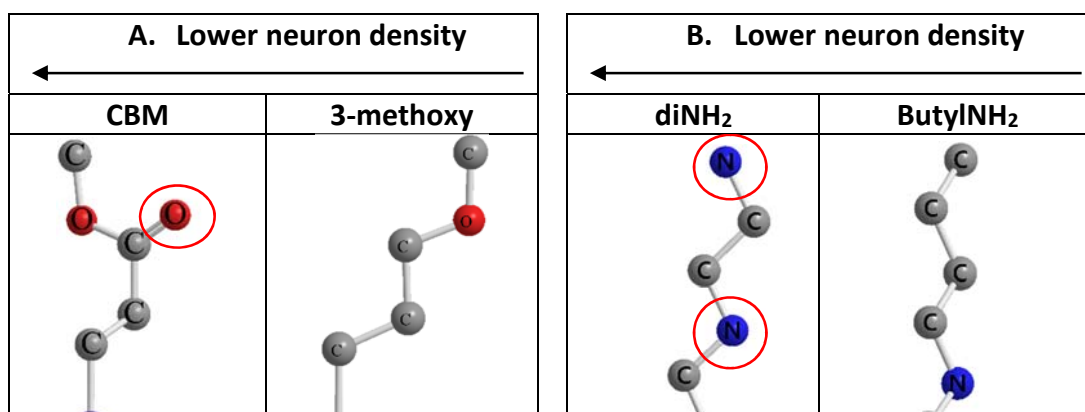


Figure 4.41: Chemical structures of carbomethoxy (CBM), 3-methoxy, diamine (diNH<sub>2</sub>), and butylNH<sub>2</sub> (butylNH<sub>2</sub>). The constituents circled in red are for comparing the position of molecule constituents between chemistries in each group.

#### 4.3.1.4 Astrocyte density and proportion

Cells tend to differentiate to a larger degree to type I astrocytes and these are excellent in proliferation than most of the other cell types in the central nervous system. The degree of astrocyte proliferation can be used as an indicator of cell stress (358,359). Extrapolating from this, lowering type I astrocyte proportion is desirable. For the other type of astrocytes (II), increasing their proportion is preferred as they are rare in *in vitro* cultures.

#### Type I astrocytes

As shown in the bottom graphs of Figure 4.16, diamine (diNH<sub>2</sub>) environments provide the lowest neuron density and the lowest neuron proportion with similar findings from previous work (264). Methyl (CH<sub>3</sub>) environments perform the lowest with highest neuron density but they do come with high neuron proportion. The best balance between the 2

cell parameters is seen on amine (NH<sub>2</sub>) environments with low neuron density and good neuron proportion. Previous work (41) found very similar trends but the values obtained are different due to different sampling methodology in cell counting.

For the early time point, the good performers of cell density are nitrogen-containing terminations (NH<sub>2</sub>, diNH<sub>2</sub>, l-diNH<sub>2</sub> and cyano). Diamine has the lowest cell density but also comes with highest proportion of type I astrocytes. Since the pKa (acidity) measure for these terminations is similar (9.8), the logP (lipophilicity) of diamine (-0.62) is hypothesised to have an effect on type I astrocyte proliferation and differentiation. These findings have also appeared in correlation tests revealing 2 correlations with logP on day 3 and 1 with pKa on day 7. The first one is a +correlation with type I astrocyte density ( $r = 0.61$ ) meaning as the logP value increases so does cell density. The other correlation is a -correlation with type I astrocyte proportion ( $r = -0.48$ ). At day 7, the -correlation with pKa and neuron density means as the pKa value increases, neuron density decreases.

Environments with lower type I astrocyte proportion are methyl and 3-methoxy and both have logP values of 1.82 and 1.51 respectively. Thiol has a similar pKa as the amines but the highest logP at 2.31. Thiol has among the high cell proportion setting the upper boundary for this chemical parameter. Adding to the hypothesis of the effect of lipophilicity on type I astrocyte proliferation and differentiation is the cell proportion of methyl (logP 1.82, pKa 4.8) in the later time point. It has the lowest cell proportion. Carboxyl environments (logP - 1.43, pKa 4.87) in this time point have the second lowest but also the highest cell density.

#### Type II astrocyte

As shown in the bottom graphs, overall diamine and long diamine (diNH<sub>2</sub>, l-diNH<sub>2</sub>) environments provide the lowest type II astrocyte density and the highest cell proportion.

Hydroxyl (OH), carboxyl (COOH), amine and methyl (CH<sub>3</sub>) environments are not good performers with low cell proportion. For the early time point, the proportion and density of cells goes up with the logP (lipophilicity) ( $r = 0.54, r = 0.51$ ) evident in amine, 3-methoxy and long diamine. For the later time point, it is not clear what drives cell differentiation to type II astrocytes. Except for amine and 3-methoxy, the good performers maximising type II astrocyte proportion have a pKa value 9.8.

In Figure 4.17, CBM returns as a good performer with type II astrocyte proportion on the later time point (day 7). This suggests type II astrocytes differentiation must have happened between day 3 and day 7 time-points for CBM environments. For cell proportion on day 3, butylamine (butylNH<sub>2</sub>) is the best performer from all synthetic environments and carbomethoxy (CBM) is the worst having no type II astrocytes. The performance order on this time point is *butylNH<sub>2</sub> < NH<sub>2</sub>prop < P/LAM*. For day 7, carbomethoxy (CBM) is the best performer and butylamine (butylNH<sub>2</sub>) is the worst from the remaining environments. The order of cell proportion performance on this time point is *CBM > NH<sub>2</sub>prop > P/LAM > butylNH<sub>2</sub>*.

Carbomethoxy (CBM) is better compared to 3-methoxy in lowering type I astrocyte density on day 3. The main difference with 3-methoxy is CBM's termination with the double bonded oxygen on the carbon (circled in Figure 4.41, A). As previously, the CBM's termination logP value is suspected to be the reason. The low performer methyl (CH<sub>3</sub>) has a logP value of 1.82 being lipophilic whereas CBM has a termination logP value around -0.08. Methyl does have the advantage with type I astrocyte proportion compared to CBM and this could be due to paracrine signalling inhibiting cell proliferation or because of low extracellular matrix resources (272,273).

The position of the amine group is important for lowering type I astrocyte proportion. For surface chemistries where the amine group is on the top such as amine and diamine, higher cell proportion is observed on day 3 except for aminohexyl. This could be from the higher carbon content of this chemistry as the methyl (CH<sub>3</sub>) environment has the lowest cell density. On surface chemistries where the amine group is lower than the termination (butylNH<sub>2</sub> and NH<sub>2</sub>prop), lower cell proportion is observed on day 3. By day 7, a similar trend is observed except for butylNH<sub>2</sub> with the highest cell proportion. This is interesting and warrants deeper investigation. Also, on this time point, all environments including the biological control, has an increase in cell proportion showing the proliferative ability of this cell type. No significant differences are observed in cell proportion for the same time points from the new environments.

Type II astrocytes (Figure 4.17), do not share the proliferative ability of type I astrocytes. For the environments with none of this cell type in day 3 but some in day 7 means there was cell differentiation to this cell type between the two time points. Cells present on day 3 but not on day 7 means astrocytes type II did not proliferate or they have died. It is possible other cell types proliferated to such degree that reduced the already low chance of sampling this cell type during image analysis.

CBM has the highest differentiation potential to type II astrocytes out of other oxygen containing self-assembly molecules (COOH, 3-methoxy, OH) on day 7. CBM's termination logP is suspected to be reason (logP -0.08) as this is the main difference spotted between the similar 3-methoxy. Although, 3-methoxy does have type II astrocytes present on day 3.

Butylamine (butylNH<sub>2</sub>) increases type I astrocyte and II proportion just by having the amine group near the head group of the self-assembly molecule. The head group is where these

molecules adhere on the surface allowing chemical modification. Propamine (NH<sub>2</sub>prop) has its amine group at a good position in the backbone of the molecule (Figure 4.43) as it offers low cell density and good proportion of type II astrocytes compared to others.

#### 4.3.1.5 Unknown type cells: density and proportion

The general trend is this cell type can appear in either time point (day 3 or day 7). For cells appearing on day and “disappear” by day 7 means they were stem cells/progenitors and have differentiated or have died. For unknown type cells appearing only on day 7 means they were missed during sampling in image analysis possibly overlaid by other cells, and have proliferated. From the left graph above, the top performing environment from the remaining surface chemistries is butylamine (butylNH<sub>2</sub>). The low performer is carbomethoxy with higher cell density on both time points. For this cell type, the cell proportion drives the trend of cell density resulting in very similar graphs. The order of performance for both cell performance metrics of new environments is *butylNH<sub>2</sub>* < *CBM* < *NH<sub>2</sub>prop*. There is significant –correlation between cell cluster area and unknown type cell proportion at day 3 time point ( $r = -0.51$ ). This adds to the hypothesis maximising cell cluster area will maximise cell differentiation to neurons and glia therefore minimise unknown type cell proportion.

Unknown type cells on carbomethoxy have a higher proportion compared to other oxygen containing surface chemistries (OH, COOH, 3-methoxy). This also gives higher cell densities especially on day 3. Figure 4.42 below shows CBM’s chemical structure compared to other oxygen containing molecules. The same double bond oxygen favouring type II astrocyte proportion may be doing so with unknown type cell types. Afterall, CBM and COOH share this and both perform worse than hydroxyl and 3-methoxy. For all of these surface

chemistries, their pKa values is around 4.41. Their termination logP value for CBM and COOH is -0.08 and -0.54 respectively. For OH and 3-methoxy, their logP values are -0.72 helping them with keep unknown type cell proportion low. Perhaps CBM is “friendlier” to the neural lineage since it is used in neuroimaging to study dopamine reuptake in Alzheimer’s Disease (379).

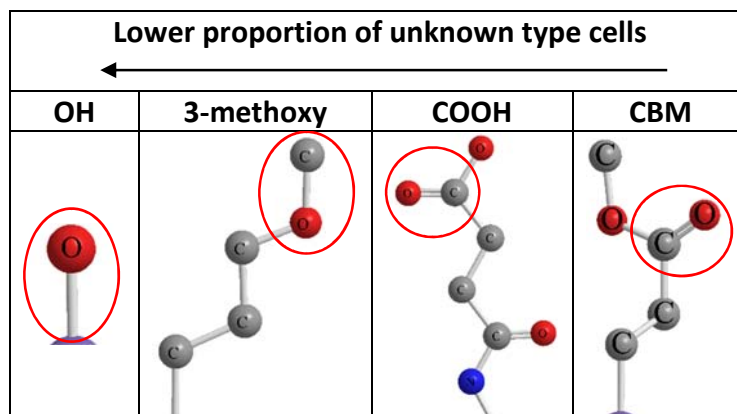


Figure 4.42: Oxygen containing surface chemistries. Ordered from lower to higher unknown type cell proportion (left to right).

From the cell proportion graph (Figure 4.18), similar trends are observed with cell density but there is an additional one. From the single amine surface chemistries, amine has the lower cell proportion and propamine (NH<sub>2</sub>prop) has the highest on day 7 (Figure 4.43). In between these two is butylamine (butylNH<sub>2</sub>). The better position for the amine group to reduce the proportion of unknown type cells is on the top or (termination) or 5 atoms down in the backbone shown in Figure 4.43:

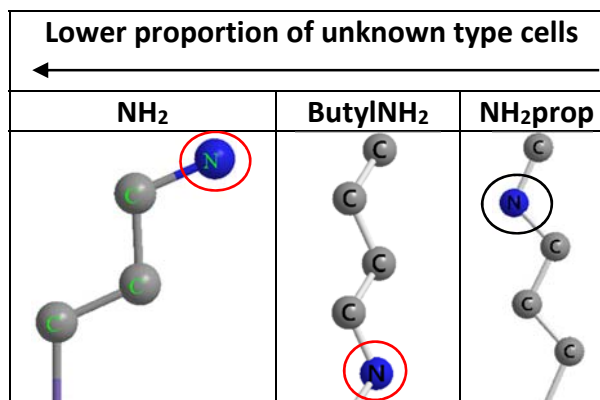


Figure 4.43: Nitrogen containing surface chemistries. Encircled in red are the constituents that help increase neurite length and in black are those that do not.

#### 4.3.1.6 Neurite length

The lowest performer from the group are carboxyl (COOH) and hydroxyl (OH) environments both considered hydrophilic and acidic (pKa 4.5) compared to the rest. Biological (P/LAM) environments have medium neurite length. This is expected because in these environments neurons have relatively high proportion and low cell density (Figure 4.14) meaning neurites do not have to extend far to find other cells. This is shown in amine environments with lower cell density than P/LAM and longer neurites. A similar trend was expected with diamine environments but this is not the case and this disagrees with previous work (41) on this time point. This difference is attributed to the sampling methodology.

The day 7 results constitute a good example of the surface chemistry's effect on neurite length. Methyl, carboxyl and amine (CH<sub>3</sub>, COOH, NH<sub>2</sub>) environments have the shortest neurites. For the first two this is expected as their neuron density and proportion is high. For amine being one of the lowest cell density environments, this outcome is unexpected. An explanation for this is neurons may be not migrated together with astrocytes in the early stages of the culture or neurons may have died. Diamine on the other hand with even lower

neuron density and proportion shows the expected behaviour with longer neurites now agreeing with previous work (41).

As shown in the bottom graph of Figure 4.20, overall thiol environments provide the slightly longer neurites compared to the rest and biological environments the shortest. There is +correlation with pKa (acidity) and neurite length on day 3 ( $r = 0.51$ ). Methyl ( $\text{CH}_3$ ) environments perform on the lower end likely due to their high cell density and high pKa value 48 setting the upper boundary for this chemical parameter.

Previous work found similar outcomes with these two environments (41) attributing the neurite elongation to molecules called epitopes present in the supportive molecule matrix (380,381). Another molecule attributed to neurite outgrowth is neural cell adhesion molecule (NCAM) present on the surface of neurons and glia (382). Biological (P/LAM) environments scored similarly with the earlier time point. For day 7, they have the shortest neurites disagreeing with previous work (41).

From Figure 4.20, propamine ( $\text{NH}_2\text{prop}$ ) shows a good position of the amine group on the surface chemistry. For molecules terminating with an amine group, shorter neurite length is observed. With respect to their acidity (pKa) values, propamine has the same (9.2) as the lowest performer amine. Their terminal logP values are also the same (-0.66) but their carbon content is slightly different. Amine has one less carbon in its backbone compared to CBM. The interesting part is amine shows the same flatness in neurite length on both time points just like biological control environments. Butylamine's amine group is lower than the other amine environments and it performs as second best. This is further evidence amine as a terminal group is not good for increasing neurite length. What is interesting is



butylamine's logP value (1.82) is the same as the lowest performer, methyl (CH<sub>3</sub>). This adds to the hypothesis that cells sense the 6<sup>th</sup> constituent of the surface chemistry.

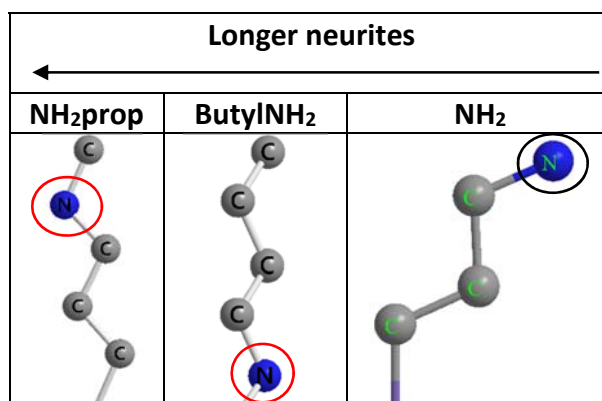


Figure 4.44: Nitrogen containing surface chemistries. Encircled in red is the position of the amine group helping increase neurite length and in black is the position that does not.

#### 4.3.1.7 Type I astrocyte area and astrocyte fibre length

Significant differences tests revealed that none of the scores is significantly different across any environment or time point. Visually, this can also be seen since the error bars (median absolute deviation) of day 3 and day 7 of the same environment meet at any point. The same goes for error bars between environments for both type I astrocyte area and astrocyte fibre length.

For both type I astrocyte area and astrocyte fibre length, none of the scores is significantly different across any environment or time point.

In this section, cell images and graphs of cell performance have been presented for all environments including the biological control (P/LAM). The chosen central tendency measure is the median and the measure of spread is the median absolute deviation. The reason for this is the gamma distribution of cell performance measurements although this are mainly for graphing purposes. Amine environments offer the largest cell cluster area, diamine environments the lowest neuron density, and methyl environments maximise

neuron differentiation. Propamine offers the highest proportion of the rare type II astrocytes and highest proportion of unknown type cells while the rest of the amines offer lower proportion of unknown type cells. Except for propamine, remaining environment show no significant differences for neurite length, type I astrocyte area and astrocyte fibre length.

The need to model cell responses computationally is to accelerate the discovery of better synthetic environments to develop nervous tissue fit enough to be used in therapies. With conventional cell culture experiments, it takes 6 months from start to finish, collecting and analysing cell data for 13 tissue-engineering environments. Done computationally, this will result in lower cost and less time. Below is a discussion for the selected models for morphological cell performance.

## 4.3.2 Computational cell models

### 4.3.2.1 Cell cluster area

Cell cluster area is related with cell sphere (neurospheres) spreading early after seeding them on modified surfaces, and with cell proliferation. Maximising the cell cluster area/neurosphere spreading is desirable as this increases cell differentiation potential (383).

The model for cell cluster area uses the time point indicator, molecular mass (chemical complexity), logP (lipophilicity) and acid dissociation constant (pKa) to predict cell cluster area. These chemical parameters are important factors to control cell cluster area. We know this as in cross-validation model selection good models have been found for both

time points. This model performance is found from splitting data in 10 parts then using 9 parts to train a model then use the remaining part to validate. This was repeated 10 times and in each iteration, the training and validation data are different. This method maximises the use of training data to construct predictive models and validating data to test the model on “future” data. In other words, as long as prediction error is minimised, the cross-validation model performance gives a good indication for model robustness. The variance of predictions is smaller than the real standard deviation meaning the confidence in the predictions is high. The logP and pKa of the terminal group were found important as well by sensitivity analysis for model inputs. The former chemical parameter was found to correlate strongly with previous (41) and new data.

#### 4.3.2.2 Neuron proportion

Successful cellular therapies to regenerate nervous tissue depend partly on the amount of neural cells delivered. Neuronal network allows function such as voluntary bodily movement. Controlling the density and proportion of transplant relevant cell populations is a key element in developing and scaling up cell-based therapy. As long as we know of the proportion of cells, cell density is derived from total cell count standardised by the cell cluster area. Neuron proportion tells us about differentiation in the early time point (day 3) and about proliferation on the later time point (day 7).

The neuron proportion model uses input variables such as time point indicator, chemical complexity (molecular mass), logP (lipophilicity), and pKa (acidity) of the terminal group of the surface. All of these are important in predicting neuron proportion also found by sensitivity analysis for model inputs. We know this as the last two were found to correlate in previous work (41) and in this work, the pKa, molecular mass and volume were found to

correlate. In addition, the 10-fold cross-validation model fit is excellent for both time points. Prediction variance is low for model due to the numerous decision trees used.

#### 4.3.2.3 **Astrocyte proportion**

Astrocytes are robust glial cells that play several roles in the central nervous system. There are two types of astrocytes, where type I has fibroblast-like morphology and type II has neuron-like morphology. They manage chemical signals (neurotransmitters) exchanged by neurons, strengthen neuron connections (synapses) (356) called long-term potentiation (275) among other functions. Astrocyte cell density and proportion can tell us about extracellular resources available in the vicinity (272,273) and cell stress (358,359). As previously mentioned, cell proportion tells us about differentiation (day 3) and proliferation (day 7).

The type I astrocyte proportion model uses the time point indicator, molecular mass and volume, logP (lipophilicity) and pKa (acidity) for prediction. These are important factors for the model also revealed by sensitivity analysis. The logP was found experimentally to correlate with cell proportion in this and previous work (41). The cross-validation model fit is good for both time points with low bias but the variance is sometimes outside the real standard deviation. More data and variance reduction techniques should help in minimising this.

The type II astrocyte proportion model uses the time point indicator, logP (lipophilicity), molecular mass and volume, and the pKa (acidity) of the terminal group of the surface chemistry. Experimentally from this work, the logP and pKa were found to correlate with this cell parameter. The model fit is good with low bias and variance but it is believed that

more data would help reduce this. Sensitivity analysis for model inputs revealed the main effector is the pKa agreeing with experimental findings.

#### 4.3.2.4 Proportion of unknown type cells

Unknown type cells are cells that did not test positive for the markers (tags) used in experiments. In other words, these cells are unidentified of type but we know they are there as their nuclei tested positive (DAPI) and they are visible in cell images. These cells could be neural stem cells/progenitors, oligodendrocytes, or ependymal cells. It is useful to know the proportion of this cell type as progenitor cells could be present in this cell group. These cells can make copies of themselves and therefore cannot enter a patient's brain in a transplant therapy (91,282).

The unknown type cell proportion model uses the time point indicator and the acidity measure of the terminal group. Experimentally from this work, the acid dissociation constant (pKa) has a strong correlation. Both model inputs are important predicting unknown type cell proportion. The model fit is great for the early time point and good on the later one. For both, there is low bias and prediction variance although the latter can be reduced with additional data. Sensitivity analysis for model inputs revealed the pKa is the more important of the two.

#### 4.3.2.5 Neurite length

Functionary nerve tissues consists of neural projections (neurites or axons) to communicate with neighbouring cells using electrical conduction across large sections of tissue. Neurite length is a good indicator of this in artificial environments (*in vitro*). One aim of neuro-regenerative biomaterials is to grow and guide neurons to specific injury areas and re-wire

compromised neural circuit to restore function. Increasing neurite length is desirable in order to connect to neighbouring cells and communicate across large sections of tissue.

The neurite length model uses mainly the logP (lipophilicity), molecular mass and volume of the untethered surface chemistry, and acidity measure (pKa) of the terminal group. Experimentally from this and previous work (41), all the model inputs were found to correlate with neurite/axon length. The model fit is very good for both time points with low bias and variance due to numerous decision trees used. Sensitivity analysis for model inputs revealed all the inputs affect the prediction agreeing with previous findings.

#### **4.3.2.6 Type I astrocyte area and astrocyte fibre length**

Astrocyte spreading is related with fibre length as astrocytes extend protrusions to interact with other cells and with the surface for migration and attachment (93). Astrocytes interact with themselves, other glial cells and neurons (194). Predicting astrocyte fibre length is important as it is an indicator of the indirect relationship astrocytes have with the culture environment. Astrocyte spreading means forming stress fibres and focal adhesions (due to Rho activation) because astrocytes are establishing and stabilising altered cytoarchitecture (357).

The type I astrocyte area model uses the lipophilicity measure (logP) of the untethered self-assembly molecules (SAMs) used to modify the presenting chemistry of culture surfaces. It also uses molecular mass and volume, and the acidity measure (pKa) of the SAM terminal group. Experimentally from this work, all the chemical parameters mentioned were found to correlate with type I astrocyte area. The model fit is excellent with for both time points with low bias and variance due to model tree rules selected that minimise the standard

deviation of the estimates. Sensitivity analysis revealed all type of chemical inputs affect the predictions agreeing with experimental findings.

The fibre length model uses the day indicator and the lipophilicity measure (logP) to predict the cell parameter. Experimentally from this and previous work (41), the logP was found to correlate with fibre length. The model fit is excellent for both time points, with low bias and variance due to combining numerous models using the gradient boosting method. This method starts with a simple prediction such as the mean then additional models are fit on the residuals left from the previous model. Sensitivity analysis revealed the logP to affect the predictions the most agreeing with experimental findings.

## 4.4 NOVELTY

### 4.4.1 Cell culture experiments

- 1) Acidic termination e.g. hydroxyl (OH) inhibit cell cluster area spreading
- 2) The further up the amine group is the larger the clusters judging by amine and diamine with similar pKa values (~10)
- 3) The lower the logP value the larger the cell cluster area (*diamine > amine > amineprop*)
- 4) In moderately hydrophilic/borderline hydrophobic surfaces with overall logP ~0, lower neuron density is observed
- 5) On surfaces with logP ~0, lower density of type I astrocyte density and higher proportion of type II astrocyte is observed
- 6) The lower the amine group is found in the backbone of surface molecules, the higher the neuron density is observed

- 7) Lower neuron proportion is observed on surfaces with higher carbon and some amine content
- 8) The higher the amine group is on the surface, the higher type I astrocyte density is observed
- 9) The further up the amine group is on the surface, the higher type I and type II astrocyte proportion and the lower the unknown type cell proportion is observed
- 10) Surface chemistries with logP values around -0.72 help with keeping the proportion of unknown type cells low
- 11) The further up the amine group is on the surface, the shorter neurite length is observed
- 12) There is evidence that cells sense the 6<sup>th</sup> constituent of the surface chemistry

#### 4.4.2 Chemical inputs used by computational models

- 1) Cell cluster area uses the time point indicator, molecular mass (chemical complexity), logP (lipophilicity) and acid dissociation constant (pKa)
- 2) The neuron proportion model uses the time point indicator, chemical complexity (molecular mass), logP (lipophilicity), and pKa (acidity) of the terminal group of the surface
- 3) Type I astrocyte proportion model uses the time point indicator, molecular mass and volume, logP (lipophilicity) and pKa (acidity) for prediction
- 4) The type II astrocyte proportion model uses the time point indicator, logP (lipophilicity), molecular mass and volume, and the pKa (acidity) of the terminal group of the surface chemistry
- 5) The unknown type cell proportion model uses the time point indicator and the acidity measure of the terminal group



- 6) Neurite length model uses mainly the logP (lipophilicity), molecular mass and volume of the untethered surface chemistry, and acidity measure (pKa) of the terminal group
- 7) The type I astrocyte area model uses the lipophilicity measure (untethered logP), molecular mass and volume, and the acidity measure (pKa) of the SAM terminal group
- 8) The fibre length model uses the day indicator and the lipophilicity measure (logP) to predict the cell parameter

## 5 SCREENING SURFACE CHEMISTRIES COMPUTATIONALLY

---

### 5.1 INTRODUCTION

Each predictive model has the captured relationship between the surface chemistry and cell performance. These were discovered using machine learning and data from cell culture experiments. The predictive models take chemical designs (in numerical form) as inputs and produce estimates of morphological cell performance shown below in Figure 5.1. The goodness of the chemical design depends on their cell performance that is compared to a target. This target is the cell performance of *in vitro* biological environments with laminin and the comparison is performed with a dissimilarity function from Bray & Curtis (260). Performing cell culture experiments using computational models then selecting candidate surface chemistries using a ranking system is mathematical optimisation. This method is the selection of the best element, with respect to some criteria, from a set of available alternatives (384). In a simple form, an optimisation problem is solved by maximising or minimising a function by systematically choosing input values from within an allowed set and computing the value of the function. To our knowledge, our work is the first to mathematically optimise the chemical design of biomaterial surfaces.



Figure 5.1: Computational cell culture methodology. The inputs are numerical surface chemistries fed into the system with predictive models and the output is cell performance estimates.

Optimisation approaches to biological problems can be classified to exact and approximate methods. The former outputs the optimal solution when convergence is achieved but this does not occur on every instance and this method is limited to small search domains. Approximate methods always output a candidate solution but there is no guarantee this is the best one (local minima). Approximate algorithms can be classified as stochastic and deterministic. Contrary to deterministic methods, stochastic use a random component and this means different solutions may be found given the same input parameters.

Related work focused in optimising culture conditions, scaffold design, and drug delivery mechanisms. In a recent study, the parameters for electrical stimulation have been optimised for cardiac tissue engineering from rat cardio-myocytes to develop transplant-grade tissue. The electrode material, amplitude and frequency of stimulation have been systematically varied to determine the optimal conditions for tissue engineering. The latter two have been optimised with models of electric fields experienced by cells found by solving of Maxwell's equations with the electro-quasistatic approximation (385). The authors discovered non-computationally that carbon electrodes exhibit the highest charge-injection capacity and produce cardiac tissues with the best structural and contractile properties. Computationally, the findings contribute to defining bioreactor design specifications and electrical stimulation regime for cardiac tissue engineering.

In scaffold design optimisation, recent work focused on the optimisation methodology rather than a specific biological target (386). Material and pore architecture (mechanical properties) of scaffolds have been optimised for use in tissue engineering. The scaffolds were fabricated with the stereolithography method being versatile and provides freedom of design. Scaffold morphology is a key factor determining tissue formation, as the pore network initially provides the spatial template for cell adhesion and proliferation and the deposition of extra-cellular matrix. Mechanical properties of the scaffolds were evaluated with in compression (mechanical loading) and numerical simulations of this were conducted utilising the nonlinear finite element to generate data. The hyperelastic model the authors made was used to predict mechanical behaviour of structures with different designed pore architectures. The model fit appears good, but the authors did not provide error measures. Nonetheless, the authors showed stereolithography fabrication methods can be used to prepare tissue-engineering scaffolds with designs that can be modelled, allowing optimisation of the properties of the structures.

### 5.1.1 Aims & Objectives

The aim is to screen surface chemistries rapidly with respect to their cell performance using a computer instead of real cell culture experiments. Model estimates of cell performance are compared with real data from new cell culture environments. The new data have not been included to train the predictive models. This allows testing the true predictive performance of the models. Cell performance from synthetic environments is compared to a target. This target is the cell performance of *in vitro* biological environments (laminin, P/LAM).

The objectives for this chapter include:

1. To compare cell performance data from new synthetic environments with environments from Chapter 4
2. To test computational models for their predictive performance on the new surface chemistries with the new cell performance data not present in model training
3. To perform mathematical optimisation to screen unexplored surface chemistries with respect to their cell performance using predictive models

## 5.2 RESULTS

### 5.2.1 Model testing

Machine learning allows computers to learn without being explicitly programmed. By learning is implied programs go through existing data and look for patterns to devise complex models and algorithms that lend themselves to prediction. These analytical models allow practitioners to produce reliable and repeatable decisions and results and uncover hidden insights by learning relationships and trends in the data.

Once the computational problem has been defined, the data need to be prepared (pre-processed) to communicate as much information as possible. This is to create effective predictive models using machine-learning programs. Once the models are trained, the next step is to test them with data that has not been used to train them. This is necessary if the models are to be used in practice. Indirectly, data pre-processing, learning algorithms and their tuned parameters are tested in the process as these have an impact on the quality of the model learnt.

The error measure on an independent test dataset is a good indicator of generalisation performance otherwise known as prediction goodness. The performance measure is taken directly from using the models on a test dataset. Large prediction errors on the test set means there is something wrong with the format of training data typically inadequate pre-processing, noise, inconsistency, not enough data, improper learning algorithm, or the model is overfit. The latter means the model is specialised to training data and performs well only on this.

A performance measure is required to test the learnt models. The one chosen is model performance ratio (MPR) derived from the absolute difference between real median and prediction standardised by the average (1) standard deviation. A ratio closer to 0 means the closer the prediction is to the real median and ratio of 1 means the prediction is outside of what is defined as acceptable. This error measure was seen previously in the previous chapter for model selection. The values for this were obtained from 10-fold cross-validation (231).

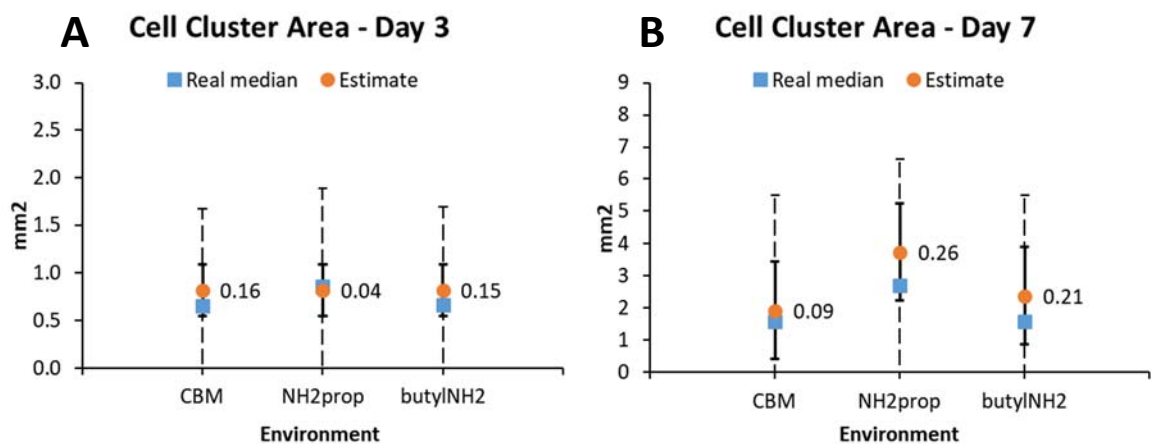
For the new test dataset, cell culture experiments were performed in the same fashion as previously for both chemistry and cell performance. As previously, the error measures of each model were collected and shown in Table 5.1:

Table 5.1: Model performance using the test dataset. This dataset consists of data from 3 chemistries that have not been used to train the models. D3 and D7 stand for day 3 and day 7 time points.

Target	Classifier	Time point	Avg. value	Average prediction	Avg. model performance ratio
Cell cluster area	Decision tree (bagging) (235)	D3	0.73 mm <sup>2</sup>	0.82 mm <sup>2</sup>	0.12
		D7	1.95 mm <sup>2</sup>	2.68 mm <sup>2</sup>	0.19
		D3	8.96 %	9.09 %	0.41

Neuron proportion	Ensemble decision trees (235)	D7	2.88 %	5.49 %	0.55
Type I astrocyte proportion	Randomised feature (239) k-Nearest Neighbours (240)	D3	89.44 %	92.43 %	0.39
		D7	96.44 %	94.72 %	0.17
Type II astrocyte proportion	Support Vector Regression (243,245)	D3	1.46 %	0.38 %	0.39
		D7	0.96 %	0.44 %	0.44
Proportion of unknown type cells	Support Vector Regression	D3	0.17 %	0.18 %	0.18
		D7	0.65 %	0.18 %	0.29
Neurite length	Randomisable ensemble of decision trees (bagging)	D3	45.37 $\mu\text{m}$	49.18 $\mu\text{m}$	0.27
		D7	105.79 $\mu\text{m}$	93.93 $\mu\text{m}$	0.73
Type I astrocyte area	Model tree (249)	D3	27.23 $\mu\text{m}^2$	27.19 $\mu\text{m}^2$	0.03
		D7	43.86 $\mu\text{m}^2$	37.49 $\mu\text{m}^2$	0.24
Astrocyte fibre length	Gradient boosted (254) decision trees (257,258)	D3	29.17 $\mu\text{m}$	30.97 $\mu\text{m}$	0.13
		D7	40.81 $\mu\text{m}$	42.95 $\mu\text{m}$	0.09

Cell data have been indicated for the time point they belong with binary (0 for day 7 and 1 for day 3). This is for practical reasons, as one model is needed to predict cell performance. Perhaps better models would have been found if the data were split for the two time-points. Regardless, all but one of the models perform within the acceptable level ( $MPR < 1$ ). The above results are dissected and shown below for each model test environment. Next, is the model testing for cell cluster area, neuron and type I astrocyte proportion:



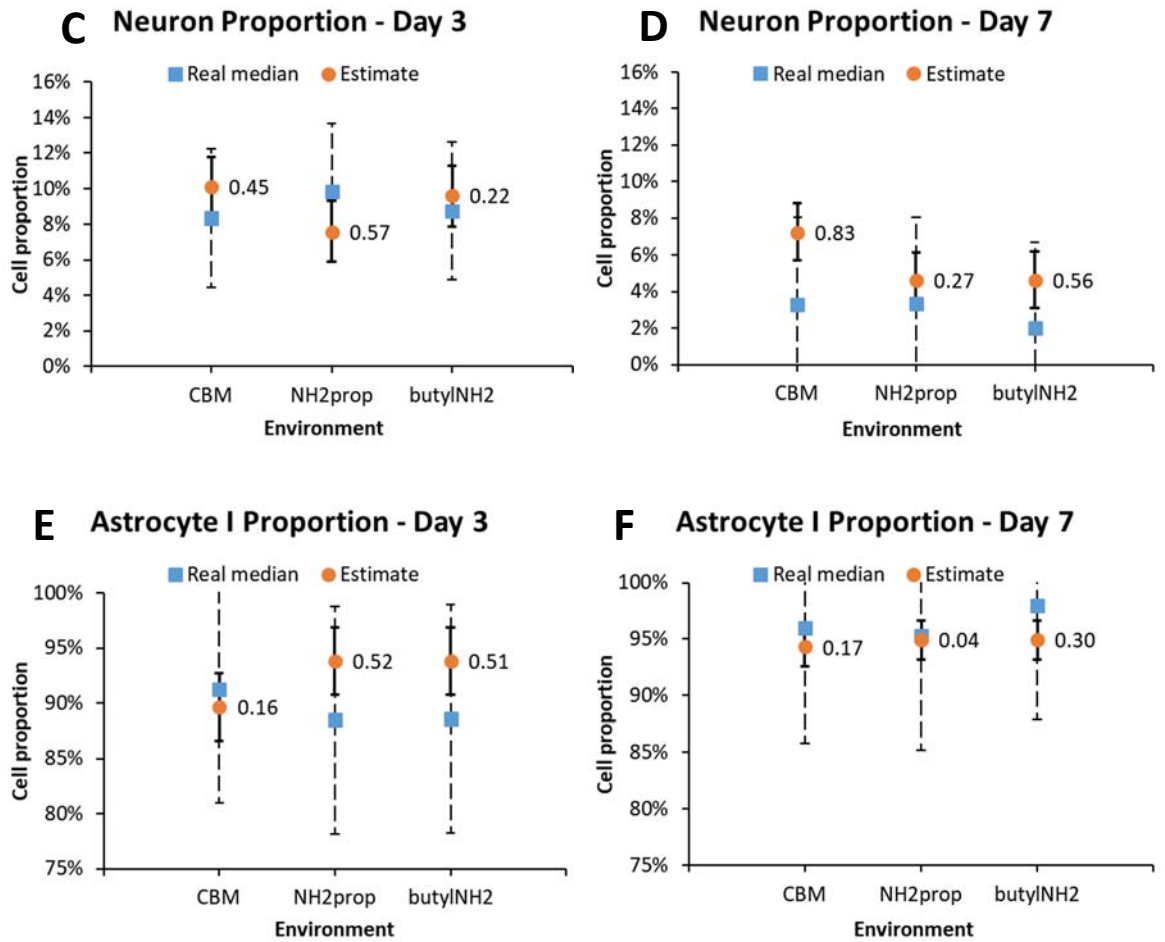


Figure 5.2: Model testing for cell cluster area, neuron and astrocyte proportion models. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. Dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates ( $n = 90$ ).

From the results above, all predictions are within the acceptable threshold for model performance ratio ( $MPR < 1$ ). Neuron proportion for carbomethoxy (CBM) on day 7 (graph D) has the largest MPR value (0.83) from the group. Figure 5.3 shows graphs of model test for cell proportion of type II astrocyte and unknown type cells, and neurite length:

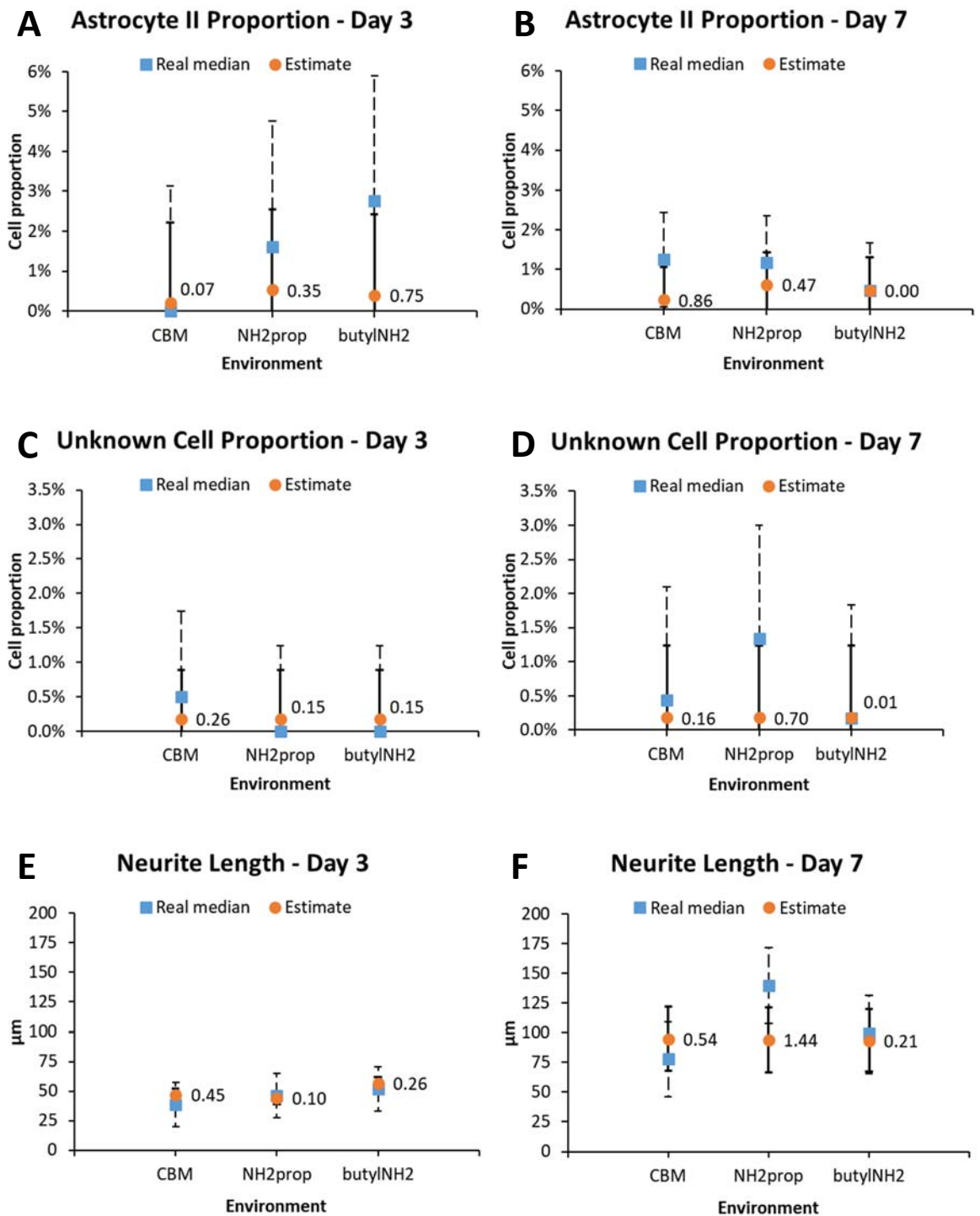


Figure 5.3: Model testing for astrocyte and unknown type cell proportion, and neurite length models. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates ( $n = 90$ ).

From Figure 5.3, CBM's MPR for type II astrocyte proportion on day 7 (graph B) is the largest from this group (0.86). Next, are model test graphs for type I astrocyte area and astrocyte fibre length:



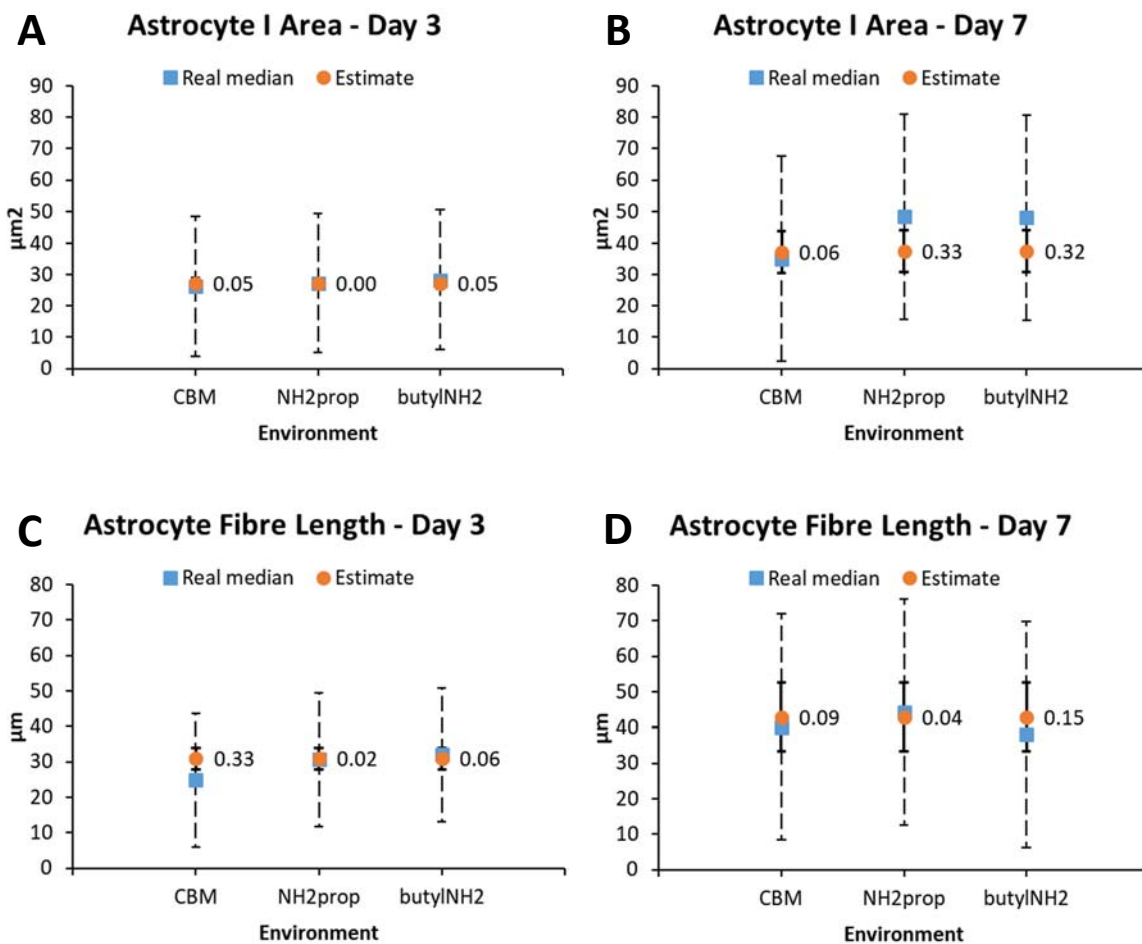


Figure 5.4: Model testing for type I astrocyte area and astrocyte fibre length models. Blue symbols represent real data and the orange symbols are the estimates. The data labels on the right handside show the model performance ratio which is a measure of model goodness compared to real values and their standard deviation. The dashed error bars represent 1 standard deviation of real data and the solid line represents the standard deviation of estimates.

This section was about testing the learnt models on new data that were not used to train the models. This is a good method to indicate their real generalisation performance. Model testing can tell whether models are underfit (high bias), overfit (high variance) or, the desirable, have the balance between the two.

Table 5.1 shows the average real values, model estimates, and model performance ratios. Subsequent graphs show the dissected model testing values for each entry in the table. In the next section are the results of computational cell culture experiments and the process of filtering the results to reconstruct new surface chemistries.

## 5.2.2 Better synthetic environments

After the models have been tested and are good to be used in practice, the next step is to conduct computational cell culture experiments. This involved generating theoretical chemical designs (in numerical form) and passing them to computational models to obtain from them cell performance estimates. The next step is to compare the estimates of synthetic environments to a target, laminin's cell performance. The problem described is mathematical optimisation since the goal is to find better chemical designs out of a set of alternatives (with criteria). The purpose can change by choosing the appropriate target e.g. maximise neuron proportion by choosing methyl ( $\text{CH}_3$ ). This whole process has been automated with made-software named 'Get-Chem'.

Both training and testing data have been unified into one dataset and this has been used to retrain the models with the same configuration as previously (Table 4.5). After this step, the theoretical chemical designs were to be defined. Each model input needs a minimum, a maximum and step values as shown in the table below (Figure 5.5).

### 5.2.2.1 Computational cell culture setup

There are thresholds for the values of each model input e.g. max for logP should be no more than 5 according to Lipinski's rule of five (209) for drug likeness. Values above this for drugs are toxic for humans. Most upper and lower limits for model inputs were determined from the min and max values of chemical variables of existing data. Since the target cell performance is from laminin environments, the pKa upper limit was adjusted by removing one outlier. Methyl's pKa value is 48 and this is over 4 times larger than the next one down. In addition, methyl environments are considered mediocre at best from this work and previous work (264).

The table below (Figure 5.5) shows the starting value table for model inputs to generate theoretical chemical designs. The lower and upper limits for each model input is the top and bottom values respectively. The step value is the increment from the previous value until the upper limit is reached. The min/max/step values together determine the granularity of mathematical optimisation and computational cost of Get-Chem.

Figure 5.5: Starting value range for model inputs when performing computational cell culture experiments. This set of data is recursively combined to provide around 15 million theoretical chemical designs in numerical form. These are given to predictive models to estimate cell performance.

	LogP1	LogP2	LogP3	LogP4	LogP5	Mol. Mass (Da)	Mol. Vol (Å)	pKa	D3
<b>Total values = 65</b>	-1.93	-1.65	-2.54	-1.22	-1.22	60	75	4.5	0
	-1.43	-1.15	-2.04	-0.72	-0.72	75	90	7.5	1
	-0.93	-0.65	-1.54	-0.22	-0.22	90	105	10.5	
	-0.43	-0.15	-1.04	0.28	0.28	105	120	13.5	
	0.07	0.35	-0.54	0.78	0.78	120			
	0.57	0.85	-0.04	1.28	1.28				
	1.07	1.35	0.46	1.78	1.78				
	1.57	1.85	0.96	2.28	2.28				
	2.07	2.35	1.46	2.78					
	2.57	2.85	1.96						
		2.46							
		2.96							
		3.46							
<b>Count</b>	10	10	13	9	8	5	4	4	2

The first five variables are the partition coefficients for the top 6 constituents of the surface chemistries (Figure 5.6). Mol. Mass and volume stand for molecular mass (Da) and volume (Å) and the pKa is the acid dissociation constant (log Ka). The last variable is the time point indicator with 1 and 0 values for day 3 and day 7 respectively. When recursively combined, this list generates 15 million chemical designs and these are passed to predictive models to provide cell performance estimates for each. Once the estimates are fully in place, they are weighted for their importance as shown in Table 2.7. Cell performance estimates for each theoretical surface chemistries are compared to the cell performance of a target

environment, laminin. This provides a cell performance index (CPI) with one value, between 0 and 1, and the closer this index is to 0 the closer the performance is to the target's. This makes it possible to sort the chemical designs in a desired order and select the ones of interest. For example, best performing environments where cell estimates are closer mathematically to laminin's.

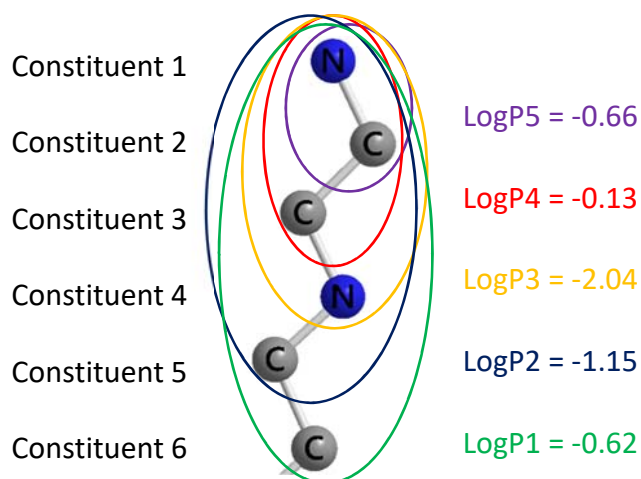


Figure 5.6: LogP values for molecule constituents. LogP5 value is the logP value for the terminal group. Moving down to logP1 being the logp value for up to the top 6 constituents of the molecule.

#### 5.2.2.2 From theoretical to synthesisable surface chemistries

After finding the cell performance index (CPI) of chemical designs, the next step was to select the ones performing equal or better (smaller index value) than our synthetic best environment for each time point (3-methoxy for day 3 or amine for day 7). The results were in the thousands for both time points combined. These however are theoretical surface chemistries, some of their chemical values may be near the real ones, and not all of them can be synthesised. Good candidate results were filtered for the unique values they can take for each model input. The results from the computational experiments provide the inclusion list of “good” range of unique chemical values to design surface chemistries (Table 5.2). These were used to redraw the surface chemistries, and these were fed into predictive models again to obtain new cell performance estimates and CPI.

Results are filtered using Python (v3.6) to create this inclusion list then chemistries are re-designed and their chemical values are extracted. Using the pKa and logP values for each, the head group and possible side chains of the molecule was first drawn. From that, one atom at a time was added to the backbone and possible side branches if necessary. The choice of atoms at each level was directed from the chemical values of results for each variable. Finally, the molecular mass and volume was calculated for the re-constructed chemistries and they were shortlisted only if they matched with those of the theoretical chemistries. Below is a table with the unique values for each model input from surviving chemical designs. Surviving here means chemical designs that are equal or better than the current synthetic best for each time point.

Table 5.2: Value range for model inputs to perform computational cell culture experiments at day 3 time point (A) and day 7 (B). This set of data contains the unique values of “good” performing cell culture environments. The performance for day 3 is predicted be better than the synthetic best (3-methoxy) with cell performance index (CPI)<0.09. For day 7, the predicted cell performance is the same as the synthetic best (amine) with CPI 0.28.

<b>A</b>	<b>LogP1</b>	<b>LogP2</b>	<b>LogP3</b>	<b>LogP4</b>	<b>LogP5</b>	<b>Mol. Mass</b>	<b>Mol. Volume</b>	<b>pKa</b>	<b>D3</b>
<b>Total values = 38</b>	-1.93	-1.65	-2.54	-1.22	-1.22	120	75	7.5	1
	-1.43	-1.15	-2.04	-0.72	-0.72		90	10.5	
	-0.93	-0.65	0.46	1.78	-0.22		120		
	-0.43	-0.15	1.96		0.28				
	0.07	0.35	2.46		0.78				
	0.57	0.85	2.96		1.28				
	1.07	1.35	3.46		2.28				
<b>Count</b>	7	7	7	3	7	1	3	2	1

<b>B</b>	<b>LogP1</b>	<b>LogP2</b>	<b>LogP3</b>	<b>LogP4</b>	<b>LogP5</b>	<b>Mol. Mass</b>	<b>Mol. Volume</b>	<b>pKa</b>	<b>D3</b>
<b>Total values = 33</b>	-1.93	0.85	-0.54	-1.22	-1.22	60	75	4.5	0
	-1.43	1.35	0.46	-0.72	-0.72	75	90	13.5	
	-0.93		1.96	0.28		90	105		
	-0.43		2.46	1.28		105			
	0.07		2.96	2.28		120			
	0.57		3.46	2.78					
<b>Count</b>	6	2	6	6	2	5	3	2	1

### 5.2.2.3 Reassessing discovered surface chemistries

The re-designed surface chemistries have different chemical values than the theoretical ones but are still within the “good” range. Reassessment is necessary as another step in the process to verify chemical designs. The re-designed chemistries are fed into the same predictive models used previously and the cell outputs with their distance to laminin’s was calculated. From the survivors, the next step is to search for chemistries as an off-the-shelf product preferably in the form of self-assembly molecules. The similarity search was conducted in [e-molecules](#) and [ChemSpider](#) with different labile groups and without. At this point, if chemistries are not found inquiries are sent to laboratories to synthesise them.

The above method for finding chemical designs was applied for better/similar to the synthetic best as well as mediocre and bottom performers. The chemistries discovered are shown in the figure below:

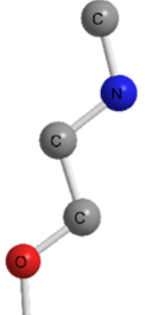
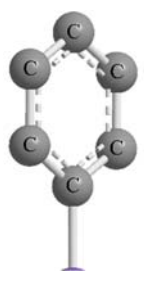
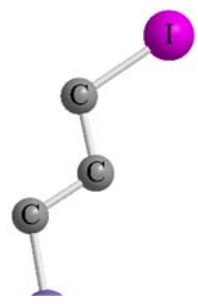
Good	Mediocre	Low
Chem1 - Oxyethanamine	Chem2 - Phenol	Chem3 - Iodine
		
Methyl-2-[(trimethylsilyl)oxy]ethanamine	Phenyl triethoxysilane	3-Iodopropyl trimethoxysilane

Figure 5.7: Discovered off-the-shelf silanes used to change the top chemistry of surfaces for use in tissue engineering. The leftmost is predicted to perform as good as our synthetic best whereas the rightmost is predicted to perform low. The middle chemical is predicted to be somewhere between the top and bottom performer.

The predicted cell performance index of the discovered surface chemistries is shown with the real cell performance index of environments used in cell culture experiments:

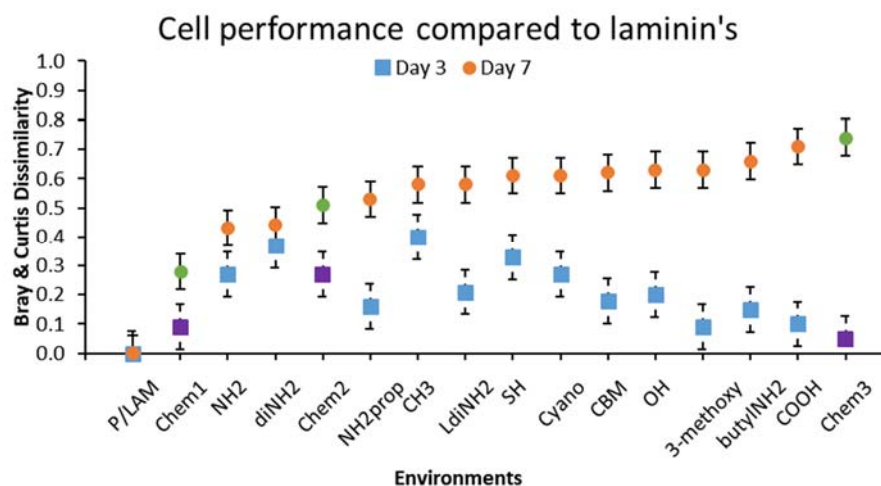


Figure 5.8: Overall cell performance ranks for cell culture environments. Cell response estimates are weighted according to their importance on morphological cell performance. These are passed to a dissimilarity metric called the Bray-Curtis (260,261) to calculate the cell performance index. The error bars indicate the standard deviation of the dissimilarity metric of real data ( $n = 32$ ).

In this section, the “good” chemical designs discovered from computational cell culture experiments have been presented in the form of tables. These tables contain value ranges for chemical parameters after reassessing reconstructed chemical designs. The reason for reconstructing the surface chemistries is due to the conversion from theoretical to “synthesisable” chemical designs. The conversion causes value changes in the chemical parameters and a method to test the resulting chemistries is feeding them in the same computational models used to discover them. Off-the-shelf surface chemistries were discovered and presented along with their predicted cell performance index.

## 5.3 DISCUSSION

### 5.3.1 Model testing

The model in (Figure 5.2, graph A and B) captured the relationship well between the surface chemistry and cell cluster area. Predicting propamine (NH<sub>2</sub>prop) is challenging as the terminal logP and pKa are identical to amine’s, one of the top synthetic environments. Even so, with the additional information (logP) of the top 6 constituents in the surface chemistry,

the model performs well and the highest MPR score for day 7 is 0.26. This is well below the acceptable threshold ( $MPR < 1$ ).

Moving to (Figure 5.2, graphs C and D) for neuron proportion. Carbomethoxy is challenging predict its cell performance more accurately due to its side-chains on the second constituent. The only other environment with side-chains is carboxyl (COOH). Although the cross-validation model performs well on COOH, there is not enough data for these surface chemistries as there is on side-chain-less environments. Propamine (NH<sub>2</sub>prop) has the same issue as with cell cluster area having the same terminal chemistry values as amine on highly influential model inputs, the terminal group logP and pKa. Here, the additional data of the remaining molecule does not help as much as it did predicting cell cluster area. Butylamine (butylNH<sub>2</sub>) MPR for day 7 is the second highest (0.56) and it is believed this is because the terminal logP of this surface chemistry is the same as best performer for this cell performance type, methyl (CH<sub>3</sub>).

In the same Figure 5.2, graphs E and F show type I astrocyte proportion, propamine is once again treated as amine as the position of the amine group is not explicit in the data. This is because the terminal logP value is for at least two atoms excluding hydrogens. The current chemical data representation does not communicate the amine group being on the very top or right below that. Regardless, propamine's MPR is around 0.5 for both time points. Butylamine on the other hand has the highest logP values for logP<sub>5</sub>-logP<sub>2</sub> (Figure 5.6) from all other environments. The closest environment in terms of high logP values is thiol and this model treats it as that. This can be rectified with cell performance data from additional surface chemistries having logP values near butylamine's logP value range. Even so, the MPR for butylamine is  $MPR = 0.4$  for both time points.



In Figure 5.3, graphs A and B show model testing for type II astrocyte cell proportion. The model for both time points show a “flat” prediction as seen previously in cross-validation of chapter 4, figures 4.17 and 4.18. It is challenging to interpret the chemical effect on this cell performance measure from real data and for learning algorithms as well. This is the effect of inadequate data for this rare cell type and chemical parameters resulting in models providing unsmoothed predictions. The model performance ratio for both time points is within the acceptable threshold ( $MPR < 1$ ).

The model for the proportion of unknown type cells (Figure 5.3, graphs C and D) shows a similar problem seen with the model of type II astrocyte proportion. Unsmoothed predictions once again as these were closer to the real values in cross-validation (chapter 4, figures 4.19 and 4.20). The model used the time indicator and the pKa values to predict this cell response. This did provide acceptable MPR for both time points in both training and testing but a confusion is now apparent. The pKa values collected were for groups with both acid dissociation constant (pKa) and base dissociation constant (pKb). Both chemical measures are helpful for predicting whether a species will donate or accept protons at a specific pH value. They describe the degree of ionisation of an acid or base. The pKa values are not necessarily for the terminal group and this confused the models apparent from the predictions above. This explains why propamine is treated as if it is amine.

Neurite length model was tested, and graphs are shown above (Figure 5.3 graphs E, F). In cross-validation model performance, the model fit appeared to be good for both time points (figure 4.21, 4.22). The propamine estimate on day 7 are outside the acceptable range ( $MPR > 1$ ). This is because there are no significant differences between environments from real data. Since there is not enough variation in cell response data, it was difficult for the machine-learning programs used to explain the outcome better. It

could be the learning algorithms cannot tell the amine group is the second atom from the top. Perhaps more predictors or different representation of the chemical data may help.

Model testing for both type I astrocyte area and astrocyte fibre length (Figure 5.4) revealed unsmoothed predictions. In other words, the model estimates vary little as shown by the prediction standard deviation in the above graphs (solid error bars on estimates). This is because there are no significant differences between the scores of these cell performance metrics on any environment on any time point (Figure 5.4). Despite this fact, the models perform well and are within the acceptable prediction threshold ( $MPR < 0.33$ ).

### 5.3.2 Reassessing discovered surface chemistries

The predicted cell performance is different from the real one e.g. amine's day 7 cell performance index (CPI) is 0.43 and it's predicted CPI is 0.28. This is due to optimising the model fit. The models selected are the ones with the lowest prediction error (mean absolute error, MAE). Fitting a model to reduce the MAE value means the fit is sometimes pulled away from some data points and put between a few (Figure 5.9). Reducing overall prediction error meant the model was adjusted in such a way where generalisation performance was maximised. Below is a graph of model fit examples comparing an overfit model (red line) with "better" fit model (green line):

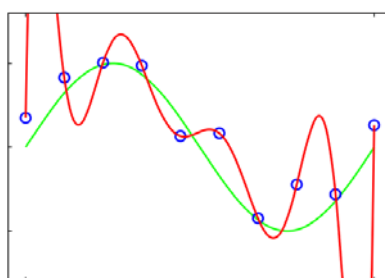


Figure 5.9: Model fit examples comparing an overfit model (red line) with a model with a better generalisation fit (green line).

Figure 5.7 shows oxyethanamine (chem1), the top performer for day 7, is very similar to a fragment of a drug called fluvoxamine (387). This drug has neurogenic properties and functions as selective serotonin reuptake inhibitor typically is used to treat obsessive-compulsive disorder. It is known to restore the balance of serotonin in the brain. Serotonin is a monoamine neurotransmitter (chemical signal) partly synthesised in serotonergic neurons of the central nervous system. Since serotonin has some cognitive functions including memory and learning (plasticity), this chemistry is interesting and warrants deeper investigation with cell culture experiments.

The position of the amine group was discussed in model testing. When the amine group is positioned in the second constituent, this provides longer neurites as well as higher proportion of unknown type cells. Then again, the values of terminal logP and pKa do not communicate the exact position of the atoms for the first 2 constituents until the third one is added. Propamine is the only surface chemistry with this kind of molecular configuration where this problem surfaced. Although propamine's data have been added into the final models, there is no way of knowing if this is resolved until new surface chemistries are tested with cell culture experiments. The fragment of the drug fluvoxamine (387) mentioned in the paragraph above, is identical to the molecule with the amine group on the very top (termination) as shown in the figure below (B).

The mediocre performer in Figure 5.7, phenyl triethoxysilane (chem2) was also found by previous work (41) to perform at a similar level as the predictions from this work. The application for this is shared with this work; surface engineering for tissue engineering using stem cells. The investigators used ventral mesencephalon neurospheres instead of cortex. For previous work (41), the real cell performance index (CPI) for day 7 is 0.11 and 0.25 for amine and phenol respectively. For this work, the predicted CPI is 0.28 and 0.51

respectively. The lower the CPI value the closer the cell performance is to laminin but because of the sampling methodology the CPI varies between the previous work (41) and this study. In addition, phenyl triethoxysilane has also been investigated for controlling the dynamics of cell transition (epithelial-mesenchymal) in heterogeneous cancer cultures (388). Cell transition is a process where epithelial cells lose their cell polarity and cell-cell adhesion and gain migration and invasive properties to become mesenchymal stem cells. Their work (388) is important for developing materials to understand cancer growth, differentiation, and invasion.

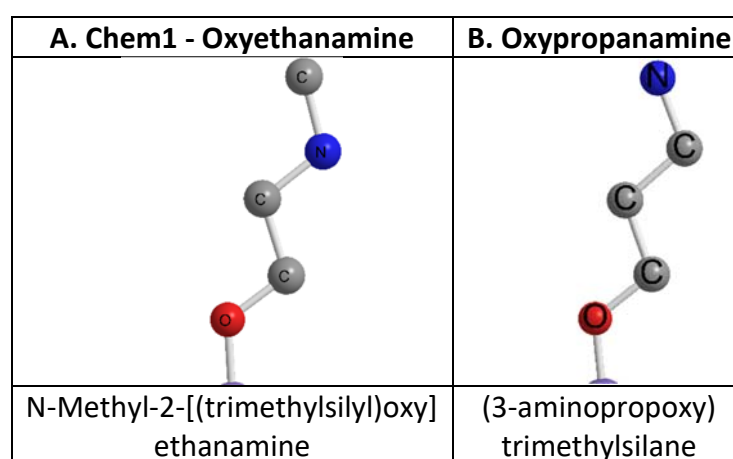


Figure 5.10: Discovered off-the-shelf self-assembly molecules used to change the top chemistry of surfaces for use in tissue engineering. In A is the discovered surface chemistry that should perform similarly to amine and in B is the fragment of a neurogenic drug called fluvoxamine.

The final surface chemistry, 3-Iodopropyl trimethoxysilane (chem3) has been used to modify the surface chemistry of magnetic nanoparticles (MNPs) (388). Functionalised MNPs were coupled with polyethylenimine as a transfection reagent for neural cells. The application of this work was for multimodal MRI-fluorescence imaging and transfection for use in neural cell replacements therapies.

From this work, the above off-the-shelf surface chemistries are already used as fragments of neurogenic drugs, tissue engineering using neural stem cells, material development for

cancer studies, and MRI imaging and transfection for neural cell replacement therapies. These discovered surface chemistries should to be tested in real cell culture experiments to ascertain the model findings. These are going to replace the three new types of environments used in model testing in this chapter. For now, the discovered surface chemistries are educated guesses for the top and bottom discovered surface chemistries (oxyethanamine, iodopropyl). For the mediocre surface chemistry (phenol), the predicted level of cell performance is very close to the real one. This concludes the final experimental chapter.

## 5.4 NOVELTY

Using the model performance ratio (MPR), the smaller this value is the lower the prediction error (mean absolute error). The MPR informs on the generalisation performance of the model and therefore whether it can be used in computational cell culture experiments. The ratio comes by standardising the difference between the prediction and the real observed median with the average standard deviation observed. MPR values between 0 and 0.33 are considered excellent, 0.34 – 0.66 as good, 0.67 – 1 as adequate, and above 1 as needs improving. Table 5.3 shows the models and their highest average MPR for both time points:

Table 5.3: Highest average model performance ratio between two time points for all models.

<b>Model</b>	<b>Performance</b>	<b>Highest average MPR</b>
Cell cluster area	Excellent	0.19
Neuron proportion	Good	0.55
Type I astrocyte proportion	Good	0.39
Type II astrocyte proportion	Good	0.44
Unknown type cell proportion	Good	0.29
Neurite length	Adequate	0.73
Type I astrocyte area	Excellent	0.24
Astrocyte fibre length	Excellent	0.19

The key novelty points are:

- 1) All the models can be used to perform computational cell culture experiments
- 2) The automated solution for computational cell culture experiments screened 15 million chemical designs in under 30 minutes on a quad core laptop with 16 GB of ram using free and open source software
- 3) One discovered chemistry has been used in a previous study with similar observed cell performance as the one predicted by this methodology
- 4) Similar chemistries as two discovered chemistries from this work are used as:
  - a) fragments of neurogenic drugs
  - b) in tissue engineering using neural stem cells
  - c) material development for cancer studies
  - d) MRI imaging and transfection for neural cell replacement therapies

## 6 GENERAL DISCUSSION

---

### 6.1 SURFACE CHEMISTRY OF *IN VITRO* CELL CULTURE ENVIRONMENTS

#### 6.1.1 Lipophobicity

Lipophobic surfaces provide larger cell clusters ( $r = -0.67$  on day 5 (41),  $r = -0.58$  on day 3) and decrease cell density ( $r = 0.79$  on day 5 and 7 (41),  $r = 0.61$  on day 3). Lipophobic surfaces also increase neuron proportion on all three time points from previous work (41). Cell clusters generally spread more on lipophobic surfaces (logP 0.4 to -0.66). This finding does not apply on amine surface chemistries. The lower the logP value the larger the cell cluster area and a similar finding was made in another study (377). It is known lipophobic environments offer enhanced cell adhesion (107,306) and cell spreading (307)

as generally they are conducive to protein adsorption (95,299,308). LogP was found to negatively correlate significantly with cell cluster area in previous (41) and this work with  $r = -0.67 \pm 0.36$  and  $r = -0.58 \pm 0.25$  respectively. Others have found a relationship between the logP of (un-tethered) amino acid surface chemistries and cell spreading (77). Lipophobic environments were found to decrease astrocyte fibre length (AFL) on day 7 ( $r = 0.79$  (41),  $r = 0.49$ ). This is related with cell spreading and from relevant work it was found these environments enhance AFL (307). A limitation of using calculated logP values is that most methods are not tested for their performance on separate test sets (106) but the chosen method for this study was one of the top performer for small molecules.

### 6.1.2 Lipophilicity

Lipophilic surfaces decrease type I astrocyte proportion and area on day 3 ( $r = -0.48$ ,  $r = -0.77$ ). Reducing the proportion of the dominant cell type means the proportion of other(s) has increased. This explains the increase type II astrocyte proportion ( $r = 0.51$ ). In a relevant study (316), embryonic stem cells from mouse and humans differentiated to neurons 1.6 and 2.4 and fold respectively on neutral ultra-low-attachment plates (LAC, 23° WCA) compared to hydrophobic PDMS surfaces (111° WCA). In another study, carbon nanotubes (CNTs) were made hydrophilic using acid treatment. On these environments enhanced laminin adsorption, cell adhesion, and neuron differentiation were observed compared to a standard type surface (poly-l-ornithine) (317).

### 6.1.3 Molecular mass/volume

Molecular mass/volume is related with chemical complexity. High molecular mass/volume was found to lower cell density on day 3 ( $r = -0.62$  (41),  $r = -0.53$ ) but this is largely due to the smaller surface chemistries (OH, CH<sub>3</sub>) having higher cell density influencing

correlation. Surface chemistries reducing cell density are the amines which happen to have more mass (Da). This effect is clearer to understand with laminin environments. Laminin is the surface chemistry with largest molecular mass and volume. A limitation here is that laminin's chemical and cell performance data were excluded from correlation testing as most of laminin's chemical values are extremes and influenced the relationships found. In computational modelling, all of laminin's data was included in the training dataset because the learning algorithms used can find non-linear relationships. This coerces the effect of higher molecular mass/volume relating with enhanced cell responses for exploring candidate surface chemistries.

#### 6.1.4 Surface charge (pKa)

Surface charge (pKa) appeared in the literature for an association with the cell membrane in cell adhesion studies (86). Higher pKa values decrease neuron and type I astrocyte density on both time points ( $r_{Neuron} = -0.68$ ,  $r_{Astrocyte} = -0.61$  in study (41) and  $r_{Astrocyte} = -0.71$  found in this study). High pKa also lowers neuron proportion on day 7 ( $r = -0.52$ ), increases neurite length on day 3 ( $r = 0.51$ ) and cell cluster area. In previous work (41), no pKa correlation was found with cell cluster area on any time point but from this work, positive and significant correlations were found ( $r = 0.53 \pm 0.25$ ) for both time points. This is attributed to a limitation due to the different source of pKa values. The pKa values for previous work (41) were collected using software (ACDlabs) whereas for this work were collected from experimental results in the literature (206,207,218–222,318).



## 6.2 PREDICTIVE MODELS

The computational models for each cell response used all or a few available model inputs to predict each. These inputs include time point indicator, molecular mass, partition coefficient (logP) and pKa. The table below (Table 6.1) shows the model inputs, marked with an X, used predictive each cell response. Following is a discussion of each model input and its known effects on the cell responses investigated.

Table 6.1: Model inputs used by computational models to predict each cell response. CCA stands for cell cluster area, NP for neuron proportion, A1P/A2P for type I and II astrocyte proportion, PUC for proportion of unknown type cells, NL for neurite length, A1A for type I astrocyte area, and AFL for astrocyte fibre length.

<b>Model input</b>	<b>CCA</b>	<b>NP</b>	<b>A1P</b>	<b>A2P</b>	<b>PUC</b>	<b>NL</b>	<b>A1A</b>	<b>AFL</b>
<b>LogP</b>	X	X	X	X		X	X	X
<b>Molecular mass/volume</b>	X	X	X	X		X	X	
<b>pKa</b>	X	X	X	X	X	X	X	
<b>Time point</b>	X	X	X	X	X			X

For cell cluster area and cell density excluding proportion of unknown type cells, what is known experimentally is discussed in the section above (6.1). The model for neurite length uses the pKa, logP and molecular mass/volume as inputs. The relationship of logP and molecular mass/volume with neurite length has not been reported in the literature. Neither has molecular mass/volume and pKa with type I astrocyte area. This could be because these chemical parameters in conjunction with the other inputs for each model share a non-linear relationship with the cell response. A limitation with machine learning is that some variables make more sense to use but in terms of explaining the cell responses these may not be as good as others. All chemical parameters were expected to affect cell performance since the cells sense up to the 6<sup>th</sup> constituent of the surface chemistry found from this work.

### 6.2.1 Model testing

From model testing (section 5.2.1), almost all predictive models perform within the acceptable standards set by biological variation (average standard deviation). This means the selection of data and pre-processing, machine-learning algorithm and parameter tuning were fit for purpose. Cell performance data is required from additional surface chemistries with side-chains to reduce the prediction errors found from neuron proportion model testing on carbomethoxy. Also, data from surface chemistries with higher logP than butylamine is required to reduce type I astrocyte proportion prediction error. A limitation here is that additional cell data is required for rare cell responses such as proportion of type II astrocyte and unknown type cells. The limited data for these makes it challenging to analyse computationally. It is believed that additional data for longer duration cell cultures using a greater number of surface chemistries would benefit predictive modelling.

Another limitation is the current chemical data representation does not communicate the exact position of an atom in the terminal group. This is because obtaining logP values requires at least two functional atoms bonded together. This caused higher prediction errors in model testing for neuron proportion and type I astrocyte proportion. The position of the oxy and amine groups can have profound effects on cell responses. This is discussed in section 5.3.1 and has also been investigated extensively in the literature (95,107,323,389–391). Also, the pKa value recorded is not always for terminal group. This was realised during testing the proportion of unknown type cells and neurite length models on propamine. The models had higher prediction errors compared to other environments in model testing especially for neurite length on day 7. This surface chemistry was treated as if its amine.

Type I astrocyte area and astrocyte fibre length have unsmoothed predictions. This is due to the real data the models were trained from are not significantly different which is a limitation. This means there is not enough variation to find a better pattern in data and factor in a predictive model. The main problem with these two cell responses is the low variance. Even in environments where significant differences are observed in other cell responses, these two do not vary analogously. Perhaps this could be one of the reasons these two cell responses do not receive as much attention in the nerve tissue regeneration literature (194).

### 6.2.2 Predicting new surface chemistries

The cell performance index (CPI) is a value calculated by comparing the cell performance of an environment compared to a target, the biological control. The closer to 0 the better the match with the cell performance of the target. The mediocre surface chemistry (phenol) discovered was also used in cell culture experiments from previous work (41). The CPI of the real data placed phenol as a mediocre performer agreeing with the predictions. This surface chemistry has also been used to control the dynamics of cell transition (epithelial-mesenchymal) in heterogeneous cancer cultures (388). The top performing surface chemistry should have similar cell performance as amine. Discovering better than amine environments requires investigating the predicted top performers. Oxyethanamine or oxypropanamine are two top performers available off the shelf. The former is similar and the latter is identical to a fragment of a drug called fluvoxamine (387) inhibiting serotonin uptake in the brain. The latter surface chemistry has been added to address a limitation. The computational method does not distinguish the atomic position in the first 2 molecule constituents. Also, another shortcoming is the performance of the top and bottom surface chemistries are predictions therefore educated guesses. Until these are

tested with real cell culture experiments there is no other way to assess the prediction accuracy although the mediocre chemistry predictions were very similar to the real values.

### 6.3 SUMMARY

The findings of this work contribute to the body of work concerned with biomaterials and surface engineering. It provides a framework for screening potential tissue engineering environments used in cell therapies for neurodegenerative diseases such as Huntington's and Parkinson's diseases. Through the application of data science methods and techniques this work accelerates the process of discovering tissue engineering environments for the cell therapy in mind. Cell differentiation, migration, and neurite length can be controlled from the surface chemistry alone. To our knowledge, this is the first study investigating the top 6 constituents of the surface chemistry and eight chemical parameters simultaneously. These include the logP (5 part), pKa, molecular mass and volume, and water/decanol contact angle. The aim of this work is to find synthetic environments that perform as they do in biological *in vitro* environments (laminin). After testing for parametric assumptions, correlation tests revealed relationships between surface chemistry and cell responses in pairs. Moved to using multiple (chemical) inputs to describe the relationships in the form of non-linear models predicting the cell responses within the discovered boundary of biological variation (average 1 standard deviation). This means the predictive models can be used in production as part of a mathematical optimisation tool (Get-Chem) where 15 million theoretical surface chemistries were screened in a matter of minutes. The cell performance of theoretical chemistries was compared with laminin's. From the results, 3 surface chemistries were selected where 1 was used in a previous study investigating cell performance and the predictions were within the acceptable boundary (41). For the other 2 discovered chemistries, similar molecules have been found to be used as in neurogenic

drugs, as reagents for tissue engineering with neural stem cells, as a biomaterial for cancer studies, and in MRI imaging and transfection for neural cell replacement therapies. The mathematical optimisation tool developed has demonstrated that it can solve a wide variety of optimisation problems using free and open source software on a modern laptop.

Further studies are required: cell performance data is required from additional surface chemistries with side-chains and higher logP values than butylamine. Additional cell response data is required for rare cell responses such as proportion of type II astrocyte and unknown type cells. The exact position of atoms in the terminal group needs to be communicated in the form of new chemical parameters. The base dissociation constant should be added to complement pKa. The performance of the top and bottom surface chemistries discovered require experimental validation from cell culture experiments. Nevertheless, the findings presented herein provide the first steps towards discovering synthetic environments with simple surface chemistries for nerve tissue engineering.

## 7 REFERENCES

---

1. NHS. Huntington's disease - NHS Choices [Internet]. Department of Health; 2017. Available from: <http://www.nhs.uk/conditions/Huntingtons-disease/Pages/Introduction.aspx>
2. Parkinson's UK. Facts for journalists [Internet]. 2017. Available from: <http://www.parkinsons.org.uk/content/facts-journalists>
3. Fineberg NA, Haddad PM, Carpenter L, Gannon B, Sharpe R, Young AH, et al. The size, burden and cost of disorders of the brain in the UK. *J Psychopharmacol*. 2013 Sep;27(9):761–70.
4. OpenStax College - Anatomy & Physiology. Human central nervous system. 2013.

5. Garrondo. Alzheimer's Disease Education and Referral Center, a service of the National Institute on Aging. 2008;
6. Cao Q, Benton RL, Whittemore SR. Stem cell repair of central nervous system injury. *J Neurosci Res.* 2002 Jun 1;68(5):501–10.
7. NHS. Huntington's disease - Diagnosis. Department of Health; 2017.
8. NHS. Alzheimer's disease - Diagnosis. Department of Health; 2017.
9. NHS. Parkinson's disease - Diagnosis. Department of Health; 2017.
10. NHS. Motor neurone disease - Diagnosis. Department of Health; 2017.
11. NHS. Multiple sclerosis - Diagnosis. Department of Health; 2017.
12. NHS. Parkinson's disease - Treatment - NHS Choices [Internet]. Department of Health; 2017. Available from: <http://www.nhs.uk/Conditions/Parkinsons-disease/Pages/Treatment.aspx>
13. Shin E, Palmer MJ, Li M, Fricker R a. GABAergic neurons from mouse embryonic stem cells possess functional properties of striatal neurons in vitro, and develop into striatal neurons in vivo in a mouse model of Huntington's disease. *Stem Cell Rev.* 2012 Jun;8(2):513–31.
14. Ratner B, Hoffman ASA, Sc D. Biomaterials science: an introduction to materials in medicine. 3rd ed. San Diego, .... 2013.
15. Dai N, Sottile V. Neural Stem Cell Approaches to CNS Repair. *Electron J Biol.* 2008;4(2):79–87.
16. Roach P, Fricker R, Kyriacou T. Computational methods for optimising stem cell differentiation. 2013. 2013;1–3.
17. Badylak SF, Nerem RM. Progress in tissue engineering and regenerative medicine. *Proc Natl Acad Sci U S A.* National Academy of Sciences; 2010 Feb 23;107(8):3285–6.
18. Hollander AP. Cell therapies and regenerative medicine - the dawn of a new age or

- more hype than hope? *Clin Transl Med*. Springer; 2012 Jul 3;1(1):12.
19. Teng YD, Lavik EB, Qu X, Park KI, Ourednik J, Zurakowski D, et al. Functional recovery following traumatic spinal cord injury mediated by a unique polymer scaffold seeded with neural stem cells. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 2002 Mar 5;99(5):3024–9.
  20. Lindvall O, Bjorklund A. Cell therapy for Parkinson disease. *NeuroRx J Am Soc Exp Neurother*. Am. Soc. for Experimental NeuroTherapeutics; 2004 Oct;1(October):382–9.
  21. Newman R, Yoo D, LeRoux M, Danilkovitch-Miagkova A. Treatment of Inflammatory Diseases with Mesenchymal Stem Cells. *Inflamm Allergy - Drug Targets*. 2009 Jun 1;8(2):110–23.
  22. Ralph Brandenberger SB, Campbell A, Fong T, Lapinskas E, Rowley JA. *Cell Therapy Bioprocessing: Integrating Process and Product Development for the Next Generation of Biotherapeutics*. 2011.
  23. Wei X, Yang X, Han Z, Qu F, Shao L, Shi Y. Mesenchymal stem cells: a new trend for cell therapy. *Acta Pharmacol Sin*. Nature Publishing Group; 2013 Jun;34(6):747–54.
  24. Parekkadan B, Milwid JM. Mesenchymal stem cells as therapeutics. *Annu Rev Biomed Eng*. NIH Public Access; 2010 Aug 15;12:87–117.
  25. Evans MJ, Kaufman MH. Establishment in culture of pluripotential cells from mouse embryos. *Nature*. Nature Publishing Group; 1981 Jul 9;292(5819):154–6.
  26. Thomson JA. Embryonic Stem Cell Lines Derived from Human Blastocysts. *Science* (80- ). 1998 Nov 6;282(5391):1145–7.
  27. Stojkovic M, Lako M, Stojkovic P, Stewart R, Przyborski S, Armstrong L, et al. Derivation of human embryonic stem cells from day-8 blastocysts recovered after three-step in vitro culture. *Stem Cells*. 2004 Sep 1;22(5):790–7.
  28. Takahashi K, Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic

- and Adult Fibroblast Cultures by Defined Factors. *Cell*. 2006 Aug 25;126(4):663–76.
29. Northoff H, Flegel WA. Fetal Calf Serum. In: *Encyclopedia of Immunology*. Elsevier; 1998. p. 896–7.
  30. Perrier AL, Tabar V, Barberi T, Rubio ME, Bruses J, Topf N, et al. Derivation of midbrain dopamine neurons from human embryonic stem cells. *Proc Natl Acad Sci U S A. National Academy of Sciences*; 2004 Aug 24;101(34):12543–8.
  31. Cho MS, Lee Y-E, Kim JY, Chung S, Cho YH, Kim D-S, et al. Highly efficient and large-scale generation of functional dopamine neurons from human embryonic stem cells. *Proc Natl Acad Sci U S A. National Academy of Sciences*; 2008 Mar 4;105(9):3392–7.
  32. Kim J-H, Auerbach JM, Rodríguez-Gómez JA, Velasco I, Gavin D, Lumelsky N, et al. Dopamine neurons derived from embryonic stem cells function in an animal model of Parkinson's disease. *Nature. Nature Publishing Group*; 2002 Jul 4;418(6893):50–6.
  33. Kim H-J. Stem cell potential in Parkinson's disease and molecular factors for the generation of dopamine neurons. *Biochim Biophys Acta - Mol Basis Dis*. 2011 Jan;1812(1):1–11.
  34. Casarosa S, Bozzi Y, Conti L. Neural stem cells: ready for therapeutic applications? *Mol Cell Ther. BioMed Central*; 2014;2:31.
  35. Kriks S, Shim J-W, Piao J, Ganat YM, Wakeman DR, Xie Z, et al. Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease. *Nature*. 2011 Nov 6;480(7378):547–51.
  36. Ribes V, Briscoe J. Establishing and interpreting graded Sonic Hedgehog signaling during vertebrate neural tube patterning: the role of negative feedback. *Cold Spring Harb Perspect Biol. Cold Spring Harbor Laboratory Press*; 2009 Aug;1(2):a002014.
  37. Ashall L, Horton CA, Nelson DE, Paszek P, Harper C V., Sillitoe K, et al. Pulsatile Stimulation Determines Timing and Specificity of NF- B-Dependent Transcription.



- Science (80- ). 2009 Apr 10;324(5924):242–6.
38. Yoon S, Kim MG, Chiu CT, Hwang JY, Kim HH, Wang Y, et al. Direct and sustained intracellular delivery of exogenous molecules using acoustic-transfection with high frequency ultrasound. *Sci Rep*. Nature Publishing Group; 2016 Feb 4;6:20477.
  39. Tay S, Hughey JJ, Lee TK, Lipniacki T, Quake SR, Covert MW. Single-cell NF- $\kappa$ B dynamics reveal digital activation and analogue information processing. *Nature*. Nature Research; 2010 Jul 8;466(7303):267–71.
  40. MacArthur BD, Ma'ayan A, Lemischka IR. Systems biology of stem cell fate and cellular reprogramming. *Nat Rev Mol Cell Biol*. Nature Publishing Group; 2009 Sep 9;10(10):672.
  41. Wright R, Roach P, Fricker R. Engineering Surfaces to Control Neurogenesis. Keele University; 2014.
  42. Eiraku M, Watanabe K, Matsuo-Takasaki M, Kawada M, Yonemura S, Matsumura M, et al. Self-Organized Formation of Polarized Cortical Tissues from ESCs and Its Active Manipulation by Extrinsic Signals. *Cell Stem Cell*. 2008;3(5):519–32.
  43. Eiraku M, Takata N, Ishibashi H, Kawada M, Sakakura E, Okuda S, et al. Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature*. Nature Research; 2011 Apr 7;472(7341):51–6.
  44. Lancaster MA, Renner M, Martin C-A, Wenzel D, Bicknell LS, Hurles ME, et al. Cerebral organoids model human brain development and microcephaly. *Nature*. Nature Research; 2013 Aug 28;501(7467):373–9.
  45. Vescovi AL, Reynolds BA, Fraser DD, Weiss S. bFGF regulates the proliferative fate of unipotent (neuronal) and bipotent (neuronal/astroglial) EGF-generated CNS progenitor cells. *Neuron*. 1993 Nov;11(5):951–66.
  46. Singec I, Knoth R, Meyer RP, Maciaczyk J, Volk B, Nikkhah G, et al. Defining the actual sensitivity and specificity of the neurosphere assay in stem cell biology. *Nat*

- Methods. 2006 Oct;3(10):801–6.
47. Kim M, Morshead CM. Distinct populations of forebrain neural stem and progenitor cells can be isolated using side-population analysis. *J Neurosci*. 2003 Nov 19;23(33):10703–9.
  48. Campos LS. Neurospheres: Insights into neural stem cell biology. *J Neurosci Res*. Wiley Subscription Services, Inc., A Wiley Company; 2004 Dec 15;78(6):761–9.
  49. Pickard MR, Barraud P, Chari DM. The transfection of multipotent neural precursor/stem cell transplant populations with magnetic nanoparticles. *Biomaterials*. 2011 Mar;32(9):2274–84.
  50. Cordey M, Limacher M, Kobel S, Taylor V, Lutolf MP. Enhancing the reliability and throughput of neurosphere culture on hydrogel microwell arrays. *Stem Cells*. 2008 Oct;26(10):2586–94.
  51. Bez A, Corsini E, Curti D, Biggiogera M, Colombo A, Nicosia RF, et al. Neurosphere and neurosphere-forming cells: morphological and ultrastructural characterization. *Brain Res*. 2003 Dec 12;993(1–2):18–29.
  52. Sakai Y, Yoshida S, Yoshiura Y, Mori R, Tamura T, Yahiro K, et al. Effect of microwell chip structure on cell microsphere production of various animal cells. *J Biosci Bioeng*. 2010 Aug;110(2):223–9.
  53. Solozobova V, Wyvekens N, Pruszek J. Lessons from the Embryonic Neural Stem Cell Niche for Neural Lineage Differentiation of Pluripotent Stem Cells. *Stem Cell Rev Reports*. 2012 Sep 25;8(3):813–29.
  54. Scadden DT. The stem-cell niche as an entity of action. *Nature*. Nature Publishing Group; 2006 Jun 28;441(7097):1075–9.
  55. Clarke DL, Johansson CB, Wilbertz J, Veress B, Nilsson E, Karlström H, et al. Generalized potential of adult neural stem cells. *Science*. 2000 Jun 2;288(5471):1660–3.

56. Rosso F, Marino G, Giordano A, Barbarisi M, Parmeggiani D, Barbarisi A. Smart materials as scaffolds for tissue engineering. *J Cell Physiol.* 2005 Jun;203(3):465–70.
57. Khan F, Tanaka M. Designing Smart Biomaterials for Tissue Engineering. *Int J Mol Sci.* Multidisciplinary Digital Publishing Institute (MDPI); 2017 Dec 21;19(1).
58. Moroni L, Elisseeff JH. Biomaterials engineered for integration. *Mater Today.* Elsevier; 2008 May 1;11(5):44–51.
59. Guilak F, Cohen DM, Estes BT, Gimble JM, Liedtke W, Chen CS. Control of Stem Cell Fate by Physical Interactions with the Extracellular Matrix. *Cell Stem Cell.* NIH Public Access; 2009 Jul 2;5(1):17–26.
60. Eshghi S, Schaffer D V. Engineering microenvironments to control stem cell fate and function. *StemBook.* Harvard Stem Cell Institute; 2008.
61. Ranieri JP, Bellamkonda R, Bekos EJ, Gardella JA, Mathieu HJ, Ruiz L, et al. Spatial control of neuronal cell attachment and differentiation on covalently patterned laminin oligopeptide substrates. *Int J Dev Neurosci.* 1994 Dec;12(8):725–35.
62. Ehrbar M, Rizzi SC, Hlushchuk R, Djonov V, Zisch AH, Hubbell JA, et al. Enzymatic formation of modular cell-instructive fibrin analogs for tissue engineering. *Biomaterials.* Elsevier; 2007 Sep 1;28(26):3856–66.
63. Adelöw C, Segura T, Hubbell JA, Frey P. The effect of enzymatically degradable poly(ethylene glycol) hydrogels on smooth muscle cell phenotype. *Biomaterials.* Elsevier; 2008 Jan 1;29(3):314–26.
64. Jones MER, Messersmith PB. Facile coupling of synthetic peptides and peptide–polymer conjugates to cartilage via transglutaminase enzyme. *Biomaterials.* Elsevier; 2007 Dec 1;28(35):5215–24.
65. Martin Ehrbar †,‡, Simone C. Rizzi †,‡,||, Ronald G. Schoenmakers §, Blanca San Miguel †, Jeffrey A. Hubbell §, Franz E. Weber † and, et al. Biomolecular Hydrogels Formed and Degraded via Site-Specific Enzymatic Reactions. *American Chemical*

Society ; 2007;

66. Voskerician G, Gingras PH, Anderson JM. Macroporous condensed poly(tetrafluoroethylene). I.In vivo inflammatory response and healing characteristics. *J Biomed Mater Res Part A*. Wiley-Blackwell; 2006 Feb;76A(2):234–42.
67. Hartgerink JD. Self-Assembly and Mineralization of Peptide-Amphiphile Nanofibers. *Science* (80- ). 2001 Nov 23;294(5547):1684–8.
68. Ellis-Behnke RG, Liang Y-X, You S-W, Tay DKC, Zhang S, So K-F, et al. Nano neuro knitting: Peptide nanofiber scaffold for brain repair and axon regeneration with functional return of vision. *Proc Natl Acad Sci*. 2006 Mar 28;103(13):5054–9.
69. Kisiday J, Jin M, Kurz B, Hung H, Semino C, Zhang S, et al. Self-assembling peptide hydrogel fosters chondrocyte extracellular matrix production and cell division: Implications for cartilage tissue repair. *Proc Natl Acad Sci*. 2002 Jul 23;99(15):9996–10001.
70. Bonzani IC, George JH, Stevens MM. Novel materials for bone and cartilage regeneration. *Curr Opin Chem Biol*. Elsevier Current Trends; 2006 Dec 1;10(6):568–75.
71. Bain CD, Whitesides GM. Modeling Organic Surfaces with Self-Assembled Monolayers. *Angew Chemie Int Ed English*. 1989 Apr;28(4):506–12.
72. Chaudhury MK, Whitesides GM. How to make water run uphill. *Science* (80- ). 1992;256(5063):1539–41.
73. Hook AL, Anderson DG, Langer R, Williams P, Davies MC, Alexander MR. High throughput methods applied in biomaterial development and discovery. *Biomaterials*. Elsevier Ltd; 2010 Jan;31(2):187–98.
74. Hook AL, Chang C-Y, Yang J, Scurr DJ, Langer R, Anderson DG, et al. Polymer microarrays for high throughput discovery of biomaterials. *J Vis Exp*. 2012

Jan;(59):e3636.

75. Yang J, Rose FRAJ, Gadegaard N, Alexander MR. A High-Throughput Assay of Cell-Surface Interactions using Topographical and Chemical Gradients. *Adv Mater.* 2009 Jan 19;21(3):300–4.
76. Dowling DP, Miller IS, Ardhaoui M, Gallagher WM. Effect of surface wettability and topography on the adhesion of osteosarcoma cells on plasma-modified polystyrene. *J Biomater Appl.* 2011 Sep 1;26(3):327–47.
77. Rawsterne RE, Todd SJ, Gough JE, Farrar D, Rutten FJM, Alexander MR, et al. Cell spreading correlates with calculated logP of amino acid-modified surfaces. *Acta Biomater.* 2007 Sep;3(5):715–21.
78. Craighead HG, Turner SW, Davis RC, James C, Perez AM, St. John PM, et al. Chemical and Topographical Surface Modification for Control of Central Nervous System Cell Adhesion. *Biomed Microdevices.* Kluwer Academic Publishers; 1998;1(1):49–64.
79. Stenger DA, Pike CJ, Hickman JJ, Cotman CW. Surface determinants of neuronal survival and growth on self-assembled monolayers in culture. *Brain Res.* 1993 Dec;630(1–2):136–47.
80. Nicolau D V., Taguchi T, Tanigawa H, Yoshikawa S. Control of the neuronal cell attachment by functionality manipulation of diazo-naphthoquinone/novolac photoresist surface. *Biosens Bioelectron.* 1996 Jan;11(12):1237–52.
81. Naka Y, Eda A, Takei H, Shimizu N. Neurite outgrowths of neurons on patterned self-assembled monolayers. *J Biosci Bioeng.* 2002 Jan;94(5):434–9.
82. Kleinfeld D, Kahler KH, Hockberger PE. Controlled outgrowth of dissociated neurons on patterned substrates. *J Neurosci.* Society for Neuroscience; 1988 Nov 1;8(11):4098–120.
83. Corey JM, Wheeler BC, Brewer GJ. Micrometer resolution silane-based patterning of hippocampal neurons: critical variables in photoresist and laser ablation processes

- for substrate fabrication. *IEEE Trans Biomed Eng.* 1996 Sep;43(9):944–55.
84. Stenger DA, Georger JH, Dulcey CS, Hickman JJ, Rudolph AS, Nielsen TB, et al. Coplanar molecular assemblies of amino- and perfluorinated alkylsilanes: characterization and geometric definition of mammalian cell adhesion and growth. *J Am Chem Soc.* 1992 Oct;114(22):8435–42.
85. Palyvoda O, Bordenyuk AN, Yatawara AK, McCullen E, Chen C-C, Benderskii A V, et al. Molecular organization in SAMs used for neuronal cell growth. *Langmuir.* 2008 Apr 15;24(8):4097–106.
86. Liu BF, Ma J, Xu QY, Cui FZ. Regulation of charged groups and laminin patterns for selective neuronal adhesion. *Colloids Surf B Biointerfaces.* 2006 Dec 1;53(2):175–8.
87. Kidambi S, Lee I, Chan C. Primary Neuron/Astrocyte Co-Culture on Polyelectrolyte Multilayer Films: A Template for Studying Astrocyte-Mediated Oxidative Stress in Neurons. *Adv Funct Mater.* Wiley-VCH Verlag; 2008 Jan 24;18(2):294–301.
88. Slaughter GE, Bieberich E, Wnek GE, Wynne KJ, Guiseppi-Elie A. Improving neuron-to-electrode surface attachment via alkanethiol self-assembly: an alternating current impedance study. *Langmuir.* 2004 Aug 17;20(17):7189–200.
89. Huttenlocher A, Horwitz AR. Integrins in cell migration. *Cold Spring Harb Perspect Biol.* 2011 Sep 1;3(9):a005074.
90. Roach P, Parker T, Gadegaard N, Alexander MR. Surface strategies for control of neuronal cell adhesion: A review. *Surf Sci Rep.* Elsevier B.V.; 2010 Jun 15;65(6):145–73.
91. Fricker-Gates RA, Gates MA. Stem cell-derived dopamine neurons for brain repair in Parkinson's disease. *Regen Med.* 2010 Mar;5(2):267–78.
92. Roach P, Eglin D, Rohde K, Perry CC. Modern biomaterials: a review - bulk properties and implications of surface modifications. *J Mater Sci Mater Med.* 2007 Jul;18(7):1263–77.

93. Lamalice L, Le Boeuf F, Huot J. Endothelial cell migration during angiogenesis. *Circ Res*. 2007 Mar 30;100(6):782–94.
94. Mei Y, Gerecht S, Taylor M, Urquhart AJ, Bogatyrev SR, Cho S-W, et al. Mapping the Interactions among Biomaterials, Adsorbed Proteins, and Human Embryonic Stem Cells. *Adv Mater*. WILEY-VCH Verlag; 2009 Jul 20;21(27):2781–6.
95. Arima Y, Iwata H. Effect of wettability and surface functional groups on protein adsorption and cell adhesion using well-defined mixed self-assembled monolayers. *Biomaterials*. 2007 Jul;28(20):3074–82.
96. Anselme K, Davidson P, Popa a M, Giazon M, Liley M, Ploux L. The interaction of cells and bacteria with surfaces structured at the nanometre scale. *Acta Biomater*. Acta Materialia Inc.; 2010 Oct;6(10):3824–46.
97. Hall PE, Lathia JD, Caldwell M a, Ffrench-Constant C. Laminin enhances the growth of human neural stem cells in defined culture media. *BMC Neurosci*. 2008 Jan;9(1):71.
98. Clark P, Britland S, Connolly P. Growth cone guidance and neuron morphology on micropatterned laminin surfaces. *J Cell Sci*. 1993;105(1).
99. Rao SS, Winter JO. Adhesion molecule-modified biomaterials for neural tissue engineering. *Front Neuroeng*. Frontiers Media SA; 2009;2:6.
100. Ruoslahti E. RGD and other recognition sequences for integrins. *Annu Rev Cell Dev Biol*. 1996 Nov;12(1):697–715.
101. Ray J, Peterson DA, Schinstine M, Gage FH. Proliferation, differentiation, and long-term culture of primary hippocampal neurons. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 1993 Apr 15;90(8):3602–6.
102. Rodin S, Domogatskaya A, Ström S, Hansson EM, Chien KR, Inzunza J, et al. Long-term self-renewal of human pluripotent stem cells on human recombinant laminin-511. *Nat Biotechnol*. Nature Research; 2010 Jun 30;28(6):611–5.

103. Ren Y-J, Zhang H, Huang H, Wang X-M, Zhou Z-Y, Cui F-Z, et al. In vitro behavior of neural stem cells in response to different chemical functional groups. *Biomaterials*. 2009 Mar;30(6):1036–44.
104. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P, et al. Automated design of ligands to polypharmacological profiles. *Nature*. *Nature Research*; 2012 Dec 12;492(7428):215–20.
105. Kruisselbrink JW, Emmerich MTM, Bäck T, Bender A, IJzerman AP, van der Horst E. Combining Aggregation with Pareto Optimization: A Case Study in Evolutionary Molecular Design. In *Springer Berlin Heidelberg*; 2009. p. 453–67.
106. Mannhold R, Poda GI, Ostermann C, Tetko I V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *J Pharm Sci*. 2009 Mar;98(3):861–93.
107. Arima Y, Iwata H. Effects of surface functional groups on protein adsorption and subsequent cell adhesion using self-assembled monolayers. *J Mater Chem*. 2007;17(38):4079.
108. Woo KM, Chen VJ, Ma PX. Nano-fibrous scaffolding architecture selectively enhances protein adsorption contributing to cell attachment. *J Biomed Mater Res A*. 2003 Nov 1;67(2):531–7.
109. Wilkins MR, Sanchez JC, Gooley AA, Appel RD, Humphery-Smith I, Hochstrasser DF, et al. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol Genet Eng Rev*. 1996;13:19–50.
110. Pavlou MP, Diamandis EP. The cancer cell secretome: A good source for discovering biomarkers? *J Proteomics*. 2010 Sep 10;73(10):1896–906.
111. Skalnikova H, Motlik J, Gadher SJ, Kovarova H. Mapping of the secretome of primary isolates of mammalian cells, stem cells and derived cell lines. *Proteomics*. 2011 Feb;11(4):691–708.



112. Barnes D, Sato G. Serum-free cell culture: a unifying approach. *Cell*. 1980 Dec;22(3):649–55.
113. Looney BM, Chernatynskaya A V, Clare-Salzler MJ, Xia C-Q. Characterization of Bone Marrow-Derived Dendritic Cells Developed in Serum-Free Media and their Ability to Prevent Type 1 Diabetes in Nonobese Diabetic Mice. *J blood Disord Transfus*. NIH Public Access; 2014 Apr;5(4).
114. Hughes C, Radan L, Chang WY, Stanford WL, Betts DH, Postovit L-M, et al. Mass Spectrometry-based Proteomic Analysis of the Matrix Microenvironment in Pluripotent Stem Cell Culture. *Mol Cell Proteomics*. 2012 Dec 1;11(12):1924–36.
115. Vogler E a. Protein adsorption in three dimensions. *Biomaterials*. Elsevier Ltd; 2012 Feb;33(5):1201–37.
116. Roach P, Farrar D, Perry CC. Interpretation of protein adsorption: surface-induced conformational changes. *J Am Chem Soc*. American Chemical Society; 2005 Jun 8;127(22):8168–73.
117. Lord MS, Cousins BG, Doherty PJ, Whitelock JM, Simmons A, Williams RL, et al. The effect of silica nanoparticulate coatings on serum protein adsorption and cellular response. *Biomaterials*. 2006 Oct;27(28):4856–62.
118. Rutishauser U, Acheson A, Hall AK, Mann DM, Sunshine J. The neural cell adhesion molecule - NCAM - as a regulator of cell-cell interactions. *Science (80- )*. American Association for the Advancement of Science; 1988;240(4848):53–8.
119. Stevens CA, Safazadeh L, Berron BJ. Thiol-yne adsorbates for stable, low-density, self-assembled monolayers on gold. *Langmuir*. 2014 Mar 4;30(8):1949–56.
120. Vericat C, Vela ME, Benitez G, Carro P, Salvarezza RC. Self-assembled monolayers of thiols and dithiols on gold: new challenges for a well-known system. *Chem Soc Rev*. The Royal Society of Chemistry; 2010 Apr 26;39(5):1805.
121. Johnson DM, LaFranzo NA, Maurer JA. Creating two-dimensional patterned

- substrates for protein and cell confinement. *J Vis Exp*. 2011 Jan;(55):e3164.
122. Cedervall T, Lynch I, Lindman S, Berggård T, Thulin E, Nilsson H, et al. Understanding the nanoparticle-protein corona using methods to quantify exchange rates and affinities of proteins for nanoparticles. *Proc Natl Acad Sci U S A*. 2007 Feb 13;104(7):2050–5.
  123. Deighan M, Pfaendtner J. Exhaustively sampling peptide adsorption with metadynamics. *Langmuir*. 2013 Jun 25;29(25):7999–8009.
  124. Marshall GM, Lopinski GP, Bensebaa F, Dubowski JJ. Electro-optic investigation of the surface trapping efficiency in n-alkanethiol SAM passivated GaAs(001). *Nanotechnology*. 2011 Jun 10;22(23):235704.
  125. Maoz R, Sagiv J. On the formation and structure of self-assembling monolayers. I. A comparative atr-wettability study of Langmuir—Blodgett and adsorbed films on flat substrates and glass microbeads. *J Colloid Interface Sci*. 1984 Aug;100(2):465–96.
  126. Gun J, Iscovici R, Sagiv J. On the formation and structure of self-assembling monolayers. *J Colloid Interface Sci*. 1984 Sep;101(1):201–13.
  127. Keselowsky BG, Collard DM, García AJ. Integrin binding specificity regulates biomaterial surface chemistry effects on cell differentiation. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 2005 Apr 26;102(17):5953–7.
  128. Barbosa JN, Martins MCL, Freitas SC, Gonçalves IC, Aguas AP, Barbosa MA. Adhesion of human leukocytes on mixtures of hydroxyl- and methyl-terminated self-assembled monolayers: effect of blood protein adsorption. *J Biomed Mater Res A*. 2010 Apr;93(1):12–9.
  129. Chaki NK, Vijayamohan K. Self-assembled monolayers as a tunable platform for biosensor applications. *Biosens Bioelectron*. 2002 Jan;17(1–2):1–12.
  130. Bigelow WC, Pickett DL, Zisman WA. Oleophobic monolayers. *J Colloid Sci*. 1946 Dec;1(6):513–38.

131. Aswal DK, Lenfant S, Guerin D, Yakhmi JV, Vuillaume D. Self assembled monolayers on silicon for molecular electronics. *Anal Chim Acta*. 2006 May 24;568(1–2):84–108.
132. Ulman A. Formation and Structure of Self-Assembled Monolayers. *Chem Rev*. 1996 Jun 20;96(4):1533–54.
133. Lundqvist M, Stigler J, Elia G, Lynch I, Cedervall T, Dawson KA. Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts. *Proc Natl Acad Sci U S A. National Academy of Sciences*; 2008 Sep 23;105(38):14265–70.
134. Mrksich M, Whitesides GM. Using Self-Assembled Monolayers to Understand the Interactions of Man-made Surfaces with Proteins and Cells. *Annu Rev Biophys Biomol Struct*. 1996 Jun;25(1):55–78.
135. Emsley JG, Mitchell BD, Kempermann G, Macklis JD. Adult neurogenesis and repair of the adult CNS with neural progenitors, precursors, and stem cells. *Prog Neurobiol*. 2005 Apr;75(5):321–41.
136. Reynolds BA, Weiss S. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science*. 1992 Mar 27;255(5052):1707–10.
137. Richards LJ, Kilpatrick TJ, Bartlett PF. De novo generation of neuronal cells from the adult mouse brain. *Proc Natl Acad Sci U S A. National Academy of Sciences*; 1992 Sep 15;89(18):8591–5.
138. Kmoch S, Stránecký V, Emes RD, Mitchison HM. Bioinformatic perspectives in the neuronal ceroid lipofuscinoses. *Biochim Biophys Acta. Elsevier B.V.*; 2013 Nov;1832(11):1831–41.
139. Kitano H. Computational systems biology. *Nature. Nature Publishing Group*; 2002 Nov 14;420(6912):206–10.
140. Hasty J, McMillen D, Isaacs F, Collins JJ. Computational studies of gene regulatory

- networks: in numero molecular biology. *Nat Rev Genet*. Nature Publishing Group; 2001 Apr 1;2(4):268–79.
141. Giannitelli SM, Accoto D, Trombetta M, Rainer A. Current trends in the design of scaffolds for computer-aided tissue engineering. Vol. 10, *Acta Biomaterialia*. 2014. p. 580–94.
  142. Aimone JB, Gage FH. Modeling new neuron function: a history of using computational neuroscience to study adult neurogenesis. *Eur J Neurosci*. 2011 Mar;33(6):1160–9.
  143. Semple J, Woolridge N, Lumsden C. Review: in vitro, in vivo, in silico: computational systems in tissue engineering and regenerative medicine. *Tissue Eng*. 2005;11(3).
  144. Azuaje F. Computational discrete models of tissue growth and regeneration. *Brief Bioinform*. 2011 Jan;12(1):64–77.
  145. Kohn J. New approaches to biomaterials design. *Nat Mater*. Nature Publishing Group; 2004 Nov;3(11):745–7.
  146. Bennett KP, Parrado-Hernández E. The Interplay of Optimization and Machine Learning Research. *J Mach Learn Res*. 2006;7(Jul):1265–81.
  147. Andrus C. *Data science*. 2012.
  148. Monsarrat P, Vergnes J-N, Planat-Bénard V, Ravaud P, Kémoun P, Sensebé L, et al. An Innovative, Comprehensive Mapping and Multiscale Analysis of Registered Trials for Stem Cell-Based Regenerative Medicine. *Stem Cells Transl Med*. AlphaMed Press; 2016 Jun;5(6):826–35.
  149. Encinas JM, Michurina TV, Peunova N, Park J-H, Tordo J, Peterson DA, et al. Division-Coupled Astrocytic Differentiation and Age-Related Depletion of Neural Stem Cells in the Adult Hippocampus. *Cell Stem Cell*. 2011;8(5):566–79.
  150. Hayes NL, Nowakowski RS. Dynamics of cell proliferation in the adult dentate gyrus of two inbred strains of mice. *Brain Res Dev Brain Res*. 2002 Mar 31;134(1–2):77–85.

151. Deng W, Aimone JB, Gage FH. New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory? *Nat Rev Neurosci*. Nature Publishing Group; 2010 May 31;11(5):339–50.
152. Aimone JB, Deng W, Gage FH. Resolving New Memories: A Critical Look at the Dentate Gyrus, Adult Neurogenesis, and Pattern Separation. *Neuron*. 2011 May;70(4):589–96.
153. Rabe M, Verdes D, Seeger S. Understanding cooperative protein adsorption events at the microscopic scale: a comparison between experimental data and Monte Carlo simulations. *J Phys Chem B*. American Chemical Society; 2010 May 6;114(17):5862–9.
154. Ercan B, Khang D, Carpenter J, Webster TJ. Using mathematical models to understand the effect of nanoscale roughness on protein adsorption for improving medical devices. *Int J Nanomedicine*. 2013 Jan;8 Suppl 1:75–81.
155. Vasina EN, Paszek E, Nicolau D V. The BAD project: data mining, database and prediction of protein adsorption on surfaces. *Lab Chip*. The Royal Society of Chemistry; 2009 Apr 7;9(7):891–900.
156. Szott LM, Horbett T a. Protein interactions with surfaces: Computational approaches and repellency. *Curr Opin Chem Biol*. Elsevier Ltd; 2011 Oct;15(5):683–9.
157. Peirce SM. Computational and Mathematical Modeling of Angiogenesis. *Microcirculation*. 2008 Jan;15(8):739–51.
158. Schiff JL. *Cellular Automata: A Discrete View of the World*. Wiley-Interscience; 2011;252.
159. Glazier, Graner. Simulation of the differential adhesion driven rearrangement of biological cells. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*. 1993 Mar;47(3):2128–54.
160. Savill NJ, Merks RMH. The Cellular Potts Model in Biomedicine. In: *Single-Cell-Based*

- Models in Biology and Medicine. Basel: Birkhäuser Basel; 2007. p. 137–50.
161. Alber MS, Kiskowski MA, Glazier JA, Jiang Y. On Cellular Automaton Approaches to Modeling Biological Cells. In Springer, New York, NY; 2003. p. 1–39.
  162. Merks RMH, Newman SA, Glazier JA. Cell-Oriented Modeling of in vitro Capillary Development. *Lect Notes Comput Sc.* Springer, Berlin, Heidelberg; 2004;3305:425–34.
  163. Merks RMH, Koolwijk P. Modeling Morphogenesis in silico and in vitro: Towards Quantitative, Predictive, Cell-based Modeling Predictive modeling of cell cultures. *Math Model Nat Phenom.* EDP Sciences; 2009 Jul 11;4(4):149–71.
  164. Merks RMH, Glazier JA. Dynamic mechanisms of blood vessel growth. *Nonlinearity.* IOP Publishing; 2006 Jan 1;19(1):C1–10.
  165. Chavez L, Jozefczuk J, Grimm C, Dietrich J, Timmermann B, Lehrach H, et al. Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. *Genome Res.* Cold Spring Harbor Laboratory Press; 2010 Oct;20(10):1441–50.
  166. Yeo GW, Coufal NG, Liang TY, Peng GE, Fu X-D, Gage FH. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol.* 2009 Feb 11;16(2):130–7.
  167. Wilson A, Laurenti E, Oser G, van der Wath RC, Blanco-Bose W, Jaworski M, et al. Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell.* 2008;135(6):1118–29.
  168. Kiel MJ, He S, Ashkenazi R, Gentry SN, Teta M, Kushner JA, et al. Haematopoietic stem cells do not asymmetrically segregate chromosomes or retain BrdU. *Nature.* 2007 Sep 13;449(7159):238–42.
  169. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell.*

- 2008;133(6):1106–17.
170. Adachi T, Osako Y, Tanaka M, Hojo M, Hollister SJ. Framework for optimal design of porous scaffold microstructure by computational simulation of bone regeneration. *Biomaterials*. 2006;27(21):3964–72.
  171. Cookson S, Ostroff N, Pang WL, Volfson D, Hasty J. Monitoring dynamics of single-cell gene expression over multiple cell cycles. *Mol Syst Biol*. 2005 Nov 22;1(1):E1–6.
  172. Müller F-J, Laurent LC, Kostka D, Ulitsky I, Williams R, Lu C, et al. Regulatory networks define phenotypic classes of human stem cell lines. *Nature*. 2008 Sep 18;455(7211):401–5.
  173. Zaman MH, Kamm RD, Matsudaira P, Lauffenburger D a. Computational model for cell migration in three-dimensional matrices. *Biophys J*. Elsevier; 2005 Aug;89(2):1389–97.
  174. N'Dri N a, Shyy W, Tran-Son-Tay R. Computational modeling of cell adhesion and movement using a continuum-kinetics approach. *Biophys J*. Elsevier; 2003 Oct;85(4):2273–86.
  175. Reekmans K, Praet J, Daans J, Reumers V, Pauwels P, Van der Linden A, et al. Current challenges for the advancement of neural stem cell biology and transplantation research. *Stem Cell Rev*. 2012 Mar;8(1):262–78.
  176. Celiz AD, Smith JGW, Langer R, Anderson DG, Winkler DA, Barrett DA, et al. Materials for stem cell factories of the future. *Nat Mater*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2014 May 21;13(6):570–9.
  177. Quinn GPG, Keough MMJ. Experimental design and data analysis for biologists. *Exp Des data Anal Biol*. 2002;i.
  178. Ishida T, Fukushima H, Mizutani W, Miyashita S, Ogiso H, Ozaki K, et al. Annealing effect of self-assembled monolayers generated from terphenyl derivatized thiols on Au(111). *Langmuir*. 2002;18(1):83–92.

179. Blau A, Weini C, Mack J, Kienle S, Jung G, Ziegler C. Promotion of neural cell adhesion by electrochemically generated and functionalized polymer films. *J Neurosci Methods*. 2001 Nov;112(1):65–73.
180. Yuan Y, Lee TR. Contact Angle and Wetting Properties. In 2013. p. 3–34.
181. Stalder AF, Melchior T, Müller M, Sage D, Blu T, Unser M. Low-bond axisymmetric drop shape analysis for surface tension and contact angle measurements of sessile drops. *Colloids Surfaces A Physicochem Eng Asp*. 2010;364(1–3):72–81.
182. Gillis J. Contact angle. 2006.
183. Haiss W, Thanh NTK, Aveyard J, Fernig DG. Determination of size and concentration of gold nanoparticles from UV-Vis spectra. *Anal Chem*. 2007;79(11):4215–21.
184. Sigma-Aldrich. Gold Nanoparticles: Properties and Applications [Internet]. 2015 [cited 2015 May 11]. Available from: <http://www.sigmaaldrich.com/materials-science/nanomaterials/gold-nanoparticles.html>
185. Pang Y, Zheng B, Kimberly SL, Cai Z, Rhodes PG, Lin RCS. Neuron-oligodendrocyte myelination co-culture derived from embryonic rat spinal cord and cerebral cortex. *Brain Behav*. Wiley-Blackwell; 2012 Jan;2(1):53–67.
186. Arimatsu Y, Miyamoto M, Nihonmatsu I, Hirata K, Uratani Y, Hatanaka Y, et al. Early regional specification for a molecular neuronal phenotype in the rat neocortex. *Proc Natl Acad Sci U S A*. 1992 Oct 1;89(19):8879–83.
187. Reynolds BA, Weiss S. Clonal and Population Analyses Demonstrate That an EGF-Responsive Mammalian Embryonic CNS Precursor Is a Stem Cell. *Dev Biol*. 1996 Apr 10;175(1):1–13.
188. Willaime-Morawek S, Seaberg RM, Batista C, Labbé E, Attisano L, Gorski JA, et al. Embryonic cortical neural stem cells migrate ventrally and persist as postnatal striatal stem cells. *J Cell Biol*. The Rockefeller University Press; 2006 Oct 9;175(1):159–68.



189. Dunnett SB, Björklund A. Dissecting Embryonic Neural Tissues for Transplantation. In 2000. p. 3–25.
190. Shirai K, Lansky AJ, Mintz GS, Costantini CO, Fahy M, Mehran R, et al. Comparison of the angiographic outcomes after beta versus gamma vascular brachytherapy for treatment of in-stent restenosis. Vol. 92, *American Journal of Cardiology*. Cambridge University Press; 2003. p. 1409–13.
191. Kamudzandu M, Yang Y, Roach P, Fricker RA. Efficient alignment of primary CNS neurites using structurally engineered surfaces and biochemical cues. *RSC Adv*. Royal Society of Chemistry; 2015 Feb 27;5(28):22053–9.
192. Jensen JB, Parmar M. Strengths and limitations of the neurosphere culture system. *Mol Neurobiol*. 2006 Dec;34(3):153–61.
193. Bussolati G, Annaratone L, Medico E, D’Armento G, Sapino A. Formalin Fixation at Low Temperature Better Preserves Nucleic Acid Integrity. Wong C-M, editor. *PLoS One*. Churchill Livingstone; 2011 Jun 15;6(6):e21043.
194. Sofroniew M V, Vinters H V. Astrocytes: biology and pathology. *Acta Neuropathol*. Springer; 2010 Jan;119(1):7–35.
195. Field A. *Discovering Statistics Using SPSS*. Vol. 58, *Statistics*. 2009. 821 p.
196. Gilbert C, Mcgregor F, Barnard P. *Asking Questions in Biology: A Guide to Hypothesis Testing, Experimental Design and Presentation in Practical Work and Research Projects*. 3rd ed. Edinburgh: Benjamin Cummings; 2007. 256 p.
197. Gorard S. Revisiting A 90-year-old debate: The advantages of the mean deviation. *Br J Educ Stud*. 2005;53:417–30.
198. Yazici B, Yolacan S. A comparison of various tests of normality. *J Stat Comput Simul*. 2007 Feb;77(2):175–83.
199. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. *J Stat Model Anal*. 2011;2:21–33.

200. Fay DS, Gerow K. A biologist's guide to statistical thinking and analysis. *WormBook*. 2013;1–54.
201. Fang F, Szleifer I. Kinetics and thermodynamics of protein adsorption: a generalized molecular theoretical approach. *Biophys J*. 2001;80(March):2568–89.
202. Lyon A. Why are Normal Distributions Normal? *Br J Philos Sci*. 2013;1990:1–25.
203. Hauke J, Kossowski T. Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *Quaest Geogr*. 2011 Jan 1;30(2).
204. Prajapati B, Dunne M, Armstrong R. Sample size estimation and statistical power analyses. *Optom Today*. 2010;
205. Kuforiji F. The investigation of surface chemical and nanotopographical cues to engineer biointerfaces. Keele University; 2014.
206. Davies MN, Toseland CP, Moss DS, Flower DR. Benchmarking pKa prediction. *BMC Biochem*. 2006;7(1):18.
207. Lee AC, Crippen GM. Predicting pKa. *J Chem Inf Model*. American Chemical Society; 2009 Sep 28;49(9):2013–33.
208. Li H, Robertson AD, Jensen JH. Very fast empirical and rationalization of protein pKa values. *Proteins*. 2005;61:704–21.
209. Kujawski J, Popielarska H, Myka A, Drabińska B, Bernard M. The log P Parameter as a Molecular Descriptor in the Computer-aided Drug Design – an Overview. *Comput Methods Sci Technol*. 2012 Sep 1;18(2):81–8.
210. Mingyu C, Kai G, Jiamou L, Yandao G, Nanming Z, Xiufang Z. Surface modification and characterization of chitosan film blended with poly-L-lysine. *J Biomater Appl*. 2004;19(1):59–75.
211. Ghose AK, Crippen GM. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J Chem Inf Comput Sci*. 1987 Feb;27(1):21–

35.

212. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain. *J Chem Inf Comput Sci.* 1989;29:163–72.
213. Pedretti A, Villa L, Vistoli G. VEGA: a versatile program to convert, handle and visualize molecular structure on Windows-based PCs. *J Mol Graph Model.* 2002 Aug;21(1):47–9.
214. Mackerell AD, Feig M, Brooks CL. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem.* 2004 Aug;25(11):1400–15.
215. Nelson MT, Humphrey W, Gursoy A, Dalke A, Kale L V., Skeel RD, et al. NAMD: a Parallel, Object-Oriented Molecular Dynamics Program. *Int J High Perform Comput Appl.* 1996 Dec 1;10(4):251–68.
216. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *J Mol Graph.* 1996 Feb;14(1):33–8.
217. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. *J Comput Chem.* Wiley Subscription Services, Inc., A Wiley Company; 2005 Dec;26(16):1781–802.
218. Rojas-Hernández A, Ibarra-Montaña EL, Rodríguez-Laguna N, Aníbal Sánchez-Hernández A. Determination of pKa Values for Acrylic, Methacrylic and Itaconic Acids by <sup>1</sup>H and <sup>13</sup>C NMR in Deuterated Water. *J Appl Solut Chem Model.* 2015 Feb 25;4(1):7–18.
219. Hall HKJ. Correlation of the Base Strengths of Amines. *J Am Chem Soc.*

- 1957;79:5441–4.
220. Megias-Alguacil D, Tervoort E, Cattin C, Gauckler LJ. Contact angle and adsorption behavior of carboxylic acids on  $\alpha$ -Al<sub>2</sub>O<sub>3</sub> surfaces. *J Colloid Interface Sci.* 2011 Jan 15;353(2):512–8.
221. Trummal A, Lipping L, Kaljurand I, Koppel IA, Leito I. Acidity of Strong Acids in Water and Dimethyl Sulfoxide. *J Phys Chem A.* American Chemical Society; 2016 May 26;120(20):3663–9.
222. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical p K a Predictions. *J Chem Theory Comput.* 2011 Feb 8;7(2):525–37.
223. Wadsworth JD, Okuno A, Strong PN. Assignment of laminin heavy chains using the lectin Ricinus communis agglutinin-1. *Biochem J.* Portland Press Ltd; 1993 Oct 15;(Pt 2):537–41.
224. Buchwald P, Bodor N. A simple, predictive, structure-based skin permeability model. *J Pharm Pharmacol.* 2001;53:1087–98.
225. Richards FM. Areas, Volumes, Packing, and Protein Structure. *Annu Rev Biophys Bioeng.* Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA; 1977 Jun;6(1):151–76.
226. Timpl R, Rohde H, Robey PG, Rennard SI, Foidart JM, Martin GR. Laminin-a glycoprotein from basement membranes. *J Biol Chem.* 1979 Oct 10;254(19):9933–7.
227. Chen CR, Makhatadze GI. ProteinVolume: calculating molecular van der Waals and void volumes in proteins. *BMC Bioinformatics.* 2015 Dec 26;16(1):101.
228. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software. *ACM SIGKDD Explor Newsl.* 2009 Nov 16;11(1):10.
229. Witten IH, Frank E, Hall MA, Pal CJ. WEKA workbench appendix. In: *Data Mining: Practical Machine Learning Tools and Techniques.* 4th ed. Morgan Kaufmann; 2016.

230. Jović a, Brkić K, Bogunović N. An overview of free software tools for general data mining. 37th Int Conv MIPRO .... 2014;(May):26–30.
231. Witten IJH, Frank E. Data Mining: Practical machine learning tools and techniques. 3rd ed. Vol. 31, ACM SIGMOD Record. Morgan Kaufmann; 2011. 76 p.
232. Hall M. Correlation-based Feature Selection for Machine Learning. Methodology. 1999;21:195–204(April):1–5.
233. Xu LXL, Yan PYP, Chang TCT. Best first strategy for feature selection. [1988 Proceedings] 9th Int Conf Pattern Recognit. IEEE Comput. Soc. Press; 1988;7–9.
234. John GH, Kohavi R, Elgin SF, Pflieger K. Irrelevant Features and the Subset Selection Problem. Mach Learn Proc Elev Int. 1994;121--129.
235. Breiman L. Random forests. Mach Learn. Kluwer Academic Publishers; 2001;45(1):5–32.
236. Frank E, Hall M, Pfahringer B. Locally Weighted Naive Bayes. Proc 19th Conf Uncertain Artif Intell. 2003 Oct 19;249–56.
237. Atkeson CG, Moore AW, Schaal S, Moore AW, Schaal S. Locally Weighted Learning. Artif Intell. Kluwer Academic Publishers; 1997;11(1/5):11–73.
238. Breiman L. Bagging Predictors. Mach Learn. Kluwer Academic Publishers-Plenum Publishers; 1996;24(2):123–40.
239. Tin Kam Ho. Random decision forests. Proc 3rd Int Conf Doc Anal Recognit. IEEE Computer Society; 1995;1:278–82.
240. Aha DW, Kibler D, Albert MK. Instance-Based Learning Algorithms. Mach Learn. Kluwer Academic Publishers; 1991 Jan;6(1):37–66.
241. Wu X, Kumar V, Ross QJ, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Vol. 14, Knowledge and Information Systems. 2008. 1-37 p.
242. Welling M. Support Vector Regression. Toronto: Toronto University; 2004. p. 203–24.

243. Smola AJAAJ, Schölkopf B. A Tutorial on Support Vector Regression. Stat Comput. Kluwer Academic Publishers; 2004 Aug;14(3):199–222.
244. Basak D, Pal S, Patranabis DC. Support Vector Regression. Neural Inf Process - Lett Rev. 2007;10:203–24.
245. Shevade SKK, Keerthi SSS, Bhattacharyya C, Murthy KRKRK. Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Networks. 2000;11(5):1188–93.
246. Platt JC. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. Adv kernel methods. 1998;185–208.
247. Ustun B, Melssen WJ, Buydens LMC. Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. Chemom Intell Lab Syst. 2006;81(1):29–40.
248. Perrone MP. Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization. Neural Networks. 1993;
249. Wang Y, Witten IH. Inducing Model Trees for Continuous Classes. Eur Conf Mach Learn. 1997;1–10.
250. Quinlan JR, Quinlan RJ. Learning With Continuous Classes. World Sci. 1992;92:343–8.
251. Holmes G, Hall M, Prank E. Generating rule sets from model trees. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer; 1999. p. 1–12.
252. Furnkranz J, Flach PA. ROC n' rule learning - Towards a better understanding of covering algorithms. Mach Learn. Kluwer Academic Publishers; 2005 Jan;58(1):39–77.
253. Frank E, Wang Y, Inglis S, Holmes G, Witten IH. Using model trees for classification.

- Mach Learn. Kluwer Academic Publishers; 1998;32(1):63–76.
254. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. Elsevier Science Publishers B. V.; 2002 Feb;38(4):367–78.
  255. Friedman JH, Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat*. 2001;29:1189--1232.
  256. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. Bayesian Forecast Dyn Model. 2009;1:1–694.
  257. Holte RC. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. *Mach Learn*. Kluwer Academic Publishers-Plenum Publishers; 1993;11(1):63–90.
  258. Ai WI, Ai WI, Langley P. Induction of One-Level Decision Trees. *Proc NINTH Int Conf Mach Learn*. 1992;233--240.
  259. Snousy MB Al, El-Deeb HM, Badran K, Khilil IA Al. Suite of decision tree-based classification algorithms on cancer gene expression data. *Egypt Informatics J*. 2011;12(2):73–82.
  260. Clarke KR, Somerfield PJ, Chapman MG. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *J Exp Mar Bio Ecol*. 2006;330(1):55–80.
  261. Bray JR, Curtis JT. An Ordination of the upland forest community of southern Wisconsin.pdf. *Ecol Monogr*. 1957;27:325–349.
  262. Burg MG, Wu CF. Differentiation and central projections of peripheral sensory cells with action-potential block in *Drosophila* mosaics. *J Neurosci*. 1986 Oct;6(10):2968–76.
  263. Ronaldson PT, Bendayan R. HIV-1 viral envelope glycoprotein gp120 produces oxidative stress and regulates the functional expression of multidrug resistance protein-1 (Mrp1) in glial cells. *J Neurochem*. 2008 Aug;106(3):1298–313.

264. Wright R. Systematic Study of Neural Stem Cell and Precursor Response to Surface Head Group Functionality. 2013.
265. Roach P, Parker T, Gadegaard N, Alexander MR. A bio-inspired neural environment to control neurons comprising radial glia, substrate chemistry and topography. *Biomater Sci*. 2013;1(1):83–93.
266. Lakard S, Herlem G, Valles-Villareal N, Michel G, Propper A, Gharbi T, et al. Culture of neural cells on polymers coated surfaces for biosensor applications. *Biosens Bioelectron*. 2005 Apr 15;20(10):1946–54.
267. Vickers AJ. Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Med Res Methodol*. 2005 Dec 3;5(1):35.
268. Frank SA. The common patterns of nature. *J Evol Biol*. NIH Public Access; 2009 Aug;22(8):1563–85.
269. Barbato G, Barini EM, Genta G, Levi R. Features and performance of some outlier detection methods. *J Appl Stat*. 2011 Oct;38(10):2133–49.
270. Edmondson JC, Hatten ME. Glial-guided granule neuron migration in vitro: a high-resolution time-lapse video microscopic study. *J Neurosci*. 1987 Jun;7(6):1928–34.
271. Jacques TS, Relvas JB, Nishimura S, Pytela R, Edwards GM, Streuli CH, et al. Neural precursor cell chain migration and division are regulated through different beta1 integrins. *Development*. 1998 Aug;125(16):3167–77.
272. Sadeghi N, Camby I, Goldman S, Gabius HJ, Balériaux D, Salmon I, et al. Effect of hydrophilic components of the extracellular matrix on quantifiable diffusion-weighted imaging of human gliomas: Preliminary results of correlating apparent diffusion coefficient values and hyaluronan expression level. *Am J Roentgenol*. American Roentgen Ray Society; 2003 Jul 23;181(1):235–41.
273. Hersel U, Dahmen C, Kessler H. RGD modified polymers: Biomaterials for stimulated



- cell adhesion and beyond. Vol. 24, Biomaterials. 2003. p. 4385–415.
274. Lindholm D, Carroll P, Tzimagiorgis G, Thoenen H. Autocrine-paracrine regulation of hippocampal neuron survival by IGF-1 and the neurotrophins BDNF, NT-3 and NT-4. *Eur J Neurosci*. 1996 Jul;8(7):1452–60.
275. Souza DG, Bellaver B, Raupp GS, Souza DO, Quincozes-Santos A. Astrocytes from adult Wistar rats aged in vitro show changes in glial functions. *Neurochem Int*. 2015 Nov;90:93–7.
276. Gabriel S, Njunting M, Pomper JK, Merschhemke M, Sanabria ERG, Eilers A, et al. Stimulus and Potassium-Induced Epileptiform Activity in the Human Dentate Gyrus from Patients with and without Hippocampal Sclerosis. *J Neurosci*. 2004;24(46).
277. Jiang T, Cadenas E. Astrocytic metabolic and inflammatory changes as a function of age. *Aging Cell*. 2014 Dec;13(6):1059–67.
278. Åkesson E, Wolmer-Solberg N, Cederarv M, Falci S, Odeberg J. Human neural stem cells and astrocytes, but not neurons, suppress an allogeneic lymphocyte response. *Stem Cell Res*. 2009 Jan;2(1):56–67.
279. Yang Y, Kamudzandu M, Roach P, Fricker R. Nanofibrous scaffolds supporting optimal central nervous system regeneration: an evidence-based review. *J Neurorestoratology*. Dove Press; 2015 Dec 2;3:123.
280. Wang F, Hao H, Zhao S, Zhang Y, Liu Q, Liu H, et al. Roles of activated astrocyte in neural stem cell proliferation and differentiation. *Stem Cell Res*. 2011 Jul;7(1):41–53.
281. Rayment EA, Williams DJ. Mind the Gap: Challenges in Characterising and Quantifying Cell- and Tissue-Based Therapies for Clinical Translation. *Stem Cells*. Wiley Subscription Services, Inc., A Wiley Company; 2010 May 1;28(5):N/A-N/A.
282. Politis M, Oertel WH, Wu K, Quinn NP, Pogarell O, Brooks DJ, et al. Graft-induced dyskinesias in Parkinson's disease: High striatal serotonin/dopamine transporter ratio. *Mov Disord*. 2011 Sep;26(11):1997–2003.

283. Weiss P. In vitro experiments on the factors determining the course of the outgrowing nerve fiber. *J Exp Zool*. Wiley Subscription Services, Inc., A Wiley Company; 1934 Aug 1;68(3):393–448.
284. Yang F, Murugan R, Wang S, Ramakrishna S. Electrospinning of nano/micro scale poly(L-lactic acid) aligned fibers and their potential in neural tissue engineering. *Biomaterials*. 2005 May;26(15):2603–10.
285. Ioannidou K, Anderson KI, Strachan D, Edgar JM, Barnett SC. Astroglial-axonal interactions during early stages of myelination in mixed cultures using in vitro and ex vivo imaging techniques. *BMC Neurosci*. 2014;15(1):59.
286. Hirrlinger J, Hulsmann S, Kirchhoff F. Astroglial processes show spontaneous motility at active synaptic terminals in situ. *Eur J Neurosci*. Blackwell Science Ltd; 2004 Oct 1;20(8):2235–9.
287. Dziewulska D, Jamrozik Z, Podlecka A, Rafałowska J. Do astrocytes participate in rat spinal cord myelination? *Folia Neuropathol*. 1999;37(2):81–6.
288. Meyer-Franke A, Shen S, Barres BA. Astrocytes Induce Oligodendrocyte Processes to Align with and Adhere to Axons. *Mol Cell Neurosci*. 1999 Oct;14(4–5):385–97.
289. Dagenais C, Avdeef A, Tsinman O, Dudley A, Beliveau R. P-glycoprotein deficient mouse in situ blood–brain barrier permeability and its prediction using an in combo PAMPA model. *Eur J Pharm Sci*. 2009 Sep 10;38(2):121–37.
290. Avdeef A, Nielsen PE, Tsinman O. PAMPA—a drug absorption in vitro model. *Eur J Pharm Sci*. 2004 Aug;22(5):365–74.
291. Sinkó B, Kökösi J, Avdeef A, Takács-Novák K. A PAMPA study of the permeability-enhancing effect of new ceramide analogues. In: *Chemistry and Biodiversity*. 2009. p. 1867–74.
292. Epa VC, Yang J, Mei Y, Hook AL, Langer R, Anderson DG, et al. Modelling human embryoid body cell adhesion to a combinatorial library of polymer surfaces. *J Mater*

Chem. Royal Society of Chemistry; 2012 Sep 18;22(39):20902.

293. Lim JY, Donahue HJ. Cell Sensing and Response to Micro- and Nanostructured Surfaces Produced by Chemical and Topographic Patterning. *Tissue Eng.* Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2007 Aug 14;13(8):1879–91.
294. Mager MD, LaPointe V, Stevens MM. Exploring and exploiting chemistry at the cell surface. *Nat Chem.* Nature Publishing Group; 2011 Jul 22;3(8):582–9.
295. Seo M, Kim J-H, Cho Y-E, Baek M-C, Suk K. Hypothermic regulation of astrocyte proteome profile in experimental stroke. *Electrophoresis.* 2012 Dec 1;33(24):3835–48.
296. Altman J, Bayer SA. Development of the brain stem in the rat. V. Thymidine-radiographic study of the time of origin of neurons in the midbrain tegmentum. *J Comp Neurol.* Wiley Subscription Services, Inc., A Wiley Company; 1981 Jun 1;198(4):677–716.
297. Potter SM, DeMarse TB. A new approach to neural cell culture for long-term studies. *J Neurosci Methods.* 2001 Sep;110(1–2):17–24.
298. Colak S, Tew GN. Amphiphilic Polybetaines: The Effect of Side-Chain Hydrophobicity on Protein Adsorption. *Biomacromolecules.* American Chemical Society; 2012 May 14;13(5):1233–9.
299. Sigal GB, Mrksich M, Whitesides GM. Effect of Surface Wettability on the Adsorption of Proteins and Detergents. *J Am Chem Soc.* American Chemical Society; 1998 Apr;120(14):3464–73.
300. Schulte VA, Hu Y, Diez M, Bünger D, Möller M, Lensen MC. A hydrophobic perfluoropolyether elastomer as a patternable biomaterial for cell culture and tissue engineering. *Biomaterials.* 2010 Nov;31(33):8583–95.
301. Ai H, Meng H, Ichinose I, Jones SA, Mills DK, Lvov YM, et al. Biocompatibility of layer-

- by-layer self-assembled nanofilm on silicone rubber for neurons. *J Neurosci Methods*. Elsevier; 2003 Sep 30;128(1–2):1–8.
302. Bartneck M, Schulte VA, Paul NE, Diez M, Lensen MC, Zwadlo-Klarwasser G. Induction of specific macrophage subtypes by defined micro-patterned structures. *Acta Biomater*. 2010 Oct;6(10):3864–72.
303. Bauer S, Park J, Mark K von der, Schmuki P. Improved attachment of mesenchymal stem cells on super-hydrophobic TiO<sub>2</sub> nanotubes. *Acta Biomater*. 2008 Sep;4(5):1576–82.
304. Gottenbos B, van der Mei HC, Klatter F, Nieuwenhuis P, Busscher HJ. In vitro and in vivo antimicrobial activity of covalently coupled quaternary ammonium silane coatings on silicone rubber. *Biomaterials*. 2002 Mar;23(6):1417–23.
305. Nisbet DR, Pattanawong S, Nunan J, Shen W, Horne MK, Finkelstein DI, et al. The effect of surface hydrophilicity on the behavior of embryonic cortical neurons. *J Colloid Interface Sci*. 2006 Jul;299(2):647–55.
306. Lampin M, Warocquier-Clérout R, Legris C, Degrange M, Sigot-Luizard MF. Correlation between substratum roughness and wettability, cell adhesion, and cell migration. *J Biomed Mater Res*. John Wiley & Sons, Inc.; 1997 Jul 1;36(1):99–108.
307. Ponsonnet L, Reybier K, Jaffrezic N, Comte V, Lagneau C, Lissac M, et al. Relationship between surface properties (roughness, wettability) of titanium and titanium alloys and cell behaviour. *Mater Sci Eng C*. 2003 Jun;23(4):551–60.
308. Rasi Ghaemi S, Harding F, Delalat B, Vasani R, Voelcker NH. Surface Engineering for Long-Term Culturing of Mesenchymal Stem Cell Microarrays. *Biomacromolecules*. American Chemical Society; 2013 Aug 12;14(8):2675–83.
309. Corey JM, Gertz CC, Wang B-S, Birrell LK, Johnson SL, Martin DC, et al. The design of electrospun PLLA nanofiber scaffolds compatible with serum-free growth of primary motor and sensory neurons. *Acta Biomater*. 2008 Jul;4(4):863–75.

310. Lewandowska K, Balachander N, Sukenik CN, Culp LA. Modulation of fibronectin adhesive functions for fibroblasts and neural cells by chemically derivatized substrata. *J Cell Physiol*. Wiley Subscription Services, Inc., A Wiley Company; 1989 Nov 1;141(2):334–45.
311. Alexander MR, Williams P. Water contact angle is not a good predictor of biological responses to materials. *Biointerphases*. American Vacuum Society ; 2017 Jun 6;12(2):02C201.
312. Celiz a. D, Smith JGW, Patel AK, Langer R, Anderson DG, Barrett D a., et al. Chemically diverse polymer microarrays and high throughput surface characterisation: a method for discovery of materials for stem cell culture. *Biomater Sci*. Royal Society of Chemistry; 2014 May 12;2(11):1604–11.
313. Bedford EE, Boujday S, Humblot V, Gu FX, Pradier C-M. Effect of SAM chain length and binding functions on protein adsorption:  $\beta$ -Lactoglobulin and apo-transferrin on gold. *Colloids Surf B Biointerfaces*. Elsevier B.V.; 2014 Apr 1;116:489–96.
314. Jeon SI, Lee JH, Andrade JD, De Gennes PG. Protein-surface interactions in the presence of polyethylene oxide. I. Simplified theory. *J Colloid Interface Sci*. 1991 Mar;142(1):149–58.
315. Jeon SI, Andrade JD. Protein-surface interactions in the presence of polyethylene oxide. II. Effect of protein size. *J Colloid Interface Sci*. 1991 Mar;142(1):159–66.
316. Valamehr B, Jonas SJ, Polleux J, Qiao R, Guo S, Gschweng EH, et al. Hydrophobic surfaces for enhanced differentiation of embryonic stem cell-derived embryoid bodies. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 2008 Sep 23;105(38):14459–64.
317. Chao T-I, Xiang S, Chen C-S, Chin W-C, Nelson AJ, Wang C, et al. Carbon nanotubes promote neuron differentiation from human embryonic stem cells. *Biochem Biophys Res Commun*. 2009 Jul;384(4):426–30.

318. Li B, Ma Y, Wang S, Moran PM. Influence of carboxyl group density on neuron cell attachment and differentiation behavior: Gradient-guided neurite outgrowth. *Biomaterials*. 2005 Aug;26(24):4956–63.
319. Chao T-I, Xiang S, Lipstate JF, Wang C, Lu J. Poly(methacrylic acid)-Grafted Carbon Nanotube Scaffolds Enhance Differentiation of hESCs into Neuronal Cells. *Adv Mater*. WILEY-VCH Verlag; 2010 Aug 24;22(32):3542–7.
320. Lee SJ, Khang G, Lee YM, Lee HB. The effect of surface wettability on induction and growth of neurites from the PC-12 cell on a polymer surface. *J Colloid Interface Sci*. 2003 Mar;259(2):228–35.
321. De Bartolo L, Rende M, Morelli S, Giusi G, Salerno S, Piscioneri A, et al. Influence of membrane surface properties on the growth of neuronal cells isolated from hippocampus. *J Memb Sci*. 2008 Nov;325(1):139–49.
322. Woerly S, Marchand R, Lavallée G. Interactions of copolymeric poly(glyceryl methacrylate)-collagen hydrogels with neural tissue: effects of structure and polar groups. *Biomaterials*. 1991 Mar;12(2):197–203.
323. Webb K, Hlady V, Tresco PA. Relationships among cell attachment, spreading, cytoskeletal organization, and migration rate for anchorage-dependent cells on model surfaces. *J Biomed Mater Res*. NIH Public Access; 2000 Mar 5;49(3):362–8.
324. Sankar S, Mahooti-Brooks N, Hu G, Madri JA. Modulation of cell spreading and migration by pp125FAK phosphorylation. *Am J Pathol*. 1995 Sep;147(3):601–8.
325. Coll JL, Ben-Ze'ev A, Ezzell RM, Rodríguez Fernández JL, Baribault H, Oshima RG, et al. Targeted disruption of vinculin genes in F9 and embryonic stem cells changes cell morphology, adhesion, and locomotion. *Proc Natl Acad Sci U S A*. 1995 Sep 26;92(20):9161–5.
326. Fernández JLR, Geiger B, Salomon D, Ben-Ze'ev A. Overexpression of vinculin suppresses cell motility in BALB/c 3T3 cells. *Cell Motil Cytoskeleton*. 1992;22(2):127–

- 34.
327. Tawil NJ, Wilson P, Carbonetto S. Expression and distribution of functional intergrins in rat CNS glia. *J Neurosci Res.* 1994 Nov 1;39(4):436–47.
328. Horbett TA, Schway MB. Correlations between mouse 3T3 cell spreading and serum fibronectin adsorption on glass and hydroxyethylmethacrylate-ethylmethacrylate copolymers. *J Biomed Mater Res.* 1988 Sep;22(9):763–93.
329. Lanza R, Langer R, Vacanti JP. *Principles of Tissue Engineering, 4th Edition.* 4th ed. Academic Press; 2013. 1776 p.
330. Holt AB, Netoff TIT. Computational modeling of epilepsy for an experimental neurologist. *Exp Neurol.* 2013 Jun;244:75–86.
331. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. *IBM J Res Dev.* 1959 Jul;3(3):210–29.
332. Allen MP. The t test for the simple regression coefficient. In: *Understanding Regression Analysis.* Boston, MA: Springer US; 1997. p. 66–70.
333. O'brien RM. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual Quant.* Springer Netherlands; 2007 Sep 11;41(5):673–90.
334. Mousavizadegan M, Mohabatkar H. An Evaluation on Different Machine Learning Algorithms for Classification and Prediction of Antifungal Peptides. *Med Chem.* 2016;12(8):795–800.
335. Bishop CM. Neural networks for pattern recognition. *J Am Stat Assoc.* 1995;92:482.
336. Dutta JR, Dutta PK, Banerjee R. Optimization of culture parameters for extracellular protease production from a newly isolated *Pseudomonas* sp. using response surface and artificial neural network models. *Process Biochem.* Elsevier; 2004 Oct 29;39(12):2193–8.
337. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial

- neural networks. *Nat Med.* NIH Public Access; 2001 Jun;7(6):673–9.
338. Sweetman MJ, Shearer CJ, Shapter JG, Voelcker NH. Dual Silane Surface Functionalization for the Selective Attachment of Human Neuronal Cells to Porous Silicon. *Langmuir.* American Chemical Society; 2011 Aug 2;27(15):9497–503.
339. Lee MH, Boettiger D, Ducheyne P, Composto RJ. Self-Assembled Monolayers Of Omega-Functional Silanes: A Platform For Understanding Cellular Adhesion At The Molecular Level. In: *Silanes and Other Coupling Agents, Volume 4.* Brill; 2007. p. 163–78.
340. Curran JM, Chen R, Hunt JA. Controlling the phenotype and function of mesenchymal stem cells in vitro by adhesion to silane-modified clean glass surfaces. *Biomaterials.* 2005 Dec;26(34):7057–67.
341. Buxboim A, Rajagopal K, Brown AEX, Discher DE. How deeply cells feel: methods for thin gels. *J Phys Condens Matter.* NIH Public Access; 2010 May 19;22(19):194116.
342. Brunetti V, Maiorano G, Rizzello L, Sorce B, Sabella S, Cingolani R, et al. Neurons sense nanoscale roughness with nanometer sensitivity. *Proc Natl Acad Sci U S A.* 2010 Apr 6;107(14):6264–9.
343. Yan H, Zhang S, He J, Yin Y, Wang X, Chen X, et al. Self-assembled monolayers with different chemical group substrates for the study of MCF-7 breast cancer cell line behavior. *Biomed Mater.* IOP Publishing; 2013 Apr 16;8(3):035008.
344. Yan H, Yuanhao W, Hongxing Y. TEOS/silane coupling agent composed double layers structure: A novel super-hydrophilic coating with controllable water contact angle value. *Appl Energy.* Elsevier; 2017 Jan 1;185:2209–16.
345. Andrić F, Héberger K. Towards better understanding of lipophilicity: Assessment of in silico and chromatographic logP measures for pharmaceutically important compounds by nonparametric rankings. *J Pharm Biomed Anal.* Elsevier; 2015 Nov 10;115:183–91.



346. Smith E, Dent G. Modern Raman Spectroscopy - A Practical Approach. Modern Raman Spectroscopy - A Practical Approach. 2005. 1-210 p.
347. Halvorson RA, Vikesland PJ. Surface-enhanced Raman spectroscopy (SERS) for environmental analyses. Vol. 44, Environmental Science and Technology. 2010. p. 7749–55.
348. Haynes CL, McFarland AD, Duyne RP Van. Surface-Enhanced Raman Spectroscopy. Anal Chem. 2005;77(17):338 A–346 A.
349. Fogarty SW, Patel II, Martin FL, Fullwood NJ. Surface-Enhanced Raman Spectroscopy of the Endothelial Cell Membrane. PLoS One. 2014;9(9):7.
350. Mussi V, Biale C, Visentin S, Barbero N, Rocchia M, Valbusa U. Raman analysis and mapping for the determination of COOH groups on oxidized single walled carbon nanotubes. Carbon N Y. Pergamon; 2010 Oct 1;48(12):3391–8.
351. A. Lorén, J. Engelbrektsson, C. Eliasson §, M. Josefson †, J. Abrahamsson ‡, M. Johansson and, et al. Internal Standard in Surface-Enhanced Raman Spectroscopy. American Chemical Society ; 2004;
352. Metwalli E, Haines D, Becker O, Conzone S, Pantano CG. Surface characterizations of mono-, di-, and tri-aminosilane treated glass substrates. J Colloid Interface Sci. 2006;298:825–31.
353. Ratner BD, Tyler BJ, Chilkoti A. Analysis of biomedical polymer surfaces: polyurethanes and plasma-deposited thin films. Clin Mater. 1993;13(1–4):71–84.
354. Zhu X, Raina AK, Smith MA. Cell cycle events in neurons. Proliferation or death? Am J Pathol. American Society for Investigative Pathology; 1999 Aug;155(2):327–9.
355. Frade JM, Ovejero-Benito MC. Neuronal cell cycle: the neuron itself and its circumstances. Cell Cycle. Taylor & Francis; 2015;14(5):712–20.
356. Madarasz E, Theodosis DT, Poulain DA. In vitro formation of type 2 astrocytes derived from postnatal rat hypothalamus or cerebral cortex. Neuroscience.

- 1991;43(1):211–21.
357. Suidan HS, Nobes CD, Hall A, Monard D. Astrocyte spreading in response to thrombin and lysophosphatidic acid is dependent on the Rho GTPase. *Glia*. 1997 Oct;21(2):244–52.
358. Buffo A, Rite I, Tripathi P, Lepier A, Colak D, Horn A-P, et al. Origin and progeny of reactive gliosis: A source of multipotent cells in the injured brain. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 2008 Mar 4;105(9):3581–6.
359. Buffo A, Vosko MR, Ertürk D, Hamann GF, Jucker M, Rowitch D, et al. Expression pattern of the transcription factor Olig2 in response to brain injuries: implications for neuronal repair. *Proc Natl Acad Sci U S A*. National Academy of Sciences; 2005 Dec 13;102(50):18183–8.
360. Hsiao TW, Tresco PA, Hlady V. Astrocyte spreading and migration on aggrecan–laminin dot gradients. *Biointerphases*. 2018 Feb 11;13(1):01A401.
361. Williams-DeVane CR, Reif DM, Hubal EC, Bushel PR, Hudgens EE, Gallagher JE, et al. Decision tree-based method for integrating gene expression, demographic, and clinical data to determine disease endotypes. *BMC Syst Biol*. BioMed Central; 2013;7(1):119.
362. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. NIH Public Access; 2008 Sep;26(9):1011–3.
363. Chen X, Ishwaran H. Random forests for genomic data analysis. *Genomics*. NIH Public Access; 2012 Jun;99(6):323–9.
364. Qi Y. Random Forest for Bioinformatics. In: *Ensemble Machine Learning*. Boston, MA: Springer US; 2012. p. 307–23.
365. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, et al. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform*. Oxford University Press; 2013 May;14(3):315–26.

366. Berrar D, Bradbury I, Dubitzky W. Instance-based concept learning from multiclass DNA microarray data. *BMC Bioinformatics*. 2006 Feb 16;7(1):73.
367. Good BM, Ha G, Ho CK, Wilkinson MD. OntoLoki: an automatic, instance-based method for the evaluation of biological ontologies on the Semantic Web. 2015 Feb 20;
368. Yang ZR. Biological applications of support vector machines. *Brief Bioinform*. 2004 Dec;5(4):328–38.
369. Cortez P, Embrechts MJ. Using sensitivity analysis and visualization techniques to open black box data mining models. *Inf Sci (Ny)*. 2013;225:1–17.
370. Helma C, Gottmann E, Kramer S. Knowledge discovery and data mining in toxicology. *Stat Methods Med Res*. Sage PublicationsSage CA: Thousand Oaks, CA; 2000 Aug;9(4):329–58.
371. Diplaris S, Tsoumakas G, Mitkas PA, Vlahavas I. Protein Classification with Multiple Algorithms. Springer, Berlin, Heidelberg; 2005;448–56.
372. Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, et al. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, Publ online 12 August 2004; | doi101038/101038/nature02771. Nature Publishing Group; 2004 Aug 12;430(7001):768.
373. Bailis R, Ezzati M, Kammen DM. Mortality and Greenhouse Gas Impacts of Biomass and Petroleum Energy Futures in Africa. *Science (80- )*. 2005;308(5718).
374. Campbell JE, Carmichael GR, Chai T, Mena-Carrasco M, Tang Y, Blake DR, et al. Photosynthetic Control of Atmospheric Carbonyl Sulfide During the Growing Season. *Science (80- )*. 2008;322(5904).
375. Chai G, Wang Y, Yasheng A, Zhao P. Beta 2-adrenergic receptor activation enhances neurogenesis in Alzheimer’s disease mice. *Neural Regen Res*. 2016;11(10):1617.
376. Ochalek A, Szczesna K, Petazzi P, Kobolak J, Dinnyes A. Generation of Cholinergic and

Dopaminergic Interneurons from Human Pluripotent Stem Cells as a Relevant Tool for In Vitro Modeling of Neurological Disorders Pathology and Therapy. *Stem Cells Int.* Hindawi; 2016;2016:5838934.

377. Li H-L, Zhang H, Huang H, Liu Z-Q, Li Y-B, Yu H, et al. The effect of amino density on the attachment, migration, and differentiation of rat neural stem cells in vitro. *Mol Cells*. Korean Society for Molecular and Cellular Biology; 2013 May;35(5):436–43.
378. Turney SG, Bridgman PC. Laminin stimulates and guides axonal outgrowth via growth cone myosin II activity. *Nat Neurosci*. Nature Publishing Group; 2005 Jun;8(6):717–9.
379. Rinne JO, Sahlberg N, Ruottinen H, Någren K, Lehikoinen P. Striatal uptake of the dopamine reuptake ligand [<sup>11</sup>C]beta-CFT is reduced in Alzheimer's disease assessed by positron emission tomography. *Neurology*. 1998 Jan;50(1):152–6.
380. Liesi P, Närvänen A, Soos J, Sariola H, Snounou G. Identification of a neurite outgrowth-promoting domain of laminin using synthetic peptides. *FEBS Lett*. 1989 Feb 13;244(1):141–8.
381. Skubitz AP, Letourneau PC, Wayner E, Furcht LT. Synthetic peptides from the carboxy-terminal globular domain of the A chain of laminin: their ability to promote cell adhesion and neurite outgrowth, and interact with heparin and the beta 1 integrin subunit. *J Cell Biol*. Rockefeller University Press; 1991 Nov 15;115(4):1137–48.
382. Noble M, Albrechtsen M, Møllert C, Lyles J, Bock E, Goidis C, et al. Glial cells express N-CAM/D2-CAM-like polypeptides in vitro. *Nature*. Nature Publishing Group; 1985 Aug 22;316(6030):725–8.
383. Wang X, Ye K, Li Z, Yan C, Ding J. Adhesion, proliferation, and differentiation of mesenchymal stem cells on RGD nanopatterns of varied nanopacings. *Organogenesis*. Taylor & Francis; 2013 Oct 1;9(4):280–6.

384. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. *Briefings in Bioinformatics* Feb 3, 2006 p. 86–112.
385. Monzón-Verona JM, Santana-Martín FJ, García-Alonso S, Montiel-Nelson JA. Electroquasistatic analysis of an electrostatic induction micromotor using the cell method. *Sensors (Basel)*. Multidisciplinary Digital Publishing Institute (MDPI); 2010;10(10):9102–17.
386. Melchels FPW, Bertoldi K, Gabbriellini R, Velders AH, Feijen J, Grijpma DW. Mathematically defined tissue engineering scaffold architectures prepared by stereolithography. *Biomaterials*. Elsevier; 2010 Sep 1;31(27):6909–16.
387. Herrera-Arozamena C, Martí-Marí O, Estrada M, de la Fuente Revenga M, Rodríguez-Franco M. Recent Advances in Neurogenic Small Molecules as Innovative Treatments for Neurodegenerative Diseases. *Molecules*. Multidisciplinary Digital Publishing Institute; 2016 Sep 1;21(9):1165.
388. Yiu HHP, Pickard MR, Olariu CI, Williams SR, Chari DM, Rosseinsky MJ. Fe<sub>3</sub>O<sub>4</sub>-PEI-RITC Magnetic Nanoparticles with Imaging and Gene Transfer Capability: Development of a Tool for Neural Cell Transplantation Therapies. *Pharm Res*. Springer US; 2012 May 2;29(5):1328–43.
389. Pop E, Oniciu DC, Pape ME, Cramer CT. Lipophilicity Parameters and Biological Activity in a Series of Compounds with Potential Cardiovascular Applications. *Croat Chem Acta*. 2004;77(1):301–6.
390. Hayashi T, Tanaka Y, Koide Y, Tanaka M, Hara M. Mechanism underlying bioinertness of self-assembled monolayers of oligo(ethyleneglycol)-terminated alkanethiols on gold: protein adsorption, platelet adhesion, and surface forces. *Phys Chem Chem Phys*. 2012 Aug 7;14(29):10196–206.
391. Martins MCL, Fonseca C, Barbosa MA, Ratner BD. Albumin adsorption on alkanethiols self-assembled monolayers on gold electrodes studied by

- chronopotentiometry. *Biomaterials*. 2003 Sep;24(21):3697–706.
392. Kotthoff L, Thornton C, Hutter F. User Guide for Auto-WEKA. 2017. p. 1–15.
393. Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J Phys Chem A*. 1998 May;102(21):3762–72.
394. Raj V, Rai A, Rawat J. In Silico Design and Computational Study of Novel 1, 3, 4-Thiadiazole Derivatives as Potential Affinity with NA/H Exchanger Receptor for Anticonvulsant Activity. *PharmaTutor*. PharmaTutor Edu Labs; 2014;2(5):113–9.
395. Perrin DD, Dempsey B, Serjeant EP. Prediction of pK a Values of Substituted Aliphatic Acids and Bases. In: pK a Prediction for Organic Acids and Bases. Dordrecht: Springer Netherlands; 1981. p. 27–43.
396. University of California. Lab Manual for Chemistry. In: Zumdahl/Zumdahl, editor. Lab Manual for Chemistry. 6th ed. Brooks Cole; 2002.
397. Hiramatsu H, Osterloh FE. pH-controlled assembly and disassembly of electrostatically linked CdSe-SiO<sub>2</sub> and Au-SiO<sub>2</sub> nanoparticle clusters. *Langmuir*. 2003;19(17):7003–11.
398. Wong SS, Joselevich E, Woolley a T, Cheung CL, Lieber CM. Covalently functionalized nanotubes as nanometre-sized probes in chemistry and biology. *Nature*. 1998;394(6688):52–5.
399. Riccardi D, Schaefer P, Cui Q. PK a calculations in solution and proteins with QM/MM free energy perturbation simulations: A quantitative test of QM/MM protocols. *J Phys Chem B*. 2005;109(37):17715–33.
400. OECS SIDS. SIDS Initian Assessment report for 13th SIAM. 2001.
401. Komatsu K, Akamatsu H, Aonuma S, Jinbu Y, Maekawa N, Takeuchi K. Formation, properties, and reactions of the 1,2:3,4:5,6-tris(bicyclo[2.2.2]octeno)tropylium ion. *Tetrahedron*. Pergamon; 1991 Aug 19;47(34):6951–66.

402. Parambil J V., Poornachary SK, Tan RBH, Heng JYY. Template-induced polymorphic selectivity: the effects of surface chemistry and solute concentration on carbamazepine crystallisation. *CrystEngComm*. The Royal Society of Chemistry; 2014 May 19;16(23):4927–30.
403. Zhu M, Lerum MZ, Chen W. How to prepare reproducible, homogeneous, and hydrolytically stable aminosilane-derived layers on silica. *Langmuir*. NIH Public Access; 2012 Jan 10;28(1):416–23.
404. Gemeinhart RA, Park H, Park K. Pore structure of superporous hydrogels. *Polym Adv Technol*. 2000;11(8–12):617–25.

## 8 APPENDICES

### 8.1 CELL CULTURE SOLUTIONS

Table 8.1: Cell culture media used in this project.

Media	Component	Volume in 50 ml	Source
Neural Progenitor Media (NPC)	Neurobasal	47.8 ml	Gibco (Life-technologies)
	B27 supplement	0.5 ml	
	Penicillin Streptomycin fungizone (PSF)	0.5 ml	Sigma-Aldrich
	L-glutamine	0.125 ml	
	30% glucose	0.375 ml	
	Basic fibroblast growth factor ( $\beta$ FGF)	100 $\mu$ l (20 ng/ml)	Sigma-Aldrich
	Heparin	50 $\mu$ l (5 ng/ml)	
Differentiation media	Neurobasal	42.5 ml	Gibco (Life-technologies)
	Foetal calf serum	5 ml	Biocera
	B27 supplement	0.5 ml	Gibco (Life-technologies)
	Glucose solution	0.375 ml	Sigma-Aldrich
	PSF	0.5 ml	
	L-glutamine	0.125 ml	Sigma-Aldrich

## 8.2 IMMUNOCYTOCHEMISTRY SOLUTIONS

Table 8.2: Immunocytochemistry antibody solutions used in this project.

Solution	Component	Volume in 50 ml	Source
Block solution	Tris Buffer Solution	47.4 ml	
	Triton X	1:500 dilution	Sigma-Aldrich
	Normal Goat Serum (NGS)	1:20 dilution	PAA Laboratories
Primary antibody solution	TBS (1:4 dilution)	49.25 ml	
	Triton X	1:500 dilution	Sigma-Aldrich
	NGS	1:100 dilution	PAA Laboratories
	III- $\beta$ -tubulin antibody (goat host, neuronal microtubule protein, murine target)	1:500 dilution	Cambridge bioscience
	gFAP antibody (rabbit host, glia fibrillary acidic protein, murine target)	1:1000 dilution	DAKO
Secondary solution	TBS (1:4 dilution)	49.2 ml	
	NGS	1:100 dilution	PAA Laboratories
	FITC tagged 490 nm goat anti-mouse (green)	1:300 dilution	Cheshire Sciences
	TRITC tagged 547 nm goat anti-rabbit antibody (red)	1:300 dilution	Cheshire Sciences

## 8.3 MACHINE LEARNING SCHEMES

Table 8.3: Machine learning algorithms explored with their hyper-parameters and value ranges (392). Default is the preset value in WEKA. Entries with citation have been used in the project. Entries marked with an asterisk were used for regression only.

Classifier	Functions		
	Parameter	[Value range]/values	Default
Gaussian Process*	L	[0001, 1]	1
	N	0, 1, 2	0
	K	NormalizedPolyKernel, PolyKernel, Puk, RBFKernel	NormalizedPolyKernel
	E	[2, 5]	0



	L	true, false	false
	S	[1, 10]	0
	O	1	0
	C	[0001, 1]	1
MultilayerPerceptron*	L	1	3
	M	1	2
	B	true, false	false
	H	a, i, o, t	a
	C	true, false	false
	R	true, false	false
	D	true, false	false
LinearRegression*	S	0, 1, 2	0
	C	true, false	false
	R	[1e-7, 10]	1e-7
SimpleLinearRegression*	N/A		
SMOreg* (243,245)	C	5	0
	N	0, 1, 2	0
	I	RegSMOImproved	RegSMOImproved
	V	true, false	false
	K	NormalizedPolyKernel, PolyKernel, Puk (247), RBFKernel	NormalizedPolyKernel
	E	[2, 5]	0
	L	true, false	false
	S	[1, 10]	0
	O	1	0
	G	[0001, 1]	1

### Trees

Classifier	Parameter	Value range	Default
DecisionStump* (257,258)	N/A		
M5P	N	true, false	false
	M	[1, 64]	4
	U	true, false	false
	R	true, false	false
RandomTree* (235)	M	[1, 64]	1
	K	0	0
	K	[2, 32]	2
	depth	0	0
	depth	[2, 20]	2
	N	0	0
	N	[2, 5]	3
REPTree*	U	true, false	false
	M	[1, 64]	2
	V	[1e-5, 1e-1]	1e-3
	L	-1	-1
	L	[2, 20]	2
P	true, false	false	

### Rules

Classifier	Parameter	Value range	Default
DecisionTable*	E	acc, rmse, mae, auc	acc
M5Rules (249–251)	N	true, false	false
	M	[1, 64]	4
	U	true, false	false
	R	true, false	false
ZeroR*		N/A	

### Instance based methods

Classifier	Parameter	Value range	Default
IBk* (240)	E	true, false	false
	K	[1, 64]	1
	X	true, false	false
	F	true, false	false
	I	true, false	false
KStar*	B	[1, 100]	20
	E	true, false	false
	M	a, d, m, n	a

### Ensemble methods

Classifier	Parameter	Value range	Default
RandomForest (235)	I	[2, 256]	10
	K	0	0
	K	[1, 32]	2
	depth	0	0
	depth	[1, 20]	2
Stacking	X	10	10
Vote	R	AVG, PROD, MAJ, MIN, MAN	AVG

### Meta-methods

Classifier	Parameter	Value range	Default
LWL (236,237)	K	[-1, 120]	-1
	A	LinearNNSearch	LinearNNSearch
AdditiveRegression (254)	I	[2, 64]	10
	S	[1, 0.3]	1
Bagging (238)	P	[10, 100]	100
	I	[2, 28]	10
	O	true, false	false
RandomCommittee (248)	I	[2, 64]	10
RandomSubSpace	I	[2, 64]	10
	P	[1, 0]	1

### Attribute evaluation and selection

Classifier	Parameter	Value range	Default
CfsSubsetEval (232)	L	true, false	false

BestFirst (233)	D	0, 1, 2	1
	N	[2, 10]	5
GreedyStepwise (234)	C	true, false	false
	B	true, false	false
	R	true, false	false
	N	[10, 1000]	30

## 8.4 NEURON PROPORTION MODEL

RandomTree

=====

```
LogP4 < -0.21
| MolMass1 < 57.58
| | D3 < 0.5 : 4.16 (9/0.39)
| | D3 >= 0.5 : 8.06 (9/2.16)
| MolMass1 >= 57.58
| | D7 < 0.5
| | | LogP1 < -0.14 : 12.6 (10/18.42)
| | | LogP1 >= -0.14 : 7.89 (10/7.01)
| | | D7 >= 0.5
| | | | LogP4 < -0.37
| | | | pKa < 8.57 : 9.36 (19/25)
| | | | pKa >= 8.57 : 8.51 (10/3.17)
| | | | LogP4 >= -0.37 : 6.33 (8/5.47)
LogP4 >= -0.21
| D3 < 0.5
| | LogP5 < 1.27
| | | LogP5 < 0.03
| | | | LogP2 < -0.11 : 2.13 (9/0.36)
| | | | LogP2 >= -0.11
| | | | | MolVol1 < 106.67
| | | | | | LogP4 < -0.11 : 3.07 (12/0.38)
| | | | | | LogP4 >= -0.11 : 2.9 (5/0.92)
| | | | | | MolVol1 >= 106.67 : 2.56 (14/0.5)
| | | | | | LogP5 >= 0.03 : 4.12 (9/1.79)
| | | | | | LogP5 >= 1.27 : 6.16 (11/0.88)
| | | | | D3 >= 0.5
| | | | | | pKa < 10.86
| | | | | | | LogP5 < 0.03
| | | | | | | | LogP4 < -0.11
| | | | | | | | | LogP3 < -0.82 : 7.19 (9/7)
| | | | | | | | | LogP3 >= -0.82
| | | | | | | | | | MolMass1 < 80.16 : 8.06 (6/0.44)
| | | | | | | | | | MolMass1 >= 80.16 : 7.9 (11/4.99)
| | | | | | | | | | | LogP4 >= -0.11 : 9.88 (6/4.6)
| | | | | | | | | | | LogP5 >= 0.03 : 5.47 (8/4.97)
| | | | | | | | | | | pKa >= 10.86 : 15.99 (5/6.74)
```

Size of the tree : 37

RandomTree

=====

```
MolMass1 < 103.64
| D7 < 0.5
| | LogP4 < 1.54
| | | MolVol1 < 104.41
| | | | MolMass1 < 85.64
```

Model 1: Neuron proportion random forest model. Double click to expand. Each node represents a logic test with the value of an attribute. At the leaves is the outcome. The figures in parenthesis next to the value of each leaf represent the: (number of instances that reached / and the mean squared error).

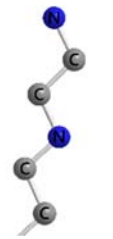


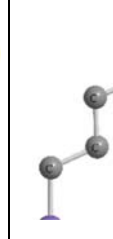
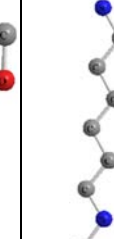


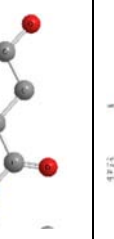
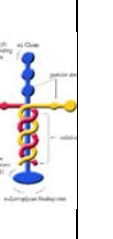
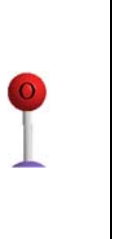
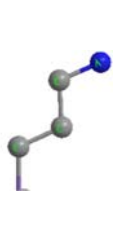
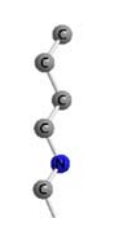
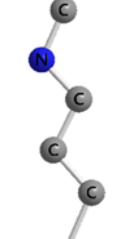
## 8.5 CHEMICAL VALUE TABLE

Table 8.4: Chemical characterisation table. These are the chemical inputs of the cell culture environment. LogP stands for partition coefficient which a lipophilicity measure; pKa stands for acid dissociation constant; WCA/DCA stand for water/decanol (lipid) contact angle.

Functionality	LogP1 ( $\pm 0.16$ ) (393)	LogP2 ( $\pm 0.16$ )	LogP3 ( $\pm 0.16$ )	LogP4 ( $\pm 0.16$ )	LogP5 ( $\pm 0.16$ )	Molecular mass (da) (394)	Molecular volume ( $\text{\AA}^3$ , $\pm 0.99$ ) (394)	pKa	WCA ( $^\circ$ )	DCA ( $^\circ$ )
Methyl (CH <sub>3</sub> )	1.82					30.07	45.15	48 (395)	74.33 $\pm$ 2.97	45.57 $\pm$ 3.22
Carboxyl (COOH)	-1.43	-0.35	0.25	-0.29	-0.54	117.1	103.18	4.87 (396)	70.65 $\pm$ 6.74	44.36 $\pm$ 2.44
Amine (NH <sub>2</sub> )	0.9		0.40	-0.13	-0.66	73.14	91.05	9.27 (397)	92.06 $\pm$ 1.83	50.53 $\pm$ 4.68
Hydroxyl (OH)	-0.72					32.04	36.84	4.5 (398)	72.42 $\pm$ 10.37	71.20 $\pm$ 0.35
Thiol (SH)	2.31		1.78	1.25	0.72	90.19	97.70	10.6 (399)	62.29 $\pm$ 3.97	50.73 $\pm$ 2.05
Diamine (diNH <sub>2</sub> )	-0.62	-1.15	-2.04	-0.13	-0.66	88.15	103.49	10.71 (400)	59.10 $\pm$ 2.2	48.26 $\pm$ 7.15
Cayno	1.15	0.62	0.08	-0.45	-0.25	83.13	95.87	9.21 (401)	62.41 $\pm$ 2.5	71.82 $\pm$ 4.2
Long diamine (L-diNH <sub>2</sub> )	1.46	0.93	0.40	-0.13	-0.66	87.17	108.02	9.27 (397)	65.69 $\pm$ 5.77	44.86 $\pm$ 7.79
3-methoxy	1.51	0.98	0.45	-0.08	-0.72	88.15	105.33	4.5 (398)	63.06 $\pm$ 4.14	57.95 $\pm$ 4.11
P/LAM	-2250.48 $\pm$ 8% (106,213)					810 kda (223)	58.86 nm <sup>3</sup> $\pm$ 2% (227)	12.5 $\pm$ 3.2 (222)	76.19 $\pm$ 4.35	38.54 $\pm$ 2.84

## 8.6 CONTACT ANGLE TABLE

Table 8.5: Chemical characterisation. Contact angle (CA) table for cell culture environments used for this study. The figure followed after  $\pm$  indicate the 1 standard deviation. WCA stands for water contact angle.

Functionality	Diamine (diNH <sub>2</sub> )	Cyano	Thiol (SH)	3-methoxy	Aminohexyl (L-diNH <sub>2</sub> )	Methyl (CH <sub>3</sub> )	Carboxyl (COOH)	P/LAM	Hydroxyl (OH)	Amine (NH <sub>2</sub> )	Butylamine (NH <sub>2</sub> )	Propamine (NH <sub>2</sub> -prop)	Carbomethoxy (CBM)
Chemical structure													
Water CA (°)	59.10 $\pm$ 2.2	62.41 $\pm$ 2.5	62.29 $\pm$ 3.97	63.06 $\pm$ 4.14	65.69 $\pm$ 5.77	74.33 $\pm$ 2.97	70.65 $\pm$ 6.74	76.19 $\pm$ 4.35	72.42 $\pm$ 10.37	92.06 $\pm$ 1.83	74.86 $\pm$ 2.88	64.97 $\pm$ 4.75	66.42 $\pm$ 3.77
Literature WCA (°)	52.3 $\pm$ 0.4 (264)	55.2 $\pm$ 1.7 (402)	74.5 $\pm$ 2.6 (343)	N/A	54 $\pm$ 2 (403)	97.6 $\pm$ 0.05 (343)	40.4 $\pm$ 2.7 (343)	73.3 $\pm$ 3 (301)	49.5 $\pm$ 3 (344)	55.2 $\pm$ 2.8(343)	N/A	N/A	59.5 $\pm$ 2.9 (404)
Decanol CA (°)	48.26 $\pm$ 7.15	71.82 $\pm$ 4.2	50.73 $\pm$ 2.05	57.95 $\pm$ 4.11	44.86 $\pm$ 7.79	45.57 $\pm$ 3.22	44.36 $\pm$ 2.44	38.54 $\pm$ 2.84	71.20 $\pm$ 0.35	50.53 $\pm$ 4.68	68.82 $\pm$ 17.26	44.28 $\pm$ 3.53	90.73 $\pm$ 5.06

## 8.7 MANUSCRIPTS IN PREPARATION

**Joseph G, Roach P, Fricker RA, Kyriacou T (2018).** The effect of partition coefficient ( $\log P$ ) of chemically defined surfaces on cell density.

**Joseph G, Roach P, Fricker RA, Kyriacou T (2018).** The depth of cell sensing of the *in vitro* surface chemistry.

**Joseph G, Roach P, Fricker RA, Kyriacou T (2018).** Tissue engineering with neural stem cells *in silico*.