Swansea University
Prifysgol Abertawe

# Investigating the use of Deep Learning Algorithms to Automatically Score Micronuclei in Human Cell Lines.

Researched by Rachel Barnes (BSc (Hons)), submitted to Swansea University in fulfilment of the requirements for the Degree of MSc in Medical and Healthcare Science by Research.

Swansea University
2023

# Abstract

The *in vitro* micronucleus (MN) assay is a globally used test to quantify DNA damage induced by test chemicals from various industries such as pharmaceuticals, cosmetics and agriculture. Currently, manual scoring is used which is extremely time-consuming and scorer subjective so causes a significant bottleneck in the use of the MN assay. This project shows that imaging flow cytometry coupled with deep learning neural networks can be reliably and accurately used with inter-laboratory function, to automatically score micronucleus events in chemically exposed human B lymphoblastoid cells called TK6 cells. Images were taken from both the cytokinesis-block micronucleus (CBMN) assay and the mononucleate MN assay at Newcastle University. Six different chemicals were tested in this study which are known genotoxic agents and known non-genotoxic agents: aroclor, carbendazim, methyl methanosulphate (MMS), vinblastine, benzo(a)pyrene, D-mannitol. These images were then inputted into a "Deep Flow" neural network, coded in the MATLAB platform which was previously trained on human-scored images assembled from the CBMN assay conducted by Cardiff and Cambridge universities, using MMS and carbendazim treated TK6 cells. Using image data from multiple laboratories in this study provides evidence that the neural network can be used to score unseen data from any laboratory. The neural network correctly scores micronucleus events for both the CBMN and mononucleate MN assays at a percentage confidence of 70% and above. Dose response data for each chemical is parallel to ECVAM guidelines. The aneugen, carbendazim, was shown by the deep learning algorithm to increase the mean dose response by 3.4-fold which shows that as the dose of carbendazim increases, the abundance of micronuclei increases. Further optimisation of the ground truth will prevent underscoring of micronuclei in binucleated cells. It can be concluded that with further optimisation and development of the neural network, this automated platform offers a great potential for the use of the *in vitro* MN assay to be widened. This method has a higher throughput and has the capability to test greater numbers of compounds and chemicals, therefore, this method will be able to keep up with the increasing demand for genotoxicity testing in industrial and pharmaceutical settings.

# Declarations

**DECLARATION**

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed: R. E. BARNES (candidate)

Date: 01/06/23

**STATEMENT 1**

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed: R. E. BARNES (candidate)

Date: 01/06/23

**STATEMENT 2**

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: R. E. BARNES (candidate)

Date: 01/06/23

**STATEMENT 3**

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed: R. E. BARNES (candidate)

Date: 01/06/23

# Contents

# Acknowledgements

Undertaking this Master's by research project has truly been a fantastic experience and one that I will never forget. Throughout the year I have been blessed to have been surrounded by some amazing people who have helped and encouraged me through all times both good and challenging.

Firstly, I would like to thank my mother and father. I am so grateful for their endless support and encouragement to be the best I can be. Without your support I would not have developed nearly as much as a person, and I feel blessed to be called your daughter.

At university, I would like to thank my supervisor Dr George Johnson. I am so grateful for your encouragement and belief in me as I am beginning my science career. You have inspired me to pursue my passion for science and to take opportunities as they arise, and I am so grateful for this. Thank you for always being the other side of emails and answering my questions despite how busy you are. I am so grateful for all of your support and encouragement through this Master's by research project, you have my utmost respect.

I would also like to express my thanks to my secondary supervisor, Professor Paul Rees. Thank you for helping me in my journey of understanding MATLAB programming. You have shown me a completely new method of research and analysis, and I am so grateful for your support. Thank you for your patience and always being happy to zoom call to help me solve the error messages in MATLAB! It has been really eye opening to understand and appreciate the huge potential that artificial intelligence and machine learning has to grow genetic research, so thank you for being there and for your support. I would like to express my gratitude to Dr John Wills from Girton College, Cambridge, for helping me to develop as a scientist and answering my questions and bettering my understanding of the role of deep learning in automating the micronucleus assay.

Lastly, I would like to thank my friends from Swansea University Choral Society, Swansea University Christian Union and Parklands Church for never failing to make me smile and for allowing me to have some time to switch off from the project. You have all made my experience in Swansea so brilliant and one that I will treasure for ever. So to everyone, thank you.

# List of Figures and Tables

## Figures

## Tables

# Abbreviations

**AIC:** Akaike Information Criterion

**BaP:** Benzo(a)pyrene

**BMD:** Benchmark Dose Analysis

**BMDL50:** Benchmark Dose (Lower confidence limit)

**BMDU50:** Benchmark Dose (Upper confidence limit)

**BMR:** Benchmark Response

**CBMN:** Cytokinesis Block Micronucleus Assay

**Cyto-B:** Cytochalasin-B

**ECVAM:** European Centre for the Validation of Alternative Methods

**ICH:** International Council for Harmonisation

**LOEL:** Lowest observable effect level

**MMS:** Methyl Methanosulphate

**MN:** Micronucleus

**MOA:** Mode of Action

**NOGEL:** No observable genotoxic effect level

**OECD:** Organisation for Economic Co-operation and Development

**PCB:** Polychlorinated Biphenyl

**POD:** Point of Departure

**ROS:** Reactive Oxygen Species

# 1. Literature Review

## i) Cell Cycle

The cell cycle has four main stages, G1, synthesis (S), G2, and mitosis (M) (Alberts, 2008). Firstly, in G1, the cell increases in size before undergoing DNA replication in the S stage. In G2, the cell then is prepared for carrying out cytokinesis by checking for any errors in DNA replication and then the cell goes into mitosis. The cell spends the majority of its time in G1, S and G2 phases and this is also called interphase. Then the cell goes into a 5-stage process called mitosis, the 5 stages of mitosis are Prophase, Pro - metaphase, Metaphase, Anaphase and Telophase. When a cell divides, it becomes prone to DNA replication errors despite the multiple cell cycle checkpoints and repair mechanisms. Some of the replication error mechanisms include nucleotide and base excision repair (NER/BER) and mismatch repair (MMR), during which mismatched nucleotides are excised using 3' to 5' exonucleases called Pol ε and Pol δ, then replaced with the correct nucleotides. One named example of a marker of DNA damage in the cell is the production of micronuclei at anaphase (Alberts, 2008).

## ii) Cellular and DNA Damage

Despite the mechanisms described above for ensuring correct alignment of nucleotides, the nucleotide error rate still exists at approximately 1 per 100,000 nucleotides which equates to approximately 120,000 mistakes every time a cell goes through one cell cycle (Pray, 2008).

DNA damage is significant in mutagenesis and carcinogenesis therefore plays a pivotal role in the onset of genetic disease. DNA damage can either be caused by endogenous sources or exogenous sources (De Bont, 2004).

Endogenous DNA damage and mutations can arise from multiple intracellular processes such as DNA replication errors (transition and transversion mutations), transposable genetic elements and metabolic processes such as oxidation, methylation, deamination and depurination of DNA bases and the production of reactive oxygen species (ROS) (Ames et al., 1993). Transition mutations are when a purine is substituted for another purine, or a pyrimidine is replaced by another pyrimidine during DNA replication. These transition mutations can be caused by deamination. For example, deaminated cytosine forms uracil which forms a complementary base pair with adenine in replication, therefore if the deamination is not reversed, a G-C pair will be converted into an A-T pair hence causing a transition mutation. Transversion mutations occur when a purine is substituted with a pyrimidine base and vice versa. ROS such as superoxide and hydroxyl radicals are routinely formed as by-products of metabolic pathways such as aerobic respiration. ROS can lead to DNA damage by oxidizing DNA bases and inducing single and double strand breaks (Palero & Crandall, 2009).

DNA damage can also be caused by exogenous sources which may be exposure to carcinogenic chemicals. These exogenous agents may include alkylating agents, aromatic amines, and radioactive chemicals. Exogenous agents can act as a structural isomers of normal DNA bases, for example, 5-bromouracil is a structural isomer of thymine so can cause a mis-pair. Alkylating agents such as ethylmethanosulfonate can bind to DNA bases and alter their structure which then leads to point mutations. Finally, radiation can cause the ionization and excitation of molecules hence lead to the production of ROS which can then damage DNA structurally. This may lead to the formation of apurinic/apyrimidinic sites or single and double strand breaks. The consequence may be point, frameshift, deletion, or duplication mutations (Griffiths et al., 2000).

There is a positive correlation between the presence of mutations and the induction of cancer, which suggests that mutagenesis drives tumour progression and metastasis. It is known that

carcinogenesis events can be started by mutations arising in genes essential for genetic stability in cells, for example, P53 which is associated with cell cycle regulation and Bcl-2 associated with apoptosis. Consequently, detecting chemicals and compounds that cause DNA mutations is essential for the protection of populations exposed to these (Loeb & Loeb, 2000). To fully test the ability of a chemical to induce mutations, DNA damage must be assessed on three levels; gene mutation, clastogenicity which covers structural chromosome aberrations and aneuploidy which assess numerical chromosome aberrations (COM, 2011). Therefore, to allow for this, genotoxicity testing has been developed to investigate the mutagenicity and hence carcinogenicity of new compounds and chemicals used in multiple industries such as cosmetics and pharmaceuticals. Regulatory bodies such as The Organisation of Economic Co-operation and Development (OECD) and The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) have published guidelines and assays for reliable and accurate genotoxicity testing.

### iii) Cancer

As genotoxicity testing has progressed over the last three to four decades, there has been a paradigm shift, showing a connection between mutagenesis, initiated by exposure to genotoxins, and carcinogenesis, the initiation of cancer formation. In 1966, Malling added a chemical hydroxylating mixture to known carcinogens; diethylnitrosamine and dimethylnitrosamine. These two compounds were found to not be mutagenic *in vitro,* however the resulting metabolites from the reaction were found to be mutagenic in the *Neurospora crassa* fungus (Malling, 1966). Later in 1971, Malling conducted another experiment of adding dimethylnitrosamine to the supernatant from mouse liver homogenate with additional cofactors (activation mixture) and again dimethylnitrosamine was found to be mutagenic in *Salmonella typhimurium* bacteria in a liquid culture (Malling, 1971). Then in 1972, Ames et al. conducted a plate incorporation assay in *Salmonella* bacteria, and this demonstrated that the metabolites of known carcinogenic agents were also mutagenic (Ames et al., 1972). This paradigm shift was further demonstrated in 1973 when Ames et al. performed a similar experiment however used a rat liver homogenate with cofactors as a metabolic activation mixture. Compounds that were previously determined as carcinogens but not mutagens were found to in fact be mutagenic. These assays led to the hypothesis that many carcinogens are innately or after metabolic activation are mutagenic, thus mutagenesis plays a pivotal role in carcinogenesis (Ames et al., 1973).

A variety of mutagens have shown to have a similar 'mutation spectrum' which means that they induce the same base substitution mutation in multiple phylogenies. Therefore, if these mutations arise in cancer-related genes (oncogenes or tumour-suppressor genes), they can drive the formation of tumours. For example, exposure to aflatoxin B, produced by many species of *Aspergillus* fungus, is strongly associated with G – T mutations in codon 259 of the *TP53* gene in liver tumours. When a mutation occurs in the *TP53* gene, it causes the under expression of tumour protein 53, a key regulator of the cell cycle. Therefore, when there is reduced tumour protein 53, cells can rapidly divide and metastasised hence leading to cancerous tumours being produced (Baan et al., 2019).

Tumours are extremely heterogeneous which means that they can contain a wide variety of different mutations, chromosomal aberrations, and aneuploidy (Parsons, 2008). However, there is now also evidence that epigenetic changes can drive carcinogenesis, therefore, genotoxic agents that induce epigenetic changes are significant in the carcinogenesis pathway. For example, mutations in DNA demethyltransferases; *DNMT1* and *DNMT3A* are associated with colorectal cancer and acute myeloid leukaemia as these mutations will cause DNA to not be methylated in the epigenetic interface (Ren et al., 2017). Additionally, mutations in histone lysine methyltransferases

(responsible for methylation of histone tails in chromatin) *HK4* and *H3K9* are present in kidney and colon cancers. Finally, mutations in histone acetyltransferases (responsible for acetylation of histone tails in chromatin) *H3K18* and *H3K27* are implicated in acute lymphoblastic leukaemia (Peltomäki, 2012).

Recognising the role of chemical and biological genotoxic agents in carcinogenesis can contribute to a model of 'agent-induced carcinogenesis'. This model can clearly illustrate the role of a genotoxic agent in carcinogenesis by firstly identifying whether it causes changes on a mutagenic or epigenetic level, the resulting changes in gene expression then the phenotypic changes it induces in the cancer cells. These are known as the hallmarks of cancer and include; mutations in oncogenes, altered gene expression, changes in cell signalling pathways, altered cell growth, evasion of apoptosis (programmed cell death), sustained angiogenesis (blood vessel production), increased genomic instability and metastasis (Hanahan & Weinberg, 2011).

## 1.1 Genetic Toxicology

Due to the wide variety of inducible mutations and genetic events from test chemicals on DNA, multiple *in vitro* and *in vivo* tests have been developed to screen for genotoxicity on the three levels listed above. Figure 1 shows the drug development pipeline and the genotoxicity tests that new pharmaceuticals are subjected to test for both carcinogenicity and genotoxicity.

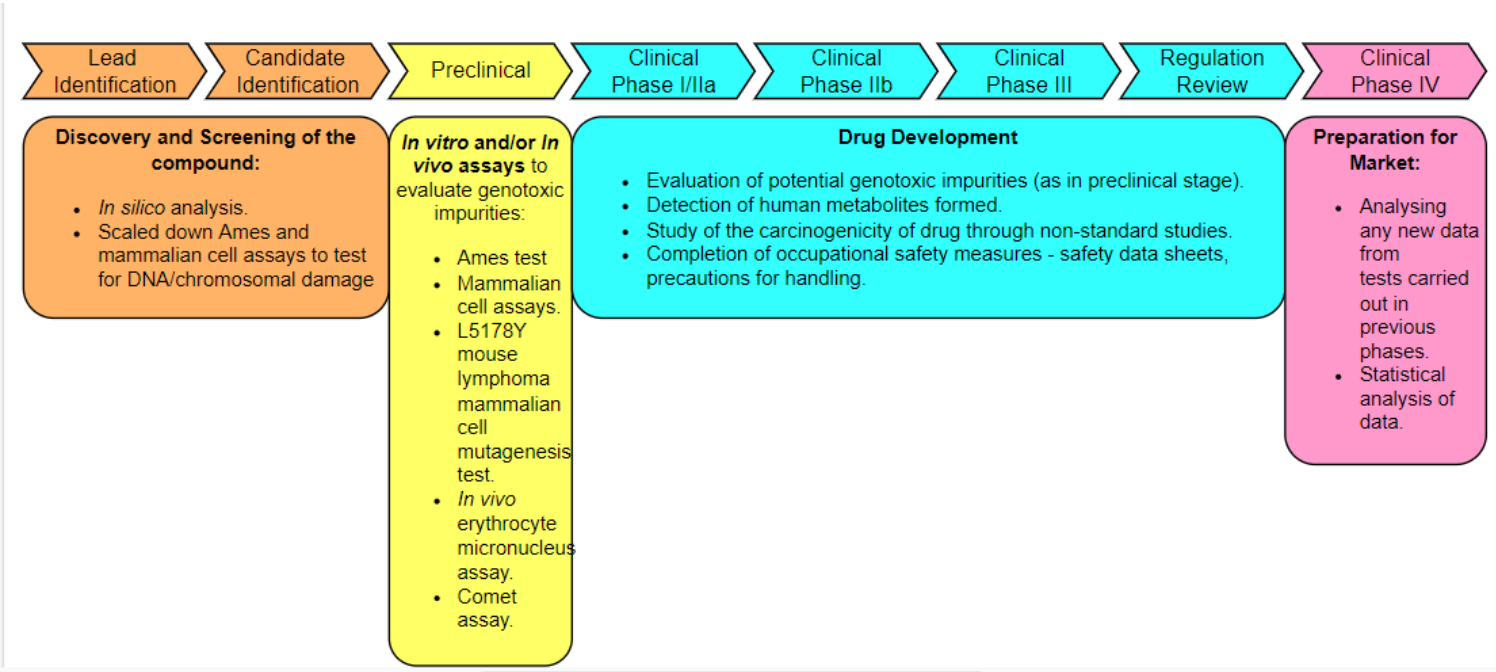



**Figure 1**. Flow diagram of the drug development pipeline from screening of potential compounds and chemicals to progressing the pharmaceuticals to market (Modified from Johnson, 2020).

In the late 1990s, the ICH achieved a consensus for a testing battery for the pharmaceuticals industry. The three tests are firstly a bacterial test for gene mutations in bacteria, an *in vitro* mammalian chromosome aberration test called the L5178Y mouse lymphoma mammalian cell mutagenesis test, and finally an *in vivo* chromosomal damage test in rodent haematopoietic cells called the *in vivo* erythrocyte micronucleus assay (MacGregor et al., 2000).

However, since this initial testing battery, regions of the world have adapted this to create specific test batteries for their region and specific industries. The European Union (EU) test battery is a three-tiered system and the type of tests carried out depend on the nature of the chemical or compound. For industrial products, the first tier requires two *in vitro* tests which are the bacterial gene mutation assay (Ames Test) and a mammalian cell assay. In contrast, for pesticides, cosmetics and food additives, the EU requires three *in vitro* tests which includes the Ames Test, either the chromosome aberration test or the *in vitro* micronucleus assay, and thirdly a mammalian cell gene mutation assay. The EU's requirements are again different for pharmaceuticals. The tests required are an Ames Test, an *in vitro* mammalian cell chromosome aberrations assay or an *in vitro* gene mutation assay in mouse lymphoma cells, and thirdly either an *in vivo* chromosome aberration test or the *in vivo* micronucleus assay (Müller et al., 1999).
Furthermore, recently, changes have needed to be made to this testing regime due to the prohibition of *in vivo* testing in cosmetic products and some pharmaceuticals. Therefore, to reduce animal testing, *in vitro* genotoxicity testing has become the forefront of chemical testing for genetic damage. This being said, more attention has been paid to increase the accuracy and sensitivity of *in vitro* assays to consequently reduce the rate of false positives (Pfuhler et al., 2014). Also, more research is being carried out to develop new *in vitro* assays which aim to follow up positive results from the *in vitro* assays. An example of a follow up test uses 3D reconstructed skin models as well as gaining a deeper understanding of the biochemistry of the chemical/compound itself and the effects it has on metabolism (Allemang et al., 2021).

## 1.2 The Micronucleus (MN) Assay

The MN assay is a globally significant, OECD approved *in vitro* assay used to quantify DNA damage at the chromosomal level, in genotoxicity testing. Therefore, it is one of the first tests that a new compound or chemical is used as a subject for, to test its genotoxic, cytotoxic and carcinogenic potential (OECD, 2016).
MN occur in dividing cells and they are composed of either whole or fragments of chromosomes that have not properly adhered to the spindle machinery on the metaphase plate and they lag behind during anaphase movement. Therefore, they are not incorporated into the daughter nuclei of the dividing cell and thus a separate nuclear envelope forms around them and they become a micronucleus. Due to their nature, MN are a very effective genotoxicity endpoints as their presence is strongly indicative of chromosome damage either of aneugenic (chromosome loss) or clastogenic (chromosome breakage) nature. Common cells to host the *in vitro* MN assay is blood lymphocytes because they divide rapidly, are widely available and they are systemically distributed around the body so they are representative of how the body would respond when exposed to certain chemicals and compounds (Luzhna et al., 2013).
Prior to the 1980s, there was one conventional MN assay, however, in the early 1980s, Fenech and Morley devised the cytokinesis-block micronucleus (CBMN) assay which used cytochalasin B to inhibit cytokinesis in the dividing cells, therefore the cells become binucleated. The mode of action of cytochalasin B is inhibition of the polymerisation of actin (Fenech, 1993). Actin is an essential protein that composes the cytoskeleton and plays a significant role in the production of the cleavage furrow which thus leads to cytokinesis (Subramanian et al., 2013).
MN are formed once a cell has divided once, however their abundance in a cell decreases when the cell undergoes multiple cell cycles. Therefore, it is advantageous to conduct the CBMN assay as the

researcher can distinguish which cells are actively dividing and those that are not by whether the cells are binucleated or not.

On the other hand, there are disadvantages to the CBMN assay. Cytochalasin B does not significantly interfere with nuclear division; however, it may contribute to the formation of MN in binucleated cells as actin also has a key role in the migration of chromosomes in anaphase. Therefore, it is important to use the correct concentration of cytochalasin B to maximise cytokinesis inhibition but - minimise interference with mitosis (Fenech, 1997).

Figure 2 illustrates aneugenic and clastogenic activity in a diploid cell and the resulting MN after cytokinesis and when cytokinesis is blocked (CBMN assay).

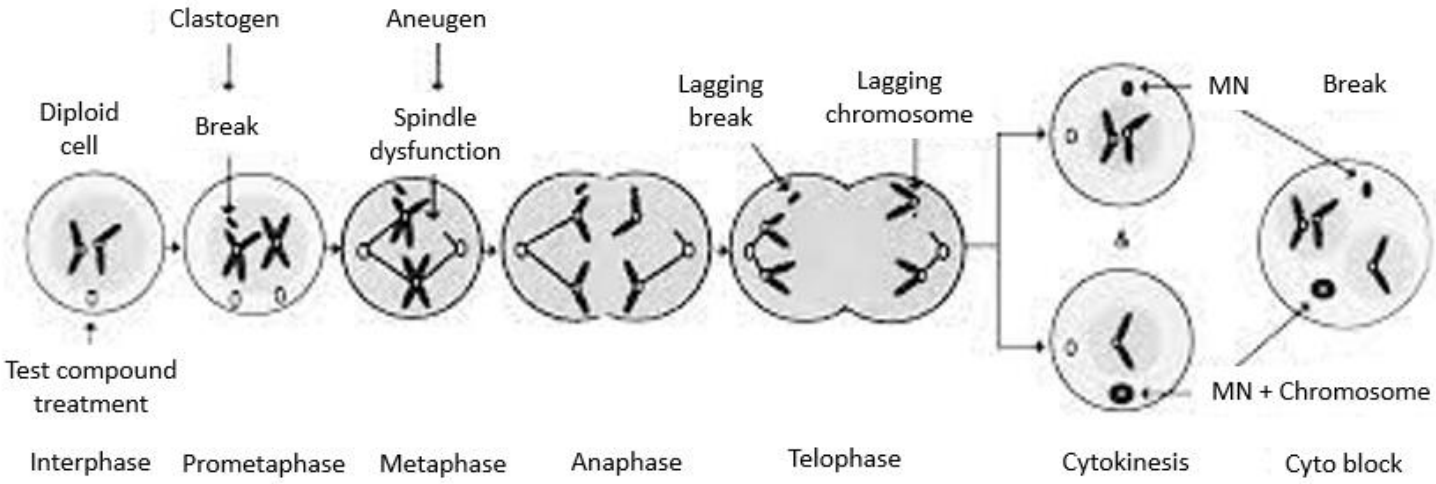



**Figure 2**. Illustrative representation of aneugenic and clastogenic activity the resulting MN formation after both cytokinesis and cytokinesis-block (CBMN assay) (Nath & Krishna, 1998). Screenshot of original figure has been imported.

## 1.3 Endpoint Analysis

As mentioned above, MN are very effective genotoxicity endpoints, however in order to gain a deep understanding of the modes of action of test chemicals, different endpoints can be used. Firstly, markers for the histone protein, H3, can be used to indicate the occurrence of chromosomal damage. Greater levels of H3 correlates to greater levels of chromosomal loss. Importantly, the use of H3 allows information to be gained of alterations taking place such as methylation, acetylation, phosphorylation, and ubiquitination. Therefore, mechanisms of a chemical's toxicity can be determined. For example, H3K9 methylation links to signalling damage to the DNA, and H3K9 acetylation links to chromatin unravelling (Hake & Allis, 2006).

The second endpoint commonly used is p53 which is a nuclear transcription factor linked with stimulating apoptosis (programmed cell death). P53 is also implicated in many cancers, and it is found to be mutated in approximately 50% of malignant cancers. P53 is known to control the arrest of the cell cycle; a cell with damaged DNA will be halted, then once the damaged DNA is repaired, the cell can continue the cell cycle. However, if a cell has significant amounts of damaged DNA, p53 will stimulate the cell to undergo apoptosis so the damaged DNA is not passed onto the daughter cells. Under healthy conditions, p53 is expressed in very low concentrations, therefore cells with DNA damage have an increase in p53 expression compared to a healthy control. This therefore alludes to the fact that p53 is a very reliable endpoint for genotoxic testing (Perri et al., 2016).

The final endpoint for genotoxicity testing is H2AX. H2AX is a histone protein which becomes phosphorylated when double stranded DNA breaks occur from chemical exposure. When H2AX is phosphorylated at the 139$^{th}$ serine residue, γH2AX is produced which can be detected in immunofluorescence-based assays. Research has shown phosphorylation of H2AX is an early response to double stranded breaks, and the γH2AX molecules surround the double stranded break to initiate opening of the chromatin at the damaged site. This therefore allows repair proteins to enter the site and repair the break. Moreover, H2AX is another reliable genotoxicity endpoints as increased levels of γH2AX indicate the presence of DNA damage (Mah et al., 2010).

## 1.4 Biochemistry of Genotoxic and Non-Genotoxic Compounds

The main purpose of the *in vitro* MN assay is to quantifiably test chemicals and compounds for its genotoxic and carcinogenic potential. When developing an assay, it is important that known compounds are used to test the accuracy, precision, reliability and thus validity of the assay. Therefore, the OECD have published a list of compounds and their known genotoxicity potential. This study uses six known compounds to test the validity of the newly developed deep learning neural network.

Firstly, carbendazim is a known aneugen of the benzimidazole family of fungicides. Therefore, the main source of human exposure to carbendazim is through residues on food crops. *In vivo* studies involving rodent exposure to carbendazim showed that mice exposed to carbendazim had an increased incidence of adenoma and carcinoma, particularly in the reproductive organs. Biochemical studies suggest that the biological effect of carbendazim is due to its interaction with cell microtubules which have a key role in intracellular transport and cell division (Davidse, Flach., 1977). In fungal species, the benomyl moiety binds to the protein tubulin, therefore tubulin polymerisation is inhibited, and the microtubules are not formed and thus cannot function correctly. Dysfunctional microtubules in turn cause interference with mitosis as the spindle fibres will no longer be able to form properly. This therefore can lead to the production of micronuclei as the chromosomes can no longer bind to the spindle and get pulled apart during anaphase so whole chromosomes will not be incorporated into the daughter nuclei which explains the aneugenic properties of carbendazim (International Programme on Chemical Safety, 1993).

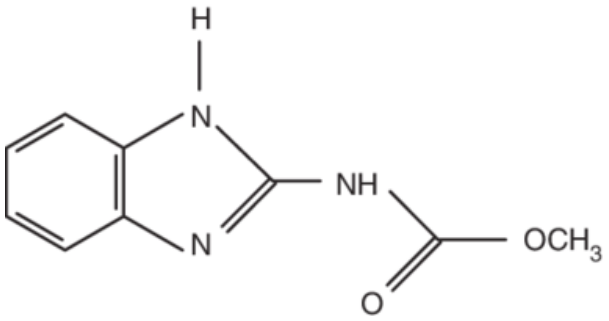



**Figure 3.** Chemical structure of carbendazim (Lu et al., 2004).

The second aneugen used in this study is vinblastine which is a type of vinca alkaloid derived from the Madagascar periwinkle plant (*Catharanthus roseus*). Vinblastine as well as other vinca alkaloids have been successfully used in management of the progression of different cancers including non-Hodgkin's lymphoma and breast cancer. Vinblastine specifically inhibits angiogenesis which is the production of blood vessels towards a tumour. Vinblastine has multiple mechanisms of action to induce genotoxic events. Similar to carbendazim, vinblastine can bind to tubulin hence interfering with spindle production. An additional mode of action for vinblastine is that it induces the production of reactive oxygen and nitrogen species. Studies have shown that vinblastine caused a significant increase in the production of 8-hydroxy-2-deoxy guanosine (8-OHdG) which is indicative of oxidative DNA damage. This also shows vinblastine induces endogenous DNA damage which causes micronucleus production by aneugenic processes (Mhaidat et al., 2016).

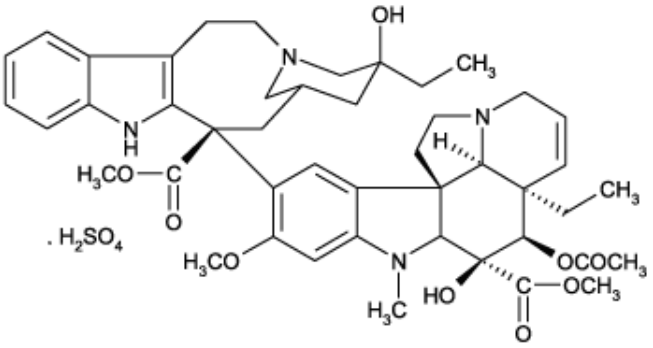



**Figure 4.** Chemical structure of vinblastine sulphate (Enzo Life Sciences, 2021).

Thirdly, MMS is a known clastogen which means it induces structural chromosome breaks. MMS is a DNA alkylating agent; therefore, its mode of action is to bind to and therefore modify DNA nucleotides. MMS specifically modifies guanine to 7-methylguanine and adenine to 3-methyladenine. This means that these nucleotides are no longer able to complementary base pair to cytosine and thymine respectively, therefore mispairing and replication blocks occur hence leading to double stranded DNA breaks. This therefore can contribute to the formation of micronuclei as fragments of the chromosomes will not be incorporated into the daughter nuclei from mitosis, hence form a micronucleus (Lundin et al., 2005).

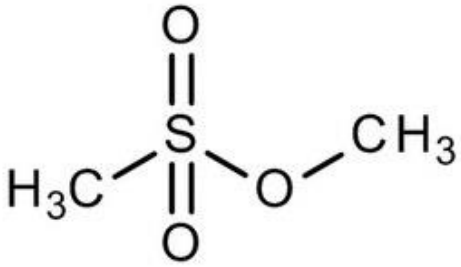



**Figure 5.** Chemical structure of MMS (Merck, 2021).

Fourthly, another known clastogen is benzo(a)pyrene (BaP), a polycyclic aromatic hydrocarbon (PAH) which are formed during the incomplete combustion or pyrolysis of organic materials. Therefore, BaP can be commonly found in air, water, soil and sediments at the source of the incompletely combusting organic material. PAHs such as BaP can be found in high concentrations in tobacco smoke and they can also be found in some foods such as broiled or smoke-cured meat, baked and fried foods that have been processed at high temperatures and vegetables that have been grown in contaminated soils. BaP can also be found in coal tar-based pharmaceuticals which are for example dermatologically applied. BaP-induced genotoxicity has been found to be due to two complementary mechanisms. The first of these is the diolepoxide mechanism in which BaP is converted to BaP-7,8-diol-9,10-epoxide through a series of metabolic transformations using the cytochrome P450 (CYP) enzymes, CYP1A1 and CYP1B1. Diolepoxides are carcinogenic because they react with the purines, deoxyguanosine and deoxyadenosine to produce bulky adducts in both the cis and trans conformations. Therefore, BaP-7,8-diol-9,10-epoxide has the potential to induce the production of 16 different DNA adducts. However, the most abundant BaP adduct found in *in vivo* studies is the $N^2$-deoxyguanosine adduct. This adduct causes G/T (transversion) and G/A (transition) mutations, therefore if these mutations occur in oncogenes such as the *Ras* gene or in tumour-suppressor genes such as the *p53* gene, it can drive the proliferation of cancer cells.

The second mechanism of BaP-induced genotoxicity is through the production of radical cations. One -electron oxidation of BaP by CYP enzymes can cause the production of a radical cation on the $6^{th}$ carbon due to the stereochemical change of the BaP molecule and the ionisation from the

oxidation reaction. Radical cations then can induce the assembly of chemically unstable covalent adducts with guanine and adenine bases. These adducts denature the bases which results in apurinic sites in the DNA which can lead to point and frameshift mutations. Similarly, to the diolepoxide mechanism, if these adducts and mutations occur in oncogenes and tumour suppressor genes a carcinogenic effect can be driven.

Due to these two mechanisms, BaP can cause chromosomal breaks which explains the clastogenic nature of the compound. The chromosomal fragments produced from BaP-induced DNA damage can lead to the production of micronuclei which are scored for in the *in vitro* MN assay (International Agency for Research on Cancer, 2012).

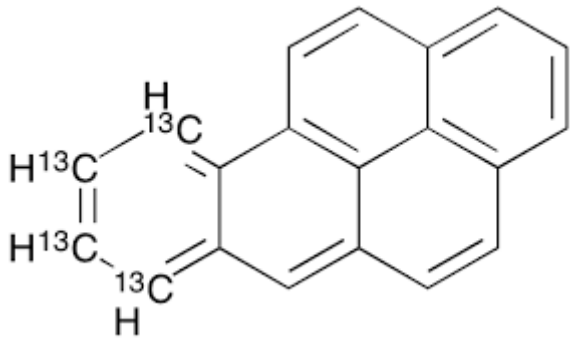



**Figure 6.** Chemical structure of benzo(a)pyrene (Toronto Research Chemicals, 2021).

Finally, two known non-genotoxic agents were used. The first of these is Aroclor which is a polychlorinated biphenyl (PCB). PCBs are commonly used as dielectric fluids in coolants, transformers and capacitors, however due to the abundance of chlorine in its structure, PCBs are acutely toxic to freshwater organisms as well as marine organisms. They are toxic to fish because PCBs are isomers to ligands of steroid nuclear receptors therefore, binding can lead to disruption of the endocrine system which can therefore lead to the onset of yolk sac oedema and haemorrhaging. This can also be called blue sac disease. PCB exposure in fish has also shown to cause abnormalities in the reproductive system such as inhibition of spermatogenesis and reduced egg production as well as other biochemical abnormalities such as hyperglycaemia and the formation of neoplasms. However, the effects of PCB are dependent on the congener that the organism has been exposed to. *In vitro* studies in mammalian cells have shown that Aroclor is not genotoxic in human cells therefore toxicity is unique to marine and freshwater organisms. This finding resulted in the termination of the use of Aroclor in the 1970s (Farrell, 2014).

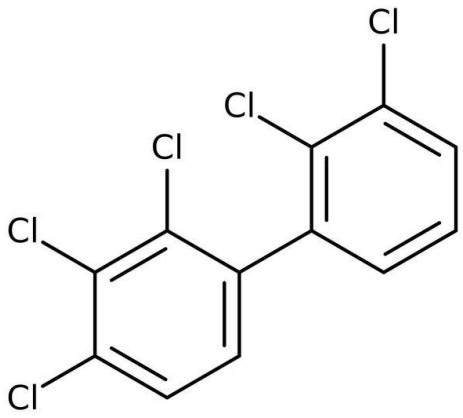


**Figure 7.** Chemical structure of Aroclor (Fisher Scientific, 2021).

The second non-genotoxic agent used in this study was D-mannitol. D-mannitol is an osmotic diuretic that naturally occurs as a sugar alcohol in fruits and vegetables. The main function of D-mannitol is to increase the blood plasma osmolality which increases the flow of water from the tissues into the interstitial fluid and blood plasma. This can therefore alleviate oedema in organs such as the lungs and the brain and it can induce diuresis to prevent the progression of kidney failure. Additional applications of D-mannitol are firstly to promote urinary excretion of toxicants and secondly enhance water flow from the eye to ease pressure in the eye which can help in the treatment of glaucoma. Thirdly, D-mannitol can establish an osmotic gradient in the epithelium of the trachea and bronchi which can push fluid into the extracellular matrix and can ease mucus clearance in cystic fibrosis patients. Finally, D-mannitol can be used as a diagnostic aide by measuring glomerular filtration rate through inducing urinary excretion of toxicants (Cruz et al., 2001).

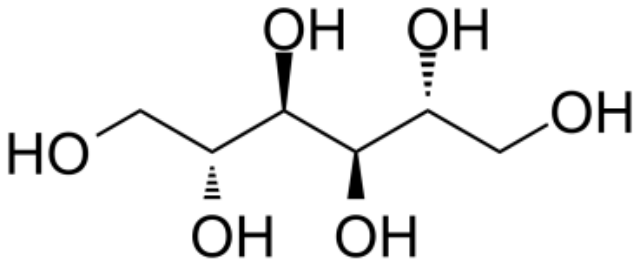



**Figure 8.** Chemical structure of D-mannitol (MedChemExpress, 2021).

Further data analysed in this project was from TK6 cells treated with quinoxaline. Quinoxaline is a heterocyclic compound which is composed of a benzene ring and a pyrazine ring. When the nitrogen atoms of the pyrazine ring are oxidised, quinoxaline 1,4-di-N-oxide (QdNO) are produced which manifest multiple beneficial biological properties including antitumoral, antibacterial, antifungal, antiprotozoal and anti-inflammatory properties, hence there is potential to use them in human medicines. For example, the use of QdNOs has shown promising effects in the treatment of antibiotic-resistant tuberculosis. QdNOs are induced to exhibit these antimicrobial properties in hypoxic conditions. In hypoxic conditions (absence of oxygen), QdNOs receive an electron and hydrogen ion, thus they form a radical which then goes on to induce DNA damage (Cheng et al., 2016).

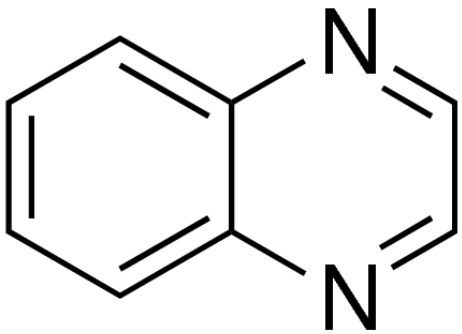



**Figure 9.** Chemical structure of Quinoxaline (Merck, 2021).

## 1.5 Nuclear Stains

When visualising cells using either manual microscopy through light microscopy, or automated microscopy through imaging flow cytometry, it is important to use nuclear stains on the sample of cells in order to aide visualisation of the nuclear DNA material. This is very important for scoring micronuclei in genotoxic tests as they are essential for distinguishing between micronuclei, main nuclei and cellular debris.

The first nuclear stain routinely used is Deep Red Anthraquinone 5 (DRAQ5) which is a fluorescent dye that stains nucleic acids. This stain allows images to be viewed in both the brightfield and nuclear fluorescence channels of a tiff file originally from the IDEAS software. The stain also causes a red fluorescence to be emitted when visualising the cells. DRAQ5 is also optimally excited at 568nm, 633nm and 647nm as shown by the black line on the emission and excitation spectrum below (Biolegend, 2021).

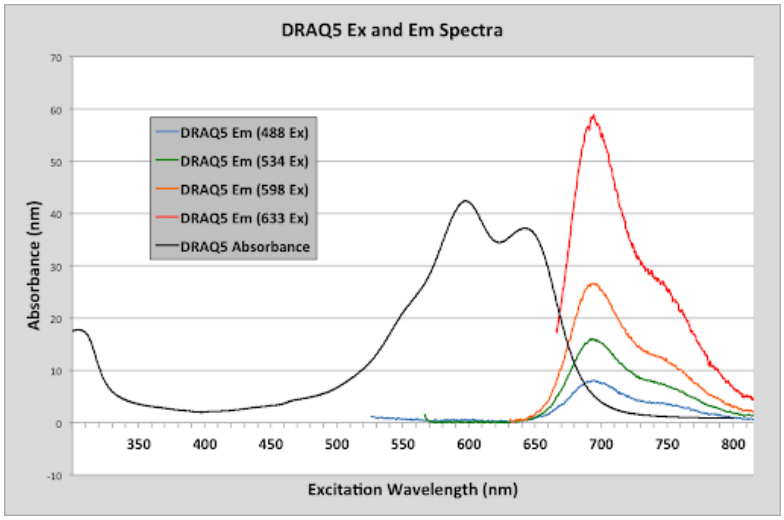



**Figure 10.** Emission and excitation spectra of DRAQ5 (Biostatus, 2021).

The second nuclear stain commonly used in genotoxicity testing is Hoechst 33342 which similarly to DRAQ5 can stain nucleic acids in live or fixed cells. Hoechst 33342 is particularly advantageous for the *in vitro* MN assay because it has a high affinity for double stranded DNA, therefore it can clearly label double stranded DNA and hence the nuclei and micronuclei it is located in. This also avoids staining of RNA molecules so this limits the staining of other cytoplasmic organelles such as ribosomes. When there has been stain-DNA binding, a blue fluorescence is emitted. Hoechst 33342 is optimally excited at 350nm and 461nm (ThermoFisher Scientific, 2021).

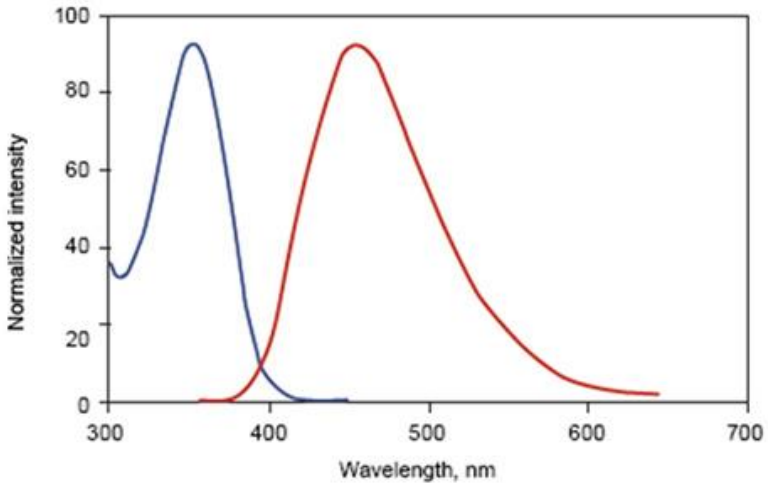


**Figure 11.** Excitation spectrum of Hoechst 33342 (Bio-Rad, 2021).

## 1.6 TK6 Cells

Human lymphoblast, thymidine kinase heterozygote is more commonly known as the TK6 cell line which was originally isolated from the lymphoblastoid cell line called HH4. The TK6 cell line is a frequently used cell line for genetic toxicology testing because they have a relatively rapid mitotic division rate, and they are morphologically larger than primary lymphocytes. It is advantageous that they are larger in size because more laboratory tests can be carried out at a decreased magnification therefore procedures involving TK6 cells are more reproducible, accessible and cost effective. TK6 cells are also cultured in suspension rather than as a monolayer which makes them an easier cell line to culture as they do not have to be treated with trypsin before subculturing. This therefore prolongs the viability of the cells *in vitro.* In addition, TK6 cells do not produce a high concentration of CYP enzymes, therefore TK6 cells tend to be treated with chemicals and compounds that do not require metabolic activation (Rees et al., 2017). If metabolic activation is required for test chemicals, other lymphoblastic cell lines such as AHH-1, which produce lots of CYP1A1, and MCL-5 which express other CYP enzymes such as CYP1A2, CYP2A6, CYP3A4 and CYP2A1 in a plasmid (Aranda et al., 2014).

## 1.7 Progression of Automating the *in vitro* MN Assay

Manual scoring of MN is the gold standard method of determining the dose response of test chemicals and compounds, however, it is a laborious and time-consuming process, as well as being scorer subjective, therefore, to keep up with the demand for genotoxicity testing, automation of MN assay is required. There are two elements of the assay that can be automated: the imaging of the exposed and treated cells, and secondly the scoring of those cells. Table 1 below presents the protocols and equipment that have been developed to automate the *in vitro* MN assay, as well as the advantages and disadvantages of these developments.

For the traditional scoring methods, the treated cells are stained then viewed under light or fluorescent microscopes. In order for cell analysis to be fair and accurate across all the test chemicals, multiple scoring measures have been implemented and approved by regulatory bodies. Firstly, when analysing binucleated cells, at least 1000 cells are required to be scored for the MN count to be valid. For mononucleated cells, this figure increases to 2000. Another implemented measure is that MN are required to be between a third and sixteenth the diameter of the main nucleus, and it should have a circular or ovular shape (Seager et al., 2014).

To increase the throughput of the assay in comparison to manual microscopy in traditional methods, automated microscopy methods have been developed. In automated microscopy, the measures mentioned above are automatically checked for therefore, the process is less laborious and scorer subjective. Metafer™ is an automated microscopy system that has been compared to manual microscopy and flow cytometry procedures as shown in table 1 (Verma et al., 2016). Metafer™ is a semi-automated system where the cells are first stained with a fluorescent dye to increase the ease and speed of scoring MN. The stained cells are then loaded for scanning and images are taken at a 10x objective. The images are then checked at with a 100x objective lens using the graticule in the display view. When this method was compared to the gold standard manual microscopy, the results were parallel, showing that this is a reliable method. Metafer™ also allows results to be stored for

further evaluation as well as for dose response calculations. This means that comparisons can be made between scorers which improves the reliability and validity of the results (Verma et al., 2016). On the other hand, there are disadvantages to this semi-automated method. Firstly, it is difficult to change the settings for checking the validity of the cells of different cell lines. For example, human lymphoblastic TK6 cells are a different size to HepG2 cells so the parameters would need to be altered. It is also difficult for the system to differentiate MN from the parent nuclei when they are overlapping. Finally, due to the use of nuclear fluorescent staining, there is reduced cytoplasmic staining therefore it is difficult to determine the cytotoxic effects of a test chemical, for example it is hard to determine if the cells are apoptotic. Therefore, due to these disadvantages, there is still a need to manually validate the images which significantly increases the time taken to score the images and reduces the automation properties of the system. Changing the parameters for each cell line can cause an underestimation of the MN frequency because unique morphologies for each cell line may not be identified and large MN may be misclassified as parent nuclei.

Following the development of automated microscopy methods, the natural next step was the development of a system to automate the scoring of MN as this was still causing a bottleneck in the use of the MN assay. This led to the introduction of the Microflow® which is a flow cytometric approach and aims to increase the throughput of the assay. The use of nuclear stains such as ethidium monoazide (EMA) allows apoptotic and necrotic bodies to be distinguished from MN which greatly improves the reliability and accuracy of the results and overcomes a significant challenge in manual microscopy methods. Flow cytometry also greatly decreases the scoring time, with 10,000 cells being scored in 1 – 2 minutes which is a very significant advantage to the scoring process and overcomes the bottleneck. This also makes the assay less laborious and scorer subjective (Verma et al., 2016).

On the contrary, unlike the Metafer™ system, the images cannot be stored for further evaluation, therefore, the images cannot be checked and if false positive or false negative results are acquired, they cannot be re-validated. This is a considerable disadvantage to this technique because it decreases the confidence of the results. Another disadvantage is that before MN scoring can begin, the cells must be lysed, therefore it cannot be determined if cells are mono-, bi-, tri or tetra-nucleated. This can subsequently lead to over- or under-estimation of MN frequency which can decrease the reliability and accuracy of the results. Over or under estimation of MN can also occur because when the cells are lysed, excess debris can be incorporated into the images and can be misclassified as MN, or MN may not be correctly identified. Cell lysis also limits the ability to determine the mode of action (MOA) of a test chemical which can decrease the significance of the assay (Fenech, 2000).

The next step in modernising and streamlining the MN assay was therefore to develop a system which does not require cell lysis, and combines automation of imaging and scoring, to form a fully automated process. The result of this was the imaging flow cytometer called Flowsight®, produced by Amnis, EMD Millipore. This machine functions as flow cytometer however, an image of each individual cell is taken, therefore the researcher is able to analyse each cell if extra checking or confirmation is required for example due to the occurrence of false positive or false negative results. The Flowsight® can also produce images with 20x magnification, which is double that that of manual microscopy. Amnis have also developed an ImageStream X Mark II® which can capture images at 40x magnification which makes it more powerful than the Flowsight® and it still has the advantages of the Flowsight®. This increased magnification means the system can be applied to smaller cell lines

such as T and B lymphocytes which would not be able to be visualised on the previous systems (Rodrigues, 2018).

A further advantage to the ImageStream X Mark II® is that scoring does not take place on the microscope, it takes place on a computer programme called IDEAS®. Scoring on a computer programme is less strenuous for the scorer which means scoring is likely to be more consistent and improves the reliability of the results. IDEAS® also has multiple functions which improve the accuracy of scoring. Firstly, the images can be viewed in 12 channels therefore, a wider variety of stains and biomarkers can be used to aid MN classification. Also, IDEAS® includes the use of masks and templates which allow the scorer to define parameters for MN identification so confusing MN events can be checked. Finally, images are taken in three channels, brightfield, darkfield and nuclear fluorescence. The brightfield channel allows the cell to be viewed on a cytoplasmic level therefore, it is easier for the scorer to distinguish the cytotoxicity of the test chemicals, and the MOA of the chemical can be determined as the researcher is able to identify if the cells are mono-, bi-, tri- or tetra- nucleated, or even apoptotic or necrotic.

However, as with the previous systems, there are disadvantages to the Flowsight® and the ImageStream X Mark II®. The cells must still be manually scored which is still laborious for the scorer. Also, these machines are very expensive, therefore it is not accessible for all laboratories, which limits its use for the *in vitro* MN assay so does not necessarily overcome the bottleneck in throughput (Rodrigues, 2018).

| MN Scoring Approach | Scoring Platform | Advantages | Disadvantages |
|---|---|---|---|
| Image Analysis | Manual (light) microscopy | Suitable for dose response and MOA analysis. | Inter-scorer variations can lead to subjective MN scoring. |
| | | Simple, economical, and adaptable. | Slow, tedious, and time-consuming. |
| | | Can be used in the presence and absence of cyto-B. | Not able to carry out multiplex assays. |
| | | Stained slides can be stored for a long time and re-analysed if needed. | Limits the number of cells that can be scored, therefore reduces statistical precision. |
| | | Suitable for analysing bi-, tri- and tetra nucleated cells. | |
| | Metafer™ (fluorescent microscopy) | Semi-automated. | Classification settings for MN must be optimised for different cell lines and test chemicals. |
| | | Allows for higher content input and therefore higher statistical precision. | Unable to stain the cytoplasm so it becomes difficult to detect small MN. |
| | | Suitable for dose response and MOA analysis. | |
| | | Images from the microscope can be stored for re-validation if needed. | |
| Flow Cytometry | Microflow® | Fully automated to score MN. | Requires cell lysis prior to MN scoring. |
| | | Suitable for dose response and MOA analysis. | Misleading MN cannot be re-validated from the same sample. |
| | | Allows for higher content input and therefore higher throughput. | Can over- and underestimate MN to expert analysis is required. |
| | | 10,000 events can be scored in 1-2 minutes. | Reduced MOA analysis with TK6 cells. |
| | | Capable of analysing the cell cycle. | |

**Table 1**. Summary of the advantages and disadvantages of manual, Metafer™ and Microflow® approaches to the *in vitro* MN assay (Verma et al., 2016).

## 1.8 Deep Learning Convolutional Neural Networks

Therefore, to fully utilise the advantages of the imaging flow cytometry approaches, full automation of the scoring process needs to be achieved. Therefore, the next and current step in the progression of the automation of the *in vitro* MN assay is to develop a deep learning convolutional neural network which applies machine learning and artificial intelligence so a network coded in MATLAB and Python programmes, can be trained to identify micronuclei in images taken from the imaging flow cytometry.
Application of machine learning can be achieved by using the guidelines identified that MN need to be between a third and sixteenth of the diameter of the parent nuclei, and that they exist as a circular or ovular shape. These guidelines can be used to train a network to classify MN and furthermore, dose response calculations can be conducted.

The deep learning convolutional neural network mimics the mechanism of neurons in the brain. The neurons link together in order to form connections, and similarly the neural network is composed of multiple layers which communicate with each other by passing on their predictions to the subsequent layer. Therefore, the more layers incorporated into the neural network, the more communications can be made, hence the better the integrity and performance of the network (Emmert-Streib et al., 2020).

In the same way as we learn to identify different objects by being taught what they are, a neural network learns in the same way. Therefore, a neural network is trained on a 'ground truth' set of images. When a network is trained on a greater number of images, the accuracy of its performance will be increased as its been taught on a wider repertoire of images. The ground truth' acts as a bank of images on which the algorithm can make decisions on unseen images. The network is trained on images called the 'training data' and then it is validated and tested using 'validating data' in which the phenotype is known. A confusion matrix is then produced by the algorithm, which shows how accurate the network was, by comparing its classification with the manual classifications. It is essential that the training and validating data are not mixed because the network's reliability will be decreased because the network may have not been validated and checked on new images (Beale et al., 2020).

For the deep learning convolutional neural network to be utilised to the maximum potential, it needs to be trained on a wide variety of cellular phenotypes so that as many different classifications of images can be correctly recognized by the network. These cellular phenotypes the neural network is trained on are mononucleates with and without MN, binucleates with and without MN, trinucleates with or without MN and tetra nucleates with and without MN. Therefore, images of cells with these phenotypes are incorporated with significant frequency in the ground truth set of images. This allows a dose response to be calculated. When the deep learning convolutional neural network has been trained to the optimum level and it is shown to produce results parallel to those of manual scoring, it is predicted that the algorithm can be used across multiple laboratories to score cells automatically, without the need for manual scoring. There is also potential for the network to be applied to multiplex labels such as H2ax, p53 and H3 to further increase the likelihood of classifying MN events. Therefore, this will significantly increase the throughput, the accuracy and reliability of the assay.
However, the main limitation of this approach is that the algorithm is trained on images that have been scored manually and is therefore subjected to interscorer subjectivity. Therefore, the accuracy

of the network is limited to the accuracy of human scoring (Beale et al., 2020). MATLAB is a computer tool that can be used for running code in order to produce algorithms for artificial intelligence, machine learning and deep learning purposes. MATLAB does not require the use of a coding language as all code is written in English. This therefore makes MATLAB more accessible for multiple users whatever their previous experience of coding is.

MATLAB comes with different toolboxes, including both machine learning and deep learning toolboxes, which aide the user. The deep learning toolbox has a variety of sections that can help the user with many aspects of deep learning. The one of focus for this project is training a deep learning network to classify new images. To achieve this, the deep learning network uses a pretrained network. This pretrained network is trained on millions of images which can be classified into categories. Therefore, the pretrained network holds huge amounts of data with rich representations for each category of images. Each of these images act as an input then the network outputs a label for the image as well as probabilities for each category. A commonly used technique for deep learning is transfer learning, where the pretrained network is used as a starting point to learn new tasks, then fine-tuning is carried out of the deep learning network to increase its accuracy and specificity. This saves the users time and convenience as they do not have to create a new algorithm, they can just update the pre-existing network (Beale et al., 2020).

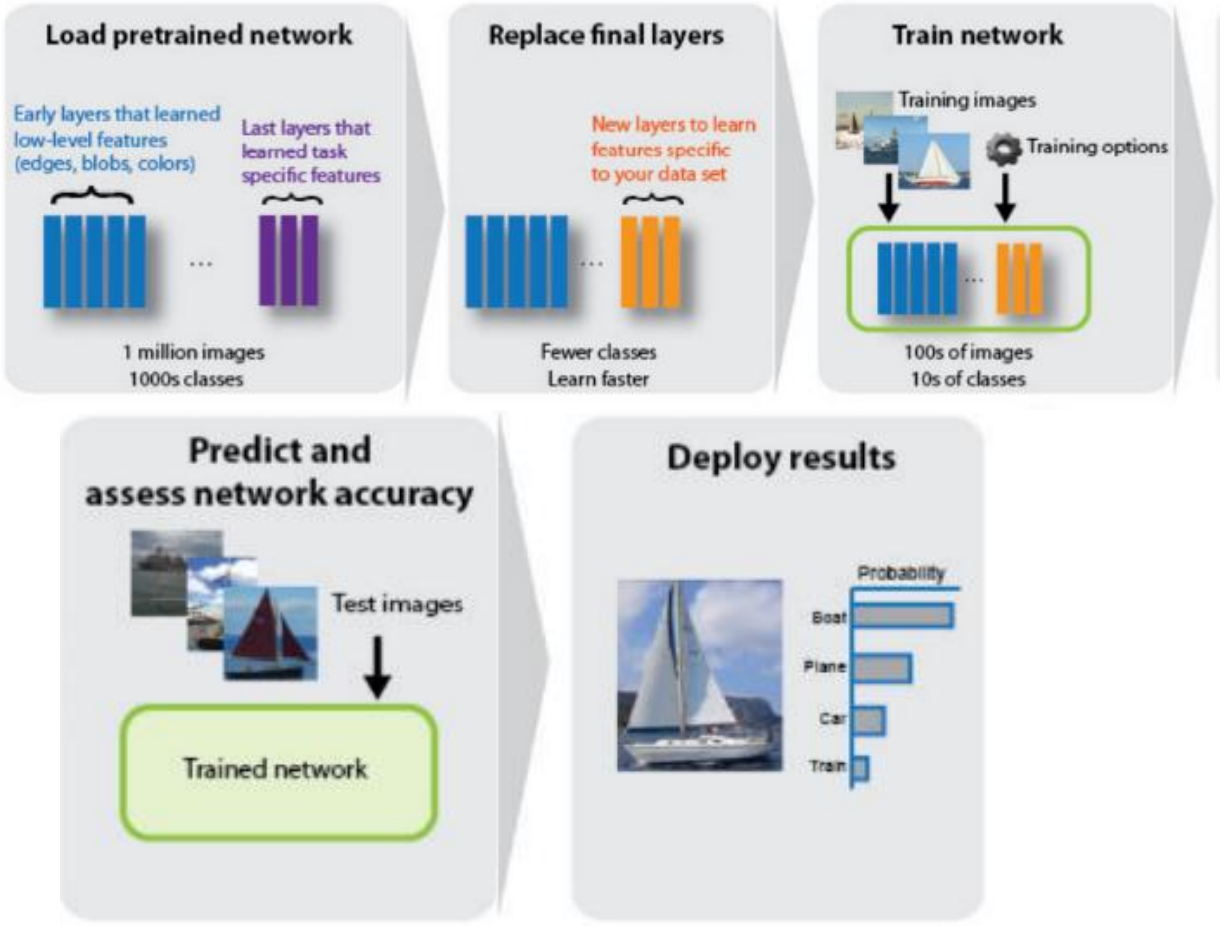



**Figure 12**. Flow diagram to illustrate transfer learning using a pretrained network (Beale et al., 2020).

The data that is loaded into the deep learning network must be split into training and validation data sets. A common adopted method undertaken to ensure training and validation data do not overlap is to use a 3:1 ratio of training images to validation images. This ratio provides a good balance for having enough images for each function. Figure 13 shows the code that is used in MATLAB programming to load a pretrained network which is then applied to the validation data set to improve the specificity and thus performance of the neural network (Beale et al., 2020).

```
% load previously trained network

load('previously trained network')

%% get tester imagesfiles
imageFolder_Validate ='Validation data'

% set up datastore
% 2 channel
imdsValidation = imageDatastore(imageFolder_Validate, 'LabelSource', 'foldernames',
'IncludeSubfolders',true,'FileExtensions','.tif','ReadFcn',@two_channel_tiff_reader
);

% print out table
tbl_validation = countEachLabel(imdsValidation)

% set image size based on trained network
inputSize = trainedNet.Layers(1).InputSize;
```

**Figure 13**. MATLAB programming code used to load a pretrained network which is subsequently applied to validation data sets.

MATLAB programming requires the use of special characters to produce a functioning script. Firstly, the percent sign (%) is used to add comments into the code as this text is non-executable. In figure 12, the % has been used to annotate the code to inform the user that a previously trained network must be loaded. Additionally, the double percent sign (%%) is used to begin a new section of the code, so any text after %% denotes the section title. The equal sign (=) is used to assign data to a named variable. In figure 12, the folder of validation data that is uploaded to the script will be assigned to the variable of 'imageFolder_Validate'. The at symbol (@) is used to indicate a class folder name. In this deep learning network, a class has been scripted called 'two channel tiff reader' which allows the network to read and therefore be trained on the two channel tiff files that are inputted from the imaging flow cytometer. Then in figure 12, a line of code is used to assign the variable 'imdsValidation' to a combination of data files, one of which is the class folder 'two-channel tiff reader'. Additional characters commonly used in MATLAB programming are a semicolon to signify the end of a row, a colon to separate elements and commands of an array and parentheses are used to enclose the data to be used in variable assignment or functions (MATLAB Operators and Special Characters, 2022).

The deep learning network is composed of layers and the first layer is known as the image input layer. This first layer denotes the properties of the images being processed. Once the images have been loaded, they go through each layer and connections are made between each layer, similar to the mode of action of neurones in the central nervous system. The following layers that make up the largest proportion of the convolutional deep learning network are pooling, rectified linear units and repeating convolutional layers. The function of these layers is to build greater connections between

the layers to aide network training, and to confirm the weighting of filters. For example, rarer MN phenotypes such as binucleates, trinucleates and tetranucleates with MN can be given a greater weighting to accommodate for the lower number of images in the training data set with these phenotypes.

Another function of MATLAB is normalising the sizes of the images in the data set. Data augmentation allows all the training images to be equally resized. Additionally, training images can be randomly flipped along the vertical axis and randomly translated up 30 pixels. This process helps to prevent the network from becoming over-trained. Over-training should be avoided because the network may start to score images incorrectly due to small differences. For example, if an image contains a MN that is micrometers different in size, the network may not correctly identify the MN. Figure 14 includes the code used in MATLAB to carry out image normalisation.

```
%% Augment the validation set if needed
          pixelRange = [-5 5];
imageAugmenter = imageDataAugmenter( ...
       'RandRotation',[0 360], ...
       'RandXReflection',true, ...
       'RandYReflection',true, ...
   'RandXTranslation',pixelRange, ...
    'RandYTranslation',pixelRange);
```

**Figure 14**. MATLAB programming code used to normalise the sizes of images in training and validation data sets.

The validation images can also be resized automatically by specifying this line of code:

```
augimdsValidation = augmentedImageDatastore(inputSize,imdsValidation);
```

**Figure 15**. MATLAB programming code used to automatically resize images in the training and validation data sets.

It is essential to set parameters for the training of a deep learning algorithm to improve the accuracy of the network. When deep learning network training begins, the error rate is high, and the accuracy is low because the network has been trained on very few images, so it has less opportunity to learn and make a prediction on new images. However, as training progresses, the accuracy increases and subsequently the error rate decreases. Low accuracy may either be due to the training data or the network itself. If the issue is with the training data, this may be because the images are of low quality, therefore the network can not be trained properly. The user can tell if the low accuracy is due to the training data through analysing the confusion matrix output. The confusion matrix displays the performance of the algorithm by calculating the accuracy for each phenotype being classified, therefore it clearly shows if there is phenotype category that requires further optimisation. On the other hand, low accuracy may be occurring due to an issue with the network itself. This may be caused by too many pooling layers in the main body of the deep learning network, which can in turn lead to over-training of the network. The first parameter to avoid low accuracy is the number of epochs. An epoch is a full training cycle of the entire data set of images. Again, to avoid over-training, the correct balance of epochs needs to be found, so that a high enough accuracy can be obtained, without over-training and increasing the error rate. The second parameter is the batch size. This is defined as the number of samples from the data set that will be inputted into the network at a time. In other words, epochs are split into batches in order for the network to go

through the data set in an organised manner. For example, if a data set contains 2000 images and the batch size is set to 100 images, the network will have to go through 20 iterations to complete an epoch and train the network. A smaller batch size is advantageous because less computational power is required, and the speed of training is faster. However, smaller batch sizes can cause lower accuracy because each training sample has less images therefore there is less opportunity for the network to correctly classify images. Overall, a balance needs to be found between computational power, time and accuracy (Beale et al., 2020).
In the context of the *in vitro* MN assay, including the CBMN assay, deep learning convolutional neural networks have recently been applied to imaging flow cytometry data using 'Amnis Artificial Intelligence' software in order to identify micro-nucleated binucleated (MNBN) cells in a 3D reconstructed skin MN assay (Allemang et al., 2021). When the MNBN cells imaged on the imaging flow cytometer were visually scored, their frequency was parallel with manual scoring which gave strong evidence that imaging flow cytometry and Amnis AI could be combined to successfully improve the throughput and robustness of the 3D reconstructed skin MN assay (Allemang et al., 2021).

An additional progression in the automation of the *in vitro* MN assay which has great clinical significance is the creation of an open framework that allows multiple laboratories to access the same deep learning network, switch the parameters for the requirements of the specific laboratory such as the calibration of the imaging flow cytometry and the nuclear stain used. This will improve the accessibility, reproducibility and accuracy of the deep learning algorithm. A study carried out in 2021 assessed the ability of deep learning algorithms to be used for MN scoring across different laboratories (Wills et al., 2021).

This masters by research project aimed to continue the optimisation of the deep learning convolutional neural network and apply it to imaging flow cytometry data from multiple laboratories including Newcastle, Swansea and Aberystwyth universities. The initial optimisation of the neural network was carried out using data from Newcastle university which was composed of TK6 cells cultured *in vitro* and treated with the six different chemicals addressed earlier.

# 2. Materials and Methods

## 2.1 Image Collection

Image data from the Amnis ImageStream X Mark II® imaging flow cytometers were collected from three different laboratories: Faculty of Medical Sciences of Newcastle University, Institute of Biological, Environmental and Rural Sciences (IBERS) of Aberystwyth University and Swansea University Medical School.

## 2.2 Chemicals

The following six compounds were purchased from Sigma Aldrich (Merck). Carbendazim (CAS registry number 10605-21-7) (working concentrations: 0.0, 0.4, 0.8, 1.0 and 1.6 µg/mL), vinblastine (CAS no. 143-67-9) (working concentrations: 0.0, 0.0002, 0.0004, 0.0006, 0.0008, 0.001 and 0.002 µg/mL), methyl methanosulphate (MMS) (CAS no. 66-27-3) (working concentrations: 0.0, 1.25, 2.5 and 5.0 µg/mL), benzo(a)pyrene (CAS no. 50-32-8) (working concentrations: 0.0, 2.0, 2.5 and 3.0 µg/mL), Aroclor (working concentrations: 0.0, 0.0004, 0.0006, 0.0008, 0.001, 0.002 and 0.004 µg/mL), D-mannitol (CAS no. 69-65-8) (working concentrations: 0.0, 500, 1000, 1500 and 2000 µg/mL) and quinoxaline (CAS no. 49-48-9)(working concentrations: 0.0, 0.5, 2.0 and 4.0 µg/mL), purchased from Fluorochem.

## 2.3 Cell Culture

TK6 cells purchased from the Health Protection Agency Culture Collections were cultured in RPMI 1640 media (#A1049101, ThermoFisher), supplemented with 100 U/mL penicillin, 100 µg/mL streptomycin and 10% heat inactivated horse serum (#2605008, ThermoFisher). The cells were seeded at a density of 2 x $10^5$ cells/mL in 25$cm^3$ culture flasks. The cultured cells were then incubated at 37 °C and 5% $CO_2$ for approximately 1.5 cell cycles (24-30 hours). Samples of cells were exposed and treated with the chemicals and half of the samples were also treated with 3 µg/mL cytochalasin B (#C6762, Sigma Aldrich). Following cell culture and treatment, the cells were pelleted and washed with phosphate buffered saline (PBS), then resuspended in FACS lysing solution (#349202, BD) for 12 minutes so the cells became permeabilised.

## 2.4 Nuclear Staining

The permeabilised cells were then incubated with 0.05 mM DRAQ5 (#564902) in PBS at room temperature for 30 minutes. The cells were then pelleted, resuspended and diluted in PBS to an optimal concentration for imaging flow cytometry which is approximately 100 µL volumes with 1 x $10^7$ cells/mL.

## 2.5 Imaging Flow Cytometry

Brightfield and nuclear fluorescence images were collected using the ImageStream X Mark II® with an x40 objective lens. The DRAQ5 stained cells were excited using 488nm laser and the brightfield images were collected in channel 1 and nuclear fluorescence images in channel 11.

These processes of cell treatment, culture and image capture were performed on a previous project so that the current study could only focus on the optimisation of the deep learning algorithm.

## 2.6 Compensated Image File (.cif) Formation in IDEAS

Once the images were taken, the raw image files (.rif) were transformed into .cif files then data analysis files (.daf) in IDEAS® 6.2 software. A master template was created in IDEAS from masks and

features in order to produce images of individual cells to make it easier to recognise the different cell phenotypes which in turn aides the creation of the ground truth data set.

### 2.7 Cif to Tif Extraction

The individual images of cell populations exported as .cif files were then converted to grayscale, three 8-bit channel tif files (composed of brightfield, nuclear fluorescence and darkfield channels). The three individual channels were (max-min) renormalised and cropped to 64x64 pixel squares which was the correct format for the tif files to be loaded into the DeepFlow convolutional neural network for incorporation into the ground truth data set and deep learning. The MATLAB script used for cif to tif extraction was downloaded from Wills et al., 2021.

### 2.8 Automated Scoring and Image Classification by the Deep Learning Network

Automated scoring and classification of the tif files was performed using the trained DeepFlow neural network. The deep neural network was previously trained on a ground truth image set representing a wide range of cell phenotypes that arose from TK6 cells treated with carbendazim and MMS. The ground truth was created using imaging flow cytometry images from Cardiff and Cambridge universities (Wills et al, 2021). The methods of cell culture in Cardiff and Cambridge universities were identical to the methods described above, however Cambridge used Hoechst 33342 nuclear staining, thus the stained cells were excited during imaging flow cytometry with 405nm lasers and brightfield images were collected in channel 4 and nuclear fluorescence images in channel 1.

To follow the guidelines for the *in vitro* MN assay, positive scores for MN were given to cells with fluorescently labelled MN that were circular or ovular in shape and were $1/3^{rd}$ – $1/16^{th}$ the size of the main nucleus. The experiment was also conducted in triplicate. When calculating the dose response for each compound, two thousand cells were scored.

### 2.9 Percentage Confidence Assay of the Deep Learning Neural Network

When the tif files were loaded into the trained deep learning neural network, a sample of 36 cell images were produced which the algorithm had scored as binucleated and BNMN and those images were then scored manually. The confidence level that the algorithm outputted was recorded and it was determined at what confidence the algorithm had given a classification when the algorithm's score matched the manual score using a light microscope.

### 2.10 Dose Response Calculation of Unanalyzed Image Data

Tif file images from the three locations were inputted into the trained DeepFlow convolutional neural network and the dose response for each compound was calculated using the following formula: (number of binucleated or mononucleated cells with MN/number of binucleated or mononucleated cells)/100 to determine the percentage frequency of MN in a sample.

### 2.11 Statistical Significance of MN Dose Responses

Assessment of the statistical significance of the MN dose response was conducted using the workflow detailed in (Johnson et al., 2014). The response data was log transformed and it was determined if the data was firstly normally distributed and homogeneously varied using the Shapiro-Wilks and Bartletts tests, respectively. If the data was N=3 and passed these tests ($p>0.05$), the one-sided Dunnett's test was run to determine the statistical significance of the MN dose response compared to the control (significant if $p<0.05$). If the data was N=3 and failed the Shapiro-Wilks and or the Bartlett's test, the one-sided Dunn's test was run which is non-parametric (Johnson et al., 2014).

When the data was only N=2, a different statistical test was used called the Fisher's Exact which compares the dose responses of each increase in dose, with the control group in a two-way contingency table.

### 2.12 Benchmark Dose Analysis

To compare the MN dose responses of the compounds, benchmark dose (BMD) analysis was used. This was carried out using the online PROAST software and the data was analysed using two models called the Exponential and Hill models. The data was analysed using dose and MN frequency as the variables, then the chemical was used as a covariate parameter. The benchmark response (BMR) was set to 50% which represents a 50% increase in dose response when compared to the control. BMD here is expressed as a range using the lower (BMDL) and upper (BMDU) confidence limits. Determining the BMDL is clinically significant as it can be used to estimate an oral or dermal exposure level of the tested compound. BMDL is interpreted as a dose in which the genotoxic or cytotoxic effect is lower than the BMR hence considered lower risk in risk assessments. An additional criterion that can be used is the Akaike Information Criterion (AIC). The AIC is a method to determine if the models are a good fit to the data. AIC can also be used as an additional parameter to check if there is statistical evidence of a dose-related trend. For a model to show statistical evidence for a dose-related trend, the AIC should be lower than the null AIC – 2.

# 3. Results

## 3.1 Selection and Justification of the Percentage Confidence Cut-off Rule

After the raw image files taken on the imaging flow cytometer from Newcastle University were processed into cif then tif files, the tif files were loaded into the pre-trained deep learning neural network to be automatically scored by the algorithm in MATLAB programming. The cell phenotypes that the network is trained to classify are mononucleates, binucleates, trinucleates and quadranucleates with and without MN events, as well as the additional 'other or unscorable' classification. When using the neural network on unseen images, it was essential that the algorithm was scoring the images correctly, even if that resulted in images being scored as other or unscorable as opposed to a specific phenotype. Therefore, it was important to investigate how accurate the algorithm was at scoring the phenotypes included in the dose response calculations for a test chemical; mononucleates and binucleates with and without MN events. The network outputted a sample of 36 images that it scored as either mononucleates or binucleates with or without MN, and it also outputted a percentage confidence level that it had in the classification. For this project, spreadsheets were produced to present when the algorithm was correct in its classification of binucleates and mononucleates, with and without MN, and at what percentage confidence level range the algorithm consistently correctly scored the images (Please see appendices 1 and 2). This study shows that at 70% confidence and above, the algorithm correctly classified the images. One exception was found in the sample of binucleates with MN, and seven exceptions were found in the sample of mononucleates with MN, however when viewing these images manually, it was found the images contained shadows which made the classification incorrect. However, the parameter of 70% confidence and above was well represented in the samples, therefore when dose response calculations were conducted later, only images that the algorithm had scored with 70% confidence and above were incorporated.

To increase the transparency of these results shown in appendices 1 and 2, and thus ensure the results are reliable, reproducible and valid, a JPEG of each of the samples of 36 images for each dosage and repetition of each chemical were added to the spreadsheets. Therefore, if researchers would like to see how the percentage confidence cut-off was concluded to be 70%, they could check the images (please see appendices 3 – 8). A number of the JPEGs are included and presented in figures 16 and 17. Figure 16 consists of images of TK6 cells of the cyto-b and non cyto-b assays, treated with 1.6 µg/mL carbendazim that have been correctly and incorrectly scored by the neural network as binucleates and mononucleates with MN respectively. This figure demonstrates that the neural network correctly classified the cells with a confidence level of 70% and above. Figure 17 gives an example of when the neural network incorrectly scored an image at the 70% confidence point.

| | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 2000 | | | N | N | N | 70.20% | 81.50% | 90.20% | | N | N | 50.20% | 60.20% | 73.70% | 80.60% | 92.80% | | | | | | | | |
| 25 | 0 | | N | N | N | 61.80% | 70.50% | N | 90.40% | | | | | | | | | | | | | | | | |
| 26 | 1.25 | | 35.70% | N | N | 64.30% | 70.20% | 80.90% | 91.80% | | | | | | | | | | | | | | | | |
| 27 | 2.5 | | 38.20% | N | 50.50% | 60.60% | 74.60% | 81.10% | 92.60% | | N | | | | | | | | | | | | | | |
| 28 | 5 | | N | N | N | 60.20% | 70.20% | 80% | 92.30% | | N | | | | | | | | | | | | | | |
| 29 | 0 | | N | N | 50.30% | 60.40% | 71.70% | 80.10% | 96.20% | | | | | | | | | | | | 50.50% | N | 70.40% | 86.60% | 91.60% |
| 30 | 0.0002 | | | N | N | 60.40% | 71.20% | 84.10% | 91.20% | | | | | | | | | | | | 50% | 61.50% | N | 80.50% | 91% |
| 31 | 0.0004 | N | 39.10% | 46.20% | 53.50% | 61.30% | 70.30% | N | 91.80% | | | | | | | | | | | | | 61.30% | 70.30% | N | 91.80% |
| 32 | 0.0006 | | | N | 51.50% | 60.20% | 70.10% | 81.50% | 90% | | | | | | | | | | | | 53.50% | N | N | 80.50% | 90% |
| 33 | 0.0008 | | | N | N | 60.10% | 70% | 80.70% | 94.80% | | | | | | | | | | | | | 62.30% | 70.30% | 81.60% | 91.60% |
| 34 | 0.001 | | | 44.50% | N | N | 71.20% | 80.50% | 90.30% | | N | | | | | | | | | | 50.20% | N | 73.60% | 85.60% | 90.50% |
| 35 | 0.002 | | | 40.50% | N | N | 70.60% | 82.80% | 94.60% | | N | | | | | | | | | | 50.80% | N | 70.80% | 82.90% | 90.90% |
| 36 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 37 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 38 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 39 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 40 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 41 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 42 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 43 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 44 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 45 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 46 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 47 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 48 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 49 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 50 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 51 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 52 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 53 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 54 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 55 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 56 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 57 | | | | | | | | | | | | | | | | | | | | | | | | | |
| 58 | | | | | | | | | | | | | | | | | | | | | | | | | |

**Rachel Barnes:**

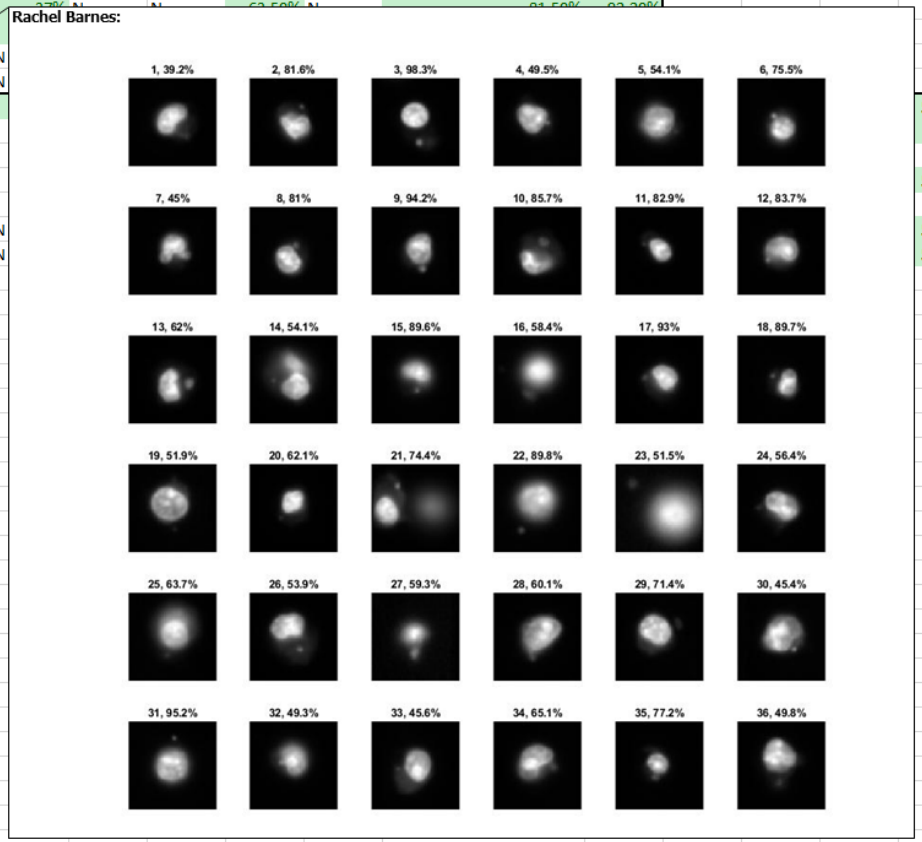


**Figure 16. Images of TK6 cells treated with 1.6 µg/mL Carbendazim from the cyto-b (A-C) and non-cyto b (D-F) assays.**

(A) TK6 cell that the neural network incorrectly scored as a binucleate with MN with a confidence level of 57.5%. After manual scoring, this cell was classified as a mononucleate with MN.
(B) TK6 cell that the neural network correctly scored as a binucleate with MN with a confidence level of 71.7%.
(C) TK6 cell that the neural network correctly scored as a binucleate with MN with a confidence level of 92.2%.
(D) TK6 cell that the neural network incorrectly scored as a mononucleate with MN with a confidence level of 54.8%. After manual scoring, this cell was classified as a mononucleate with 2 micronuclei.
(E) TK6 cell that the neural network correctly scored as a mononucleate with MN with a confidence level of 76.7%.
(F) TK6 cell that the neural network correctly scored as a mononucleate with MN, with a confidence level of 95.7%.

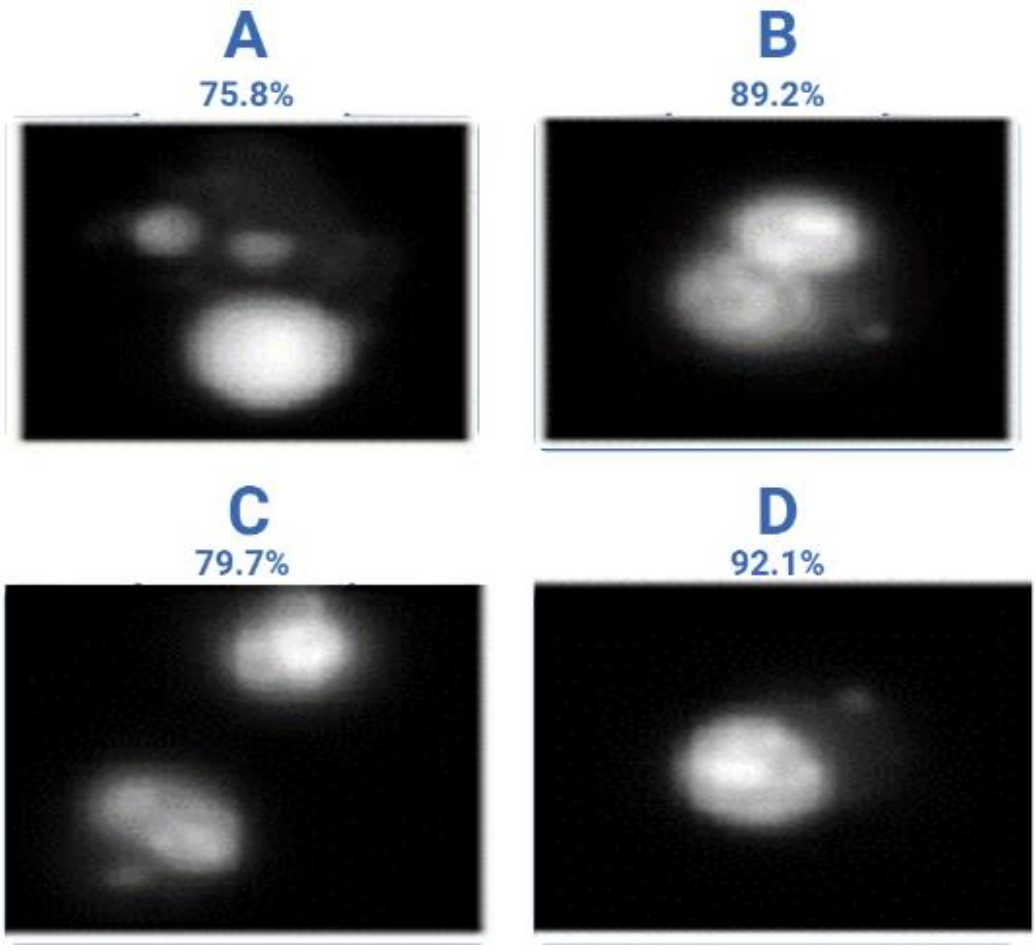



**Figure 17. Images that the neural network incorrectly scored with a confidence level above 70% compared to images scored correctly.**
(A) TK6 cell of the cyto-b assay treated with 2.5 µg/mL benzo(a)pyrene, incorrectly scored as a binucleate with MN at a confidence level of 75.8%. After manual scoring this cell was classified as a mononucleate with 2 micronuclei.
(B) TK6 cell of the cyto-b assay treated with 2.5 µg/mL benzo(a)pyrene, correctly scored as a binucleate with MN at a confidence level of 89.2%.
(C) TK6 cell of the non cyto-b assay treated with 1000 µg/mL D-mannitol, incorrectly scored as a mononucleate with MN at a confidence level of 79.7%. After manual scoring this cell was classified as a binucleate with MN.
(D) TK6 cell of the non cyto-b assay treated with 1000 µg/mL D-mannitol, correctly scored as a mononucleate with MN at a confidence level of 92.1%.

### 3.2 Dose Response Data from collaboration with Newcastle University

#### 3.2.1 Demonstration that the neural network produces valid dose response data.

Since determining the accuracy of the pretrained deep learning neural network, the algorithm could be used to calculate the dose response of aneugenic, clastogenic and non-genotoxic chemicals from images collected from multiple laboratories.

Firstly, Fig. 18 demonstrate the aneugenic effect of carbendazim. he control has a mean MN frequency of 0.351% for binucleates with MN and 0.499% for mononucleates. Then as the dose increases to 1.6 µg/mL, the mean MN frequency increases for binucleates to 1.198% and for mononucleates the MN frequency increases to 1.784%. Since the collected data was N=2, the non-parametric Fisher's Exact test could be applied for statistical analysis. The increase in MN frequency from 0.8 – 1.6 µg/mL was found to be statistically significant in both the binucleate and mononucleate experiments ($p<0.0001$).

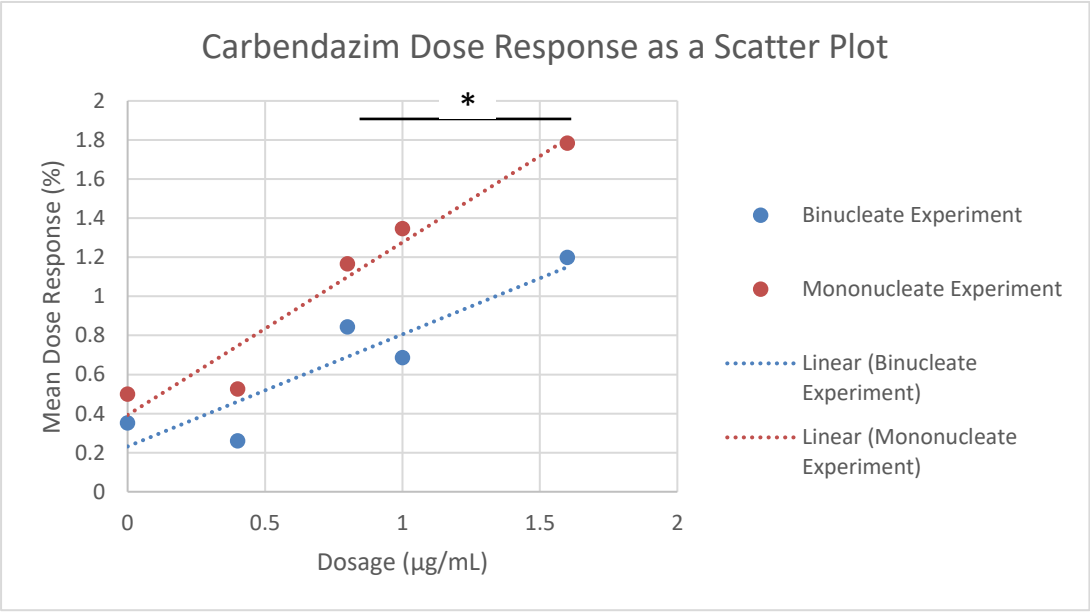



**Figure 18.** Comparison of mean MN dose response of the cyto-b and non-cyto b MN assay of carbendazim treated TK6 cells. N=2,. * $p<0.05$ using Fisher's Exact.

Secondly, Fig.19 shows the aneugenic effect of vinblastine. The control has a mean MN frequency of 0.309% for binucleates with MN and 0.602% for mononucleates. Then as the dose increases to 0.002 µg/mL, the mean MN frequency increases a small amount for binucleates to 0.361% and for mononucleates the MN frequency increases more greatly to 1.134%. The Shapiro-Wilks test showed that the mononucleates MN dose response data was normally distributed, however the Bartlett's test showed that the variants through the dose range were not homogeneous, therefore a non-parametric one-sided Dunn's test was applied since N=3. The increase in MN frequency was not statistically significant. On the other hand, for the cyto-b experiment, the data (N=3) passed both the Shapiro-Wilks and the Bartlett's tests, so the one-sided Dunnett's test was used. The binucleated MN dose response data was not found to be statistically significant ($p>0.05$).

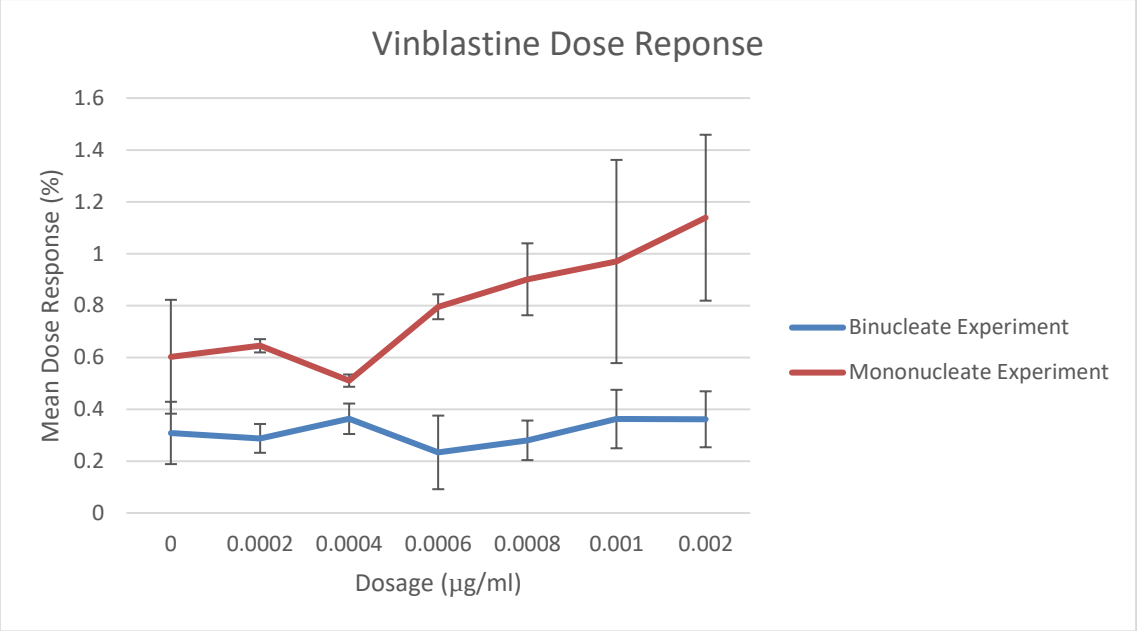



**Figure 19.** Comparison of MN dose response of the cyto-b and non-cyto b MN assay of vinblastine treated TK6 cells. N=3, mean = +/- standard deviation.

Fig. 20 validates the clastogenic effect of MMS. The graph shows a bell-shaped curve where the MN frequency increases then decreases again. The control has a mean MN frequency of 0.337% for binucleates with MN and 0.776% for mononucleates. Then as the dose increases in the cyto-b experiment to 1.25 µg/mL, the mean MN frequency increases to 1.198% and as the dose increases again to 5.0 µg/mL, the mean MN frequency decreases again to 0.188%. On the other hand, for mononucleate experiment, as the dose increases to 2.5 µg/mL the MN frequency increases to 1.221%, then when the dose increases further to 5.0 µg/mL, the mean MN frequency decreases to 1.093%. Both the binucleate and mononucleate experiment data was N=2, therefore, the Fisher's Exact test was used on the data. The increase in MN frequency from 1.25 – 5.0 µg/mL in the mononucleate experiment was found to be statistically significant ($p<0.05$). However, the dose response curve from the binucleate experiment was not found to be statistically significant ($p>0.05$).

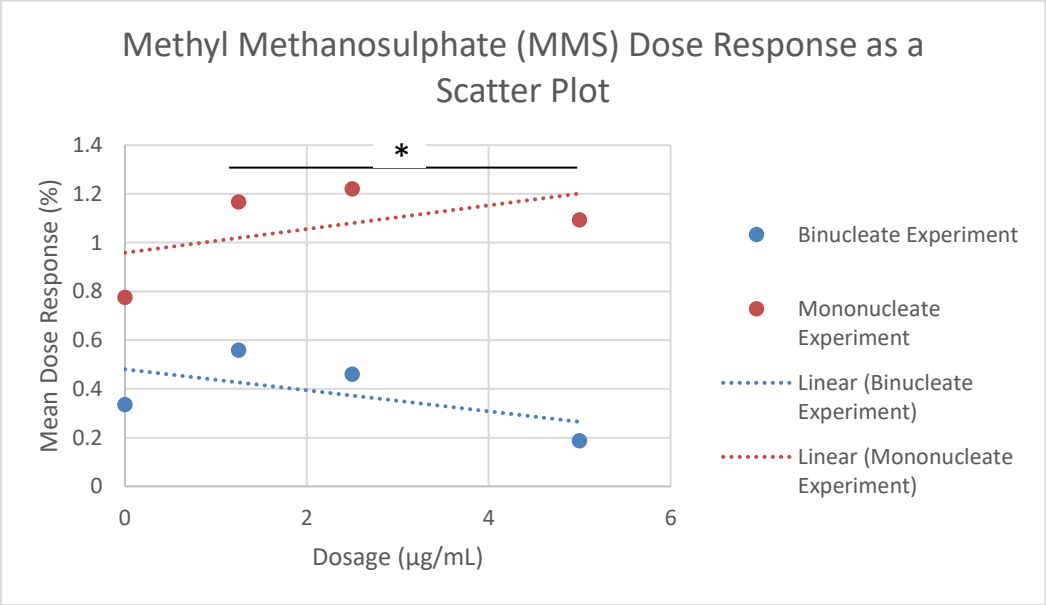


**Figure 20.** Comparison of mean MN dose response of the cyto-b and non-cyto b MN assay of MMS treated TK6 cells. N=2,. * p<0.05 using Fisher's Exact.

Fourthly, Fig. 21 shows the clastogenic effect of benzo(a)pyrene. The control has a mean MN frequency of 0.721% for binucleates with MN and 2.666% for mononucleates with MN. In the mononucleate experiment, as the dose increases to 3.0 µg/mL, the mean MN frequency increases to 4.934%. However, in the cyto-b experiment, the initial increase in dose to 2.0 µg/mL causes a decrease in MN frequency to 0.508%, then as the dose increases again to 3.0 µg/mL, the frequency of MN in binucleated cells increased to 0.721%. The N=3 data from the mononucleate experiment failed the Shapiro-Wilks test, showing that the data is not normally distributed, however it did pass the Bartlett's test which means the individual MN frequencies in the dose range are comparable. Therefore, the non-parametric one-sided Dunn's test was applied. The increase in MN frequency was not found to be statistically significant (p>0.05). However, the binucleated cell data (also N=3) passed the Shapiro-Wilks and Bartlett's tests concluding it is normally distributed and homogeneously varied. The one-sided Dunnett's test on the other hand showed the data was not statistically significant (p>0.05).

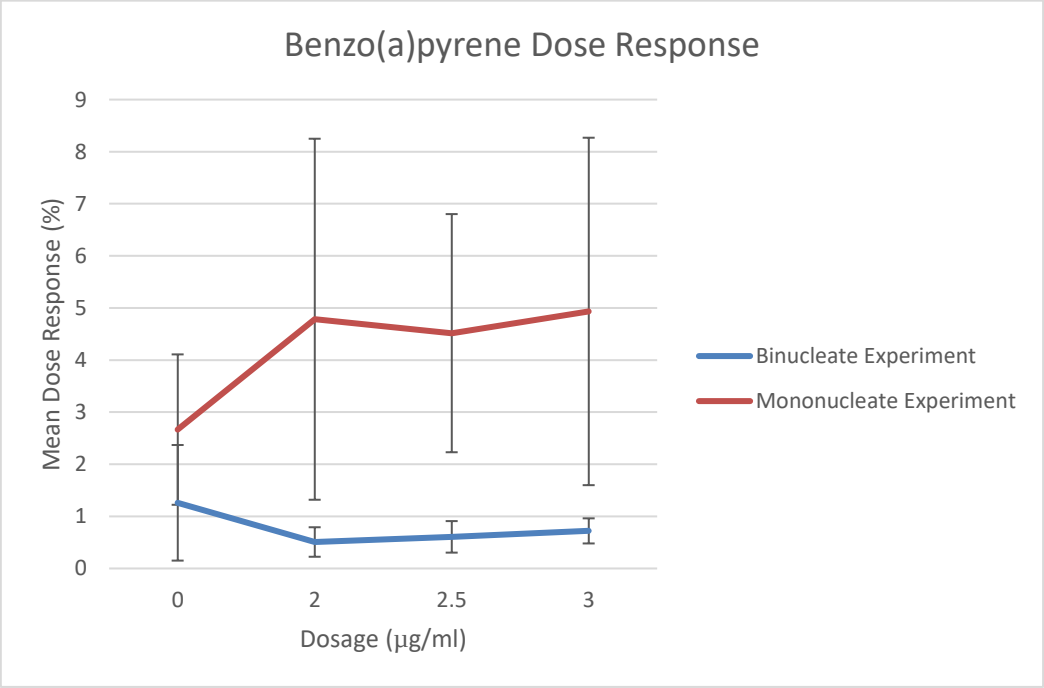



**Figure 21.** Comparison of MN dose response of the cyto-b and non-cyto b MN assay of benzo(a)pyrene treated TK6 cells. N=3, mean = +/- standard deviation.

Fig. 22 validates the non-genotoxic properties of Aroclor. The scatter plot below shows that there is no correlation between the dose of aroclor that the cultured TK6 cells were exposed to, and the percentage frequency of MN identified by the neural network.

The control has a mean MN frequency of 0.292% for binucleates with MN and 1.138% for mononucleates with MN. In the mononucleate experiment, as the dose increases to 0.0006 µg/mL, the mean MN frequency decreases to 0.396%. Then as the dose increases further to the top dose of 0.004 µg/mL, the MN frequency of mononucleates with MN increased to 0.992%. On the other hand, in the cyto-b experiment, the frequency of MN remained moderately constant. As the dose increased to 0.004 µg/mL, the MN frequency slightly decreased to 0.187%. The mononucleate experiment data was N=2, therefore the non-parametric Fisher's Exact test was conducted. This test showed that the dose response curve was statistically significant ($p<0.0001$) for the dose range 0.0004 – 0.0006 µg/mL, and $p<0.05$ for the dose range 0.0008 – 0.004 µg/mL. The binucleate cell experiment data on the other hand was found to not be statistically significant, using the Fishers Exact test ($p>0.05$).

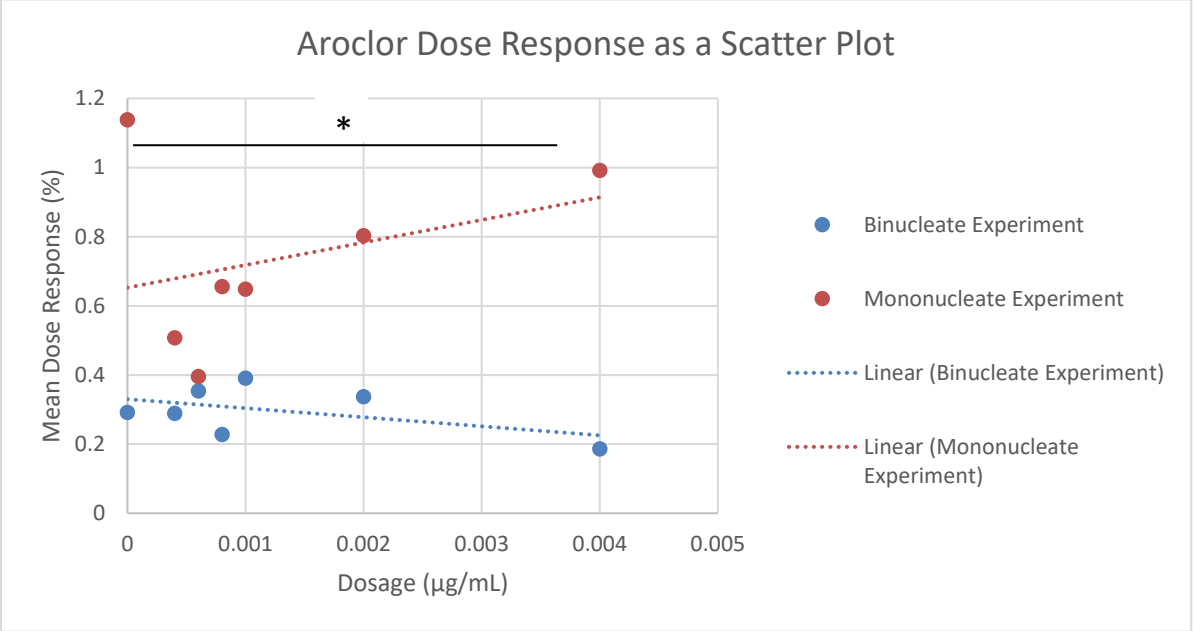



**Figure 22.** Comparison of mean MN dose response of the cyto-b and non-cyto b MN assay of aroclor treated TK6 cells. N=2, * $p<0.05$ using Fisher's Exact.

Finally, Fig. 23 shows the non-genotoxicity of D-Mannitol. In a similar fashion to the results of Aroclor exposure, the scatter plot below shows no correlation between the dose of D-Mannitol and the percentage frequency of MN identified by the neural network. This is therefore evidence that D-Mannitol has no genotoxic properties.

The control has a mean MN frequency of 0.596% for binucleates with MN and 0.290% for mononucleates with MN. In the cyto-b experiment, there is a slight decrease in MN frequency as the dose increases. When the dose increases to 2000 µg/mL, the MN frequency in binucleated cells decreases to 0.520%. However, for the mononucleate experiment, a different pattern is exhibited. As the dose increases to 1000 µg/mL, the mean MN frequency in the mononucleated cell sample increases initially to 0.433%. Then when the dose increases to 2000 µg/mL, the MN frequency in mononucleated cells decreases again to 0.313% which is only an 0.023% increase from the control. Both experiment data sets were N=2 so the Fisher's Exact test was conducted. In both experiments, the p values were greater than 0.05, therefore the results are not statistically significant.

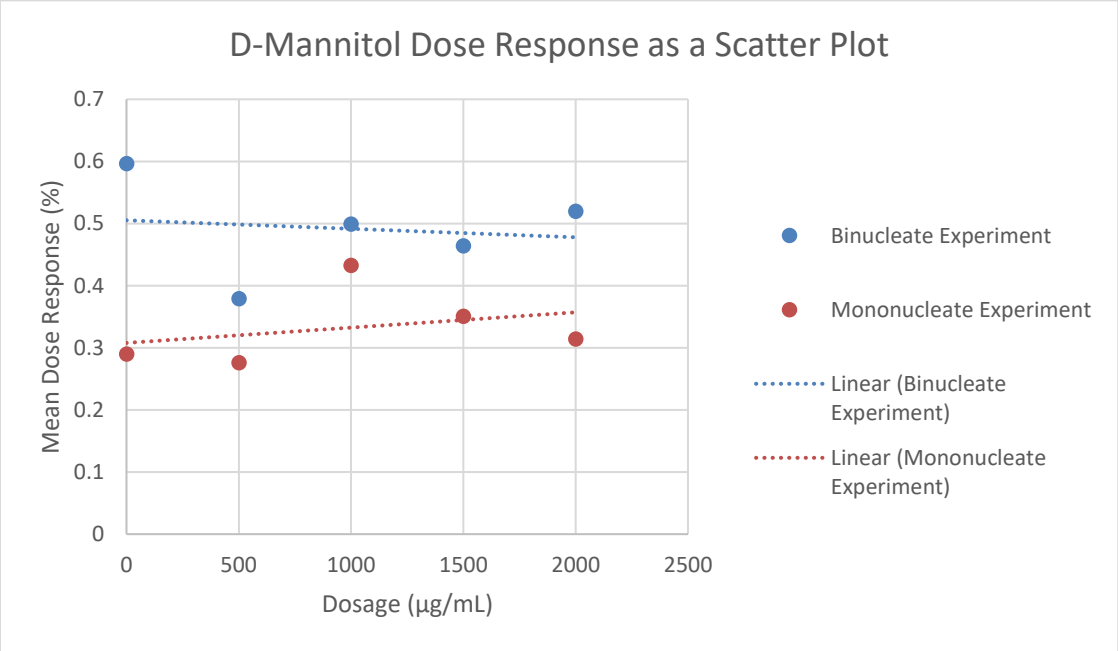



**Figure 23.** Comparison of mean MN dose response of the cyto-b and non-cyto b MN assay of D-Mannitol treated TK6 cells. N=2,.

### 3.2.2 Investigating Statistical Significance by Benchmark Dose Analysis

The pairwise and two-way contingency statistical testing methods, using the Dunnett's, Dunn's and Fisher's Exact tests did not show concordance in statistical significance. A common finding was that the tests did not show statistical significance for the dose response curves on the binucleate experiments but did show significance on the mononucleate experiments. Therefore, it was essential to carry out further statistical tests as alone these figures are not informative to draw conclusions regarding the project hypothesis.

To better understand the applications of the calculated dose responses above to genotoxicology, the exponential and Hill models for assessment of continuous genotoxicity data using BMD were used to analyse the mononucleate and binucleate experiment data of Fig. 18 -23. Fig. 24 shows the BMD analysis of the six chemicals tested using the mononucleate (non cyto-b) experiment and the dotted lines show the prediction of the BMR at 50% to in turn estimate the BMD50. The AIC for the exponential model was 98.76 and for the Hill model the AIC was defined as 98.94. The following table outlines the predicted BMD50, BMDL50 and BMDU50 values for each chemical according to each model.

| COMPOUND | FITTED MODEL | BMD50 (µg/mL) | BMDL50 (µg/mL) | BMDU50 (µg/mL) |
|---|---|---|---|---|
| Aroclor | Exponential | 0.005 | 0.002 | Infinity (Inf*) |
| | Hill | 0.006 | 0.002 | Inf |
| Benzo(a)pyrene | Exponential | 2.280 | 0.763 | Inf |
| | Hill | 2.202 | 0.726 | Inf |
| Carbendazim | Exponential | 0.428 | 0.167 | 0.655 |
| | Hill | 0.407 | 0.158 | 0.640 |
| D-Mannitol | Exponential | 8230.000 | 1400.000 | Inf |
| | Hill | 8286.000 | 1330.000 | Inf |
| MMS | Exponential | 14.280 | 2.950 | Inf |
| | Hill | 14.180 | 2.790 | Inf |
| Vinblastine | Exponential | 0.001 | 0.001 | 0.002 |
| | Hill | 0.001 | 0.001 | 0.002 |

**Table 2.** BMD50, BMDL and BMDU values for compounds tested in Newcastle University. *Inf is abbreviation of infinity.

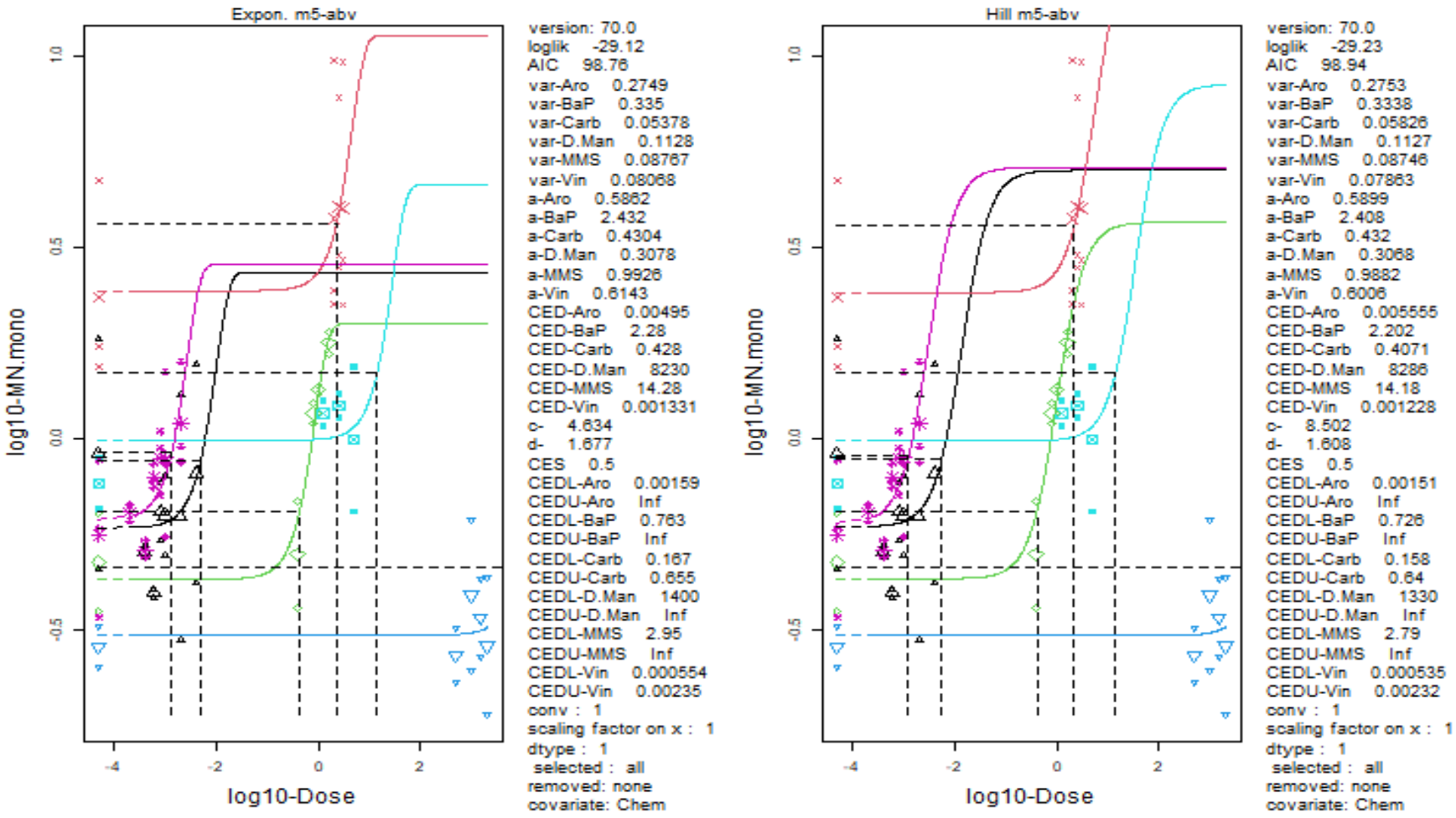



**Figure 24.** Benchmark dose (BMD) analysis using Exponential (left panel) and Hill (right panel) models. The curves represent non-cyto b MN assay dose response data for all six compounds. Carbendazim (green), vinblastine (pink), MMS (light blue), benzo(a)pyrene (red), Aroclor (black), D-mannitol (dark blue). Both models use covariate dependent parameters. Horizontal and vertical dashed lines represent the benchmark response at 50% to calculate the BMD50.

Fig. 25 shows the range of BMDL50 and BMDU50 from fig. 24 in a graphical form. When the dotted lines are further apart, this demonstrates that the data does not fit well with the exponential and Hill models and there is therefore a greater amount of variation within the data.

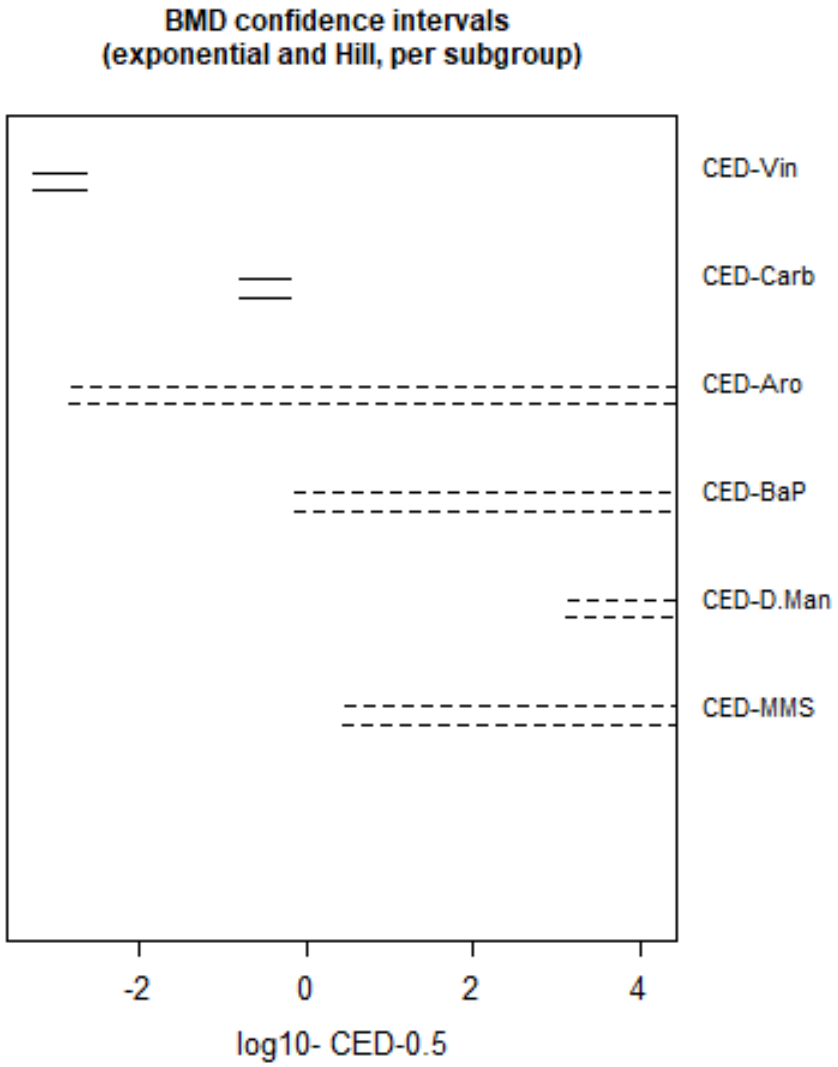



**Figure 25.** BMD confidence intervals of the exponential and hill models of BMD analysis. Represents the range of upper and lower BMD50 values.

### 3.3 Dose Response Data collected at Swansea University

The second set of data the DeepFlow deep learning neural network was applied to was collected from the ImageStream X Mark II® at Swansea University Medical School. Images of the carbendazim and MMS treated TK6 cells were collected as raw image files and subsequently processed as per the method explained above. Fig. 36 is the mononucleated MN dose response of the aneugen, carbendazim. The control has a mean MN frequency of 0.386%. Then as the dose increases to 1.6 µg/mL, the mean MN frequency increases to 2.881%. The Shapiro-Wilks and Bartletts tests showed that the MN dose response data was normally distributed and homogeneously varied so a one-sided parametric Dunnett's test could be applied. The increase in MN frequency from 0.0 – 1.6 µg/mL was found to be statistically significant ($p<0.05$).

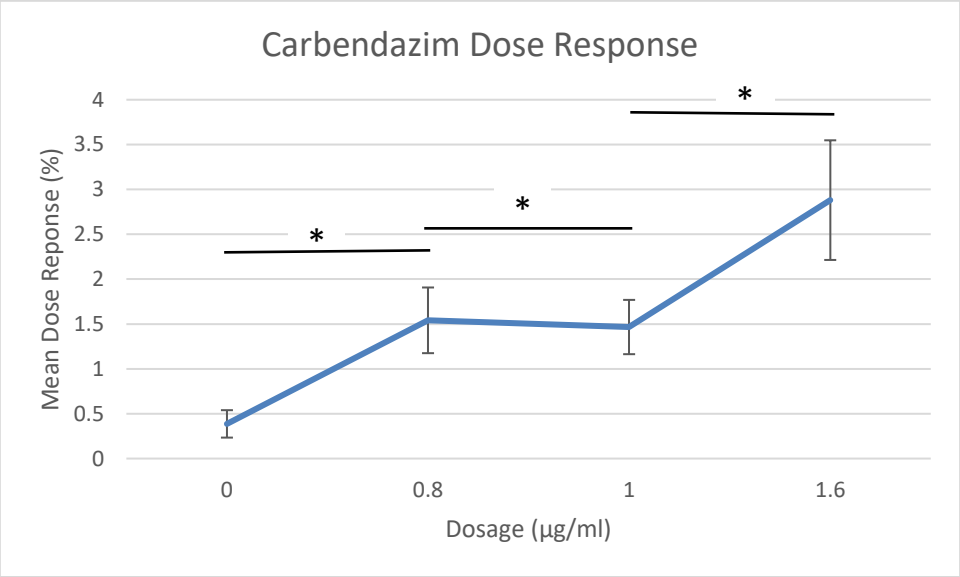



**Figure 26.** MN dose response of the non-cyto b MN assay of Carbendazim treated TK6 cells. N=3, mean = +/- standard deviation. *p<0.05 using one-sided Dunnett's test.

Secondly, Fig. 27 is the mononucleated MN dose response of the clastogen, MMS. Similarly, to findings from Newcastle University, the dose response for MMS had a bell-shaped curve where the MN frequency initially increases, then decreases again at the top dose. The control has a mean MN frequency of 0.279%. Then as the dose increases to 2.5 µg/mL, the mean MN frequency increases to 0.707%. As the dose increases further to 5.0 µg/mL, the mean MN frequency decreases again to 0.555%. The data passed both the Shapiro-Wilks and Bartletts tests which showed that the MN dose response data was normally distributed and homogeneously varied so a one-sided parametric Dunnett's test could be applied. The increase in MN frequency from 1.25 – 2.5 µg/mL was found to be statistically significant ($p<0.05$). However, the other dose response points within the dose range were not statistically significant ($p>0.05$).

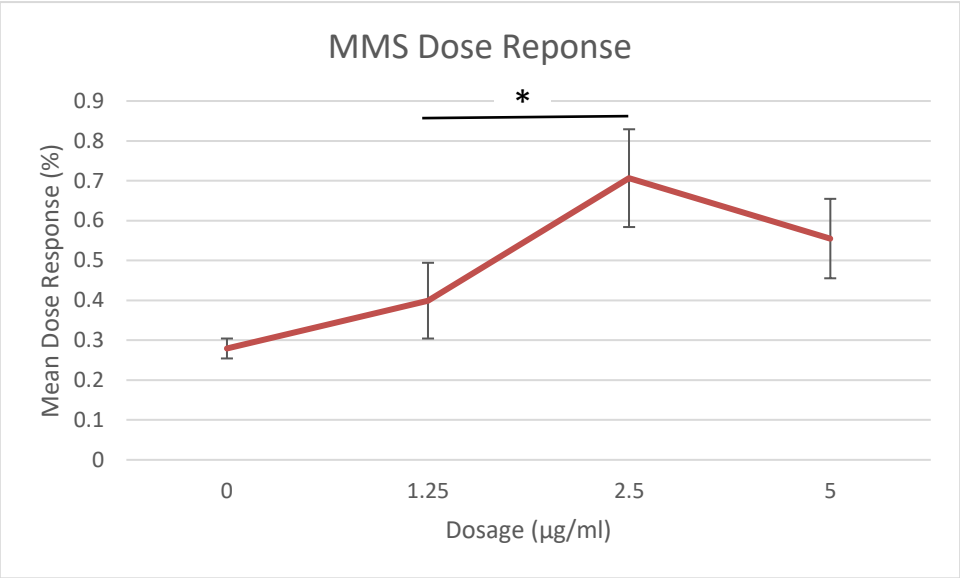



**Figure 27.** MN dose response of the non-cyto b MN assay of MMS treated TK6 cells. N=2, mean = +/- standard deviation. *p<0.05 using one-sided Dunnett's test.

BMD analysis was then carried out which is presented in Fig. 28. The AIC for the exponential model was 17.52 and for the Hill model the AIC was defined as 17.56. The following table shows the estimation of the BMD50, BMDL50 and BMDU50 for the MN dose responses of each chemical.

| COMPOUND | FITTED MODEL | BMD50 (µg/mL) | BMDL50 (µg/mL) | BMDU50 (µg/mL) |
|---|---|---|---|---|
| Carbendazim | Exponential | 0.084 | 0.004 | 0.263 |
| | Hill | 0.086 | 0.004 | 0.263 |
| MMS | Exponential | 1.437 | 0.210 | 6.980 |
| | Hill | 1.453 | 0.213 | 6.980 |

**Table 3.** BMD50, BMDL and BMDU values for compounds tested in Swansea University.

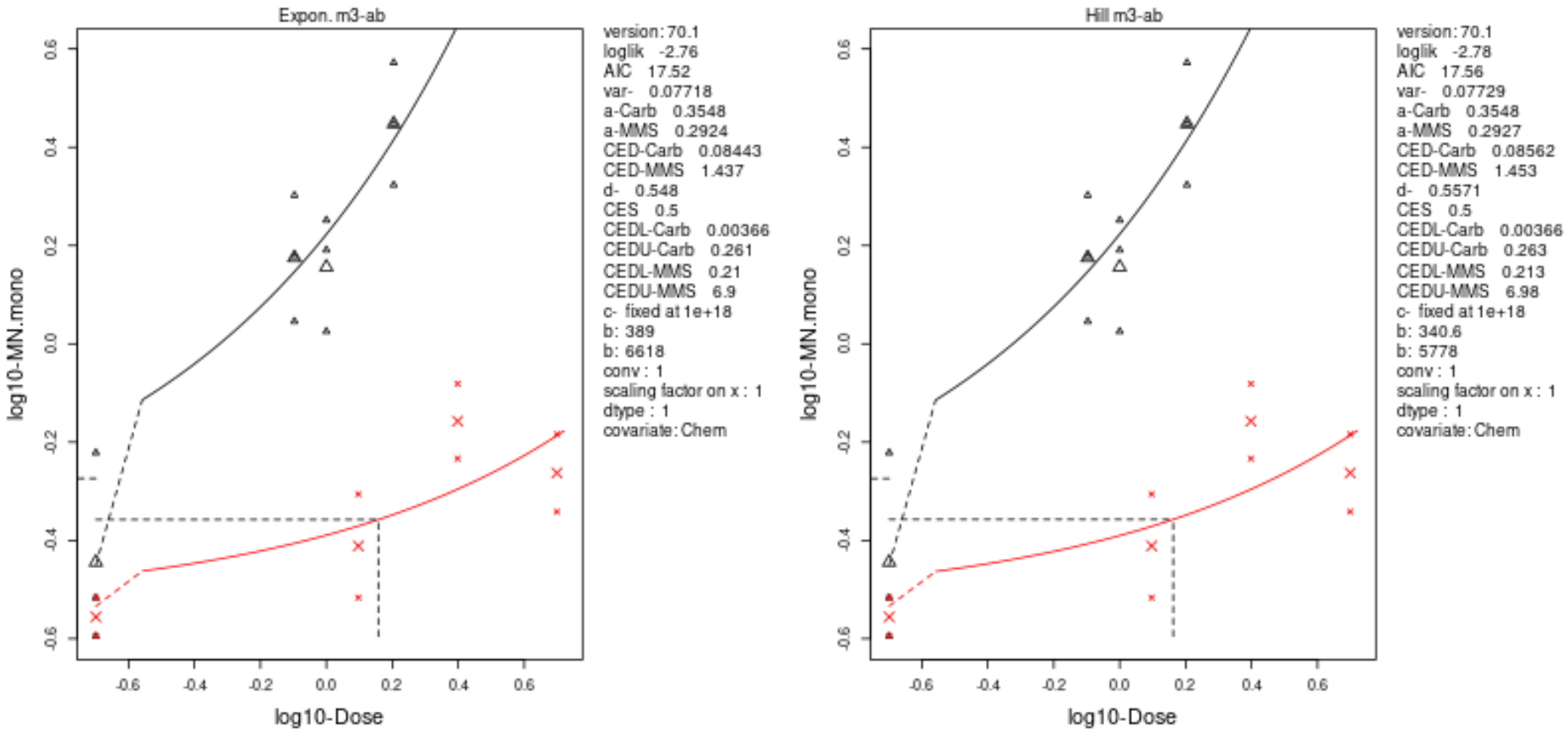



**Figure 28.** Benchmark dose (BMD) analysis using Exponential (left panel) and Hill (right panel) models. The curves represent non-cyto b MN assay dose response data for carbendazim (black) and MMS (red). Both models use covariate dependent parameters. Horizontal and vertical dashed lines represent the benchmark response at 50% to calculate the BMD50.

### 3.4 Dose Response Data from collaboration with Aberystwyth University

The final set of data the DeepFlow deep learning neural network was applied to was images taken at Aberystwyth University. Fig. 29 is the mononucleated MN dose response of the non-genotoxic compound, quinoxaline. The data shows that the MN frequency does not deviate far from the control dose point which highlights that quinoxaline is non-genotoxic. The control has a mean MN frequency of 0.445%. Then as the dose increases to 0.5 µg/mL, the mean MN frequency increases to 0.563%. As the dose further increases to the top dose of 4.0 µg/mL, the mean MN frequency decreases again to 0.397% which is 0.048% different from the control. The data failed both the Shapiro-Wilks and Bartletts tests hence showing that the MN dose response data was not normally distributed and the individual MN frequencies within the dose range were not comparable. Therefore, a one-sided non-parametric Dunn's test was applied. The change in MN frequency within the dose range of quinoxaline was not found to be statistically significant ($p>0.05$).

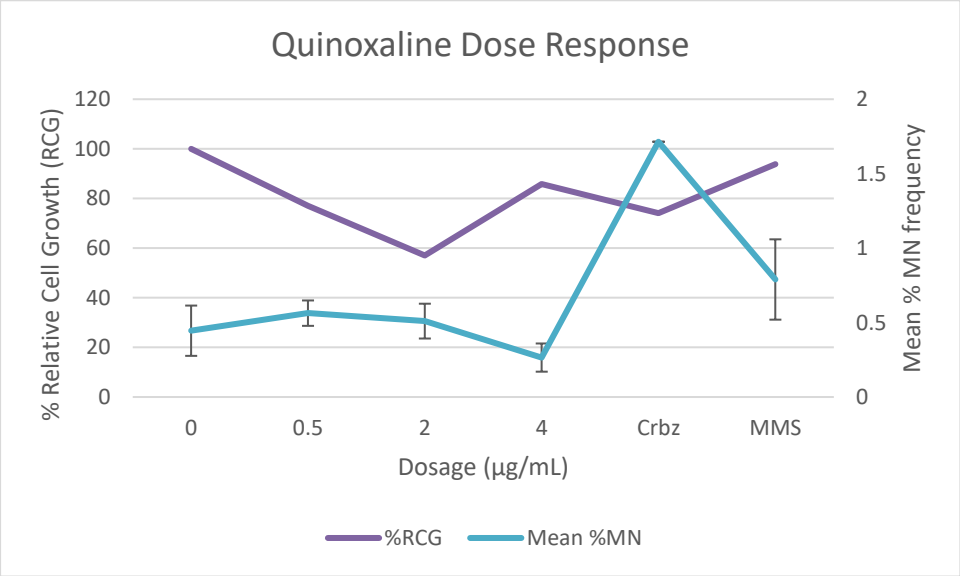



**Figure 29.** MN dose response of the non-cyto b MN assay of Quinoxaline treated TK6 cells. N=3, mean = +/- standard deviation. Relative Cell Growth (%RCG) is also displayed (Padalino et al., 2021).

BMD analysis was again carried out on the data and is presented in Fig. 30. The AIC for the exponential and Hill models was 18.22. However, due to the non-genotoxic nature of quinoxaline and the wide variation within the data, the data did not fit at all to the exponential and Hill models, therefore no BMD50, BMDL50 and BMDU50 values could be estimated.

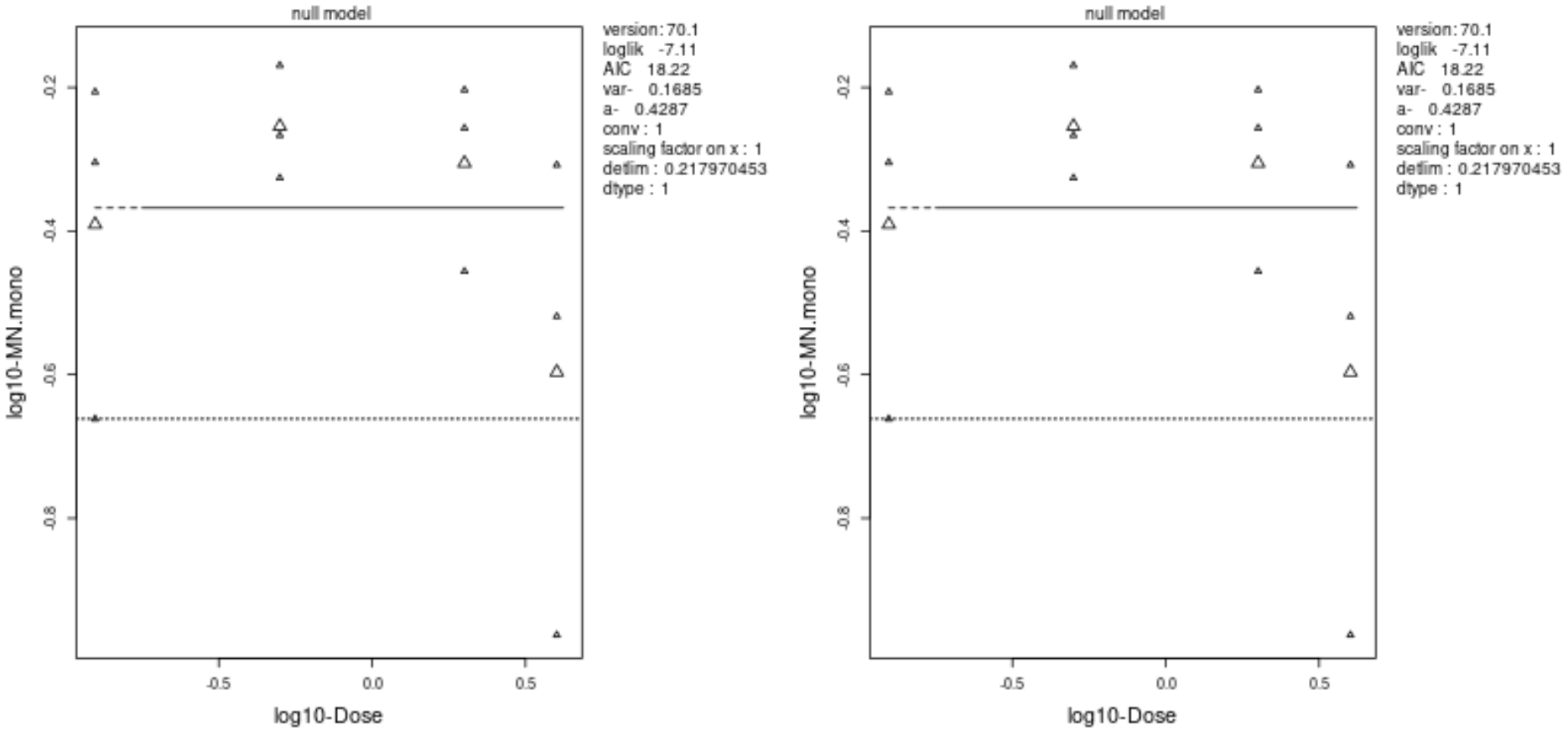



**Figure 30.** Benchmark dose (BMD) analysis using Exponential (left panel) and Hill (right panel) models. The curves represent non-cyto b MN assay dose response data for quinoxaline. Both models use covariate dependent parameters.

# 4. Discussion

The *in vitro* MN assay is a globally used method used to quantify the DNA and chromosomal damage induced by test chemicals found in developing pharmaceuticals, cosmetics and agricultural products. The manual scoring method is both time-consuming and scorer-subjective, therefore this method that is heavily relied upon creates a bottleneck in the use of the assay. Therefore, it has been significant in this project to show that imaging flow cytometry coupled with deep learning image scoring can represent a reliable and accurate method for automating the *in vitro* MN assay.

To successfully allow the progression of automating the *in vitro* MN assay using imaging flow cytometry and deep learning neural networks, it was firstly essential to assess how accurately the algorithm was performing. During this study, it was determined at what confidence level the algorithm assigned to correctly scored images. Appendices 1 and 2 show that for both binucleates and mononucleates with and without MN, when the algorithm was 70% confident and above, the images were consistently scored correctly. However, it is notable that there was one exception to this finding in the binucleated cells and seven exceptions in the mononucleated cell data, which are identified as being incorrectly scored at 70% confidence above. When an image is regarded as having been scored correctly, this means that the phenotype of the image can be coherently classified and the phenotypic classification that the algorithm assigned the image matches what the manual classification was. If the algorithm's classification of an image does not match the classification given manually, the image is viewed again and if required, a more experienced scorer also views the image, and a conclusion is made if the algorithm has scored the image incorrectly or correctly. Manual visualisation of these images however showed that there were shadows and debris which interfere with the neural network's training of MN identification. Therefore, this must be considered in future study and therefore an aim can be made to train the network to identify MN within images containing shadows and debris, through either defining more specific parameters for MN classification or by developing techniques to use multiplex labels to train an algorithm. This assay however has demonstrated that the trained neural network is correctly scoring MN events in sample images which validates its applicability to its routine use in genotoxicity testing of new compounds and chemicals.

Once the accuracy of the pretrained deep learning neural network was determined, the algorithm was used to calculate dose responses of multiple compounds from different laboratories. It must be highlighted that the scored images used in the dose response calculation were only included if they were scored with 70% confidence and above. Figures 18 - 23 firstly show that the algorithm produced results that were parallel to the findings of the ECVAM guidelines from images tested in Newcastle, Swansea and Aberystwyth universities. This is significant for the prospects of this method being routinely used in genotoxicity testing. In 2008, ECVAM published a list of chemicals and compounds and the results of their genotoxic test. Compounds could either be labelled as genotoxic, non-genotoxic or as negative *in vitro* but have induced gene mutations in mouse lymphoma cells. This list of recommendations was published in order to reduce the abundance of misleading positive results (Kirkland et al., 2016). Therefore, it is significant that the network's MN dose response curves and subsequent classification of the genotoxic or non-genotoxic nature of a compound agree with the ECVAM classifications because it shows that the neural network is less likely to produce misleading positive results when unseen chemicals and compounds are tested with this technology.

In addition to the results aligning with the ECVAM guidelines, it is also important for the negative control data to be compared to a historical negative control data set. The use of a historical negative control data set is significant because it helps to determine if the laboratory is competent to conduct

the *in vitro* MN assay. The OECD have stated that in addition to evaluating the statistical significance of the dose response data, at least one of the dose response data points must fall outside of the range of the historical negative control data. The Health and Environmental Sciences Institute (HESI) Genetic Toxicology Technical Committee (GTTC) conducted a study in 2018 in which they collected dose response data from 13 laboratories using TK6 cells to perform the *in vitro* MN assay to produce a historical negative control database. The mean frequency of micro-nucleated cells in 1000 cells ranged from 3.2 – 13.8. This 4.31-fold difference does show that there was a high degree of variation but no laboratory in the study was found to be an outlier as no individual data sets had abnormally high variability. Therefore, this historical control database is valid and can be used to compare this project's negative control data to. The negative control data collected from Newcastle and Swansea universities across the 6 tested compounds ranged from 6 – 16.7 micro-nucleated cells in 2000 cells. These results have a lower variance than the historical negative control database from the HESI GTTC, however, to improve the significance of this comparison, the number of replicates should be increased. (To generate the historical negative control database, one laboratory conducted an experiment with 795 replicates) (Lovell et al., 2018). Despite this, it is encouraging that the negative control data from this project can be compared to a historical negative control database as this gives significant confidence in the reliability and conformity of the neural network as an automated scoring method for the *in vitro* MN assay.

The dose response curves of the seven compounds tested in this project, showed that all of the compounds apart from MMS showed either an increase or decrease in MN frequency as dose increased. However, interestingly, MMS exhibited a bell-shaped curve in data collected from Newcastle and Swansea Universities, where the MN frequency increased then decreased at the top dose. This leads to the suggestion that MMS in fact has cytotoxic properties as well as genotoxicity. Previous publications have shown that when cells from multiple cell lines such as 10T1 cells are exposed to MMS, the cells were no longer able to divide and form colonies in culture. There is evidence to suggest that MMS is an alkylating agent which specifically methylates cellular macromolecules which in turn leads to the production of nucleophiles. These nucleophiles then actively attack carbon atoms in the structures of cellular molecules, including plasma membrane proteins. Therefore, at the higher doses of MMS, there is increased nucleophilic damage to the plasma membranes of cells hence sending the cell into apoptosis and necrosis (Smith & Grisham, 1983).

In figures 18 -23 which were produced from images of compounds tested in Newcastle university, a common yet notable phenomenon found was that the percentage MN frequency in mononucleated cells was higher than that of binucleated cells. The only exception to this trend was the dose responses of D-mannitol.

From the perspective of the mechanism of MN production, increased MN in mononucleated cells is unexpected. Cell division is necessary to produce MN once a chemical or compound has induced damage in the cells. The CBMN assay is very effective as it can detect clastogenic and aneugenic damage in cells that have previously divided. This is because cytochalasin B inhibits cytokinesis. However, the mononucleated assay can only detect MN in mononucleated cells which may not have gone through division since damage was induced (MNi). A reason why a lower frequency of micronucleated binucleated cells has been observed in this study is that there may have been a delay in the division of cells with damaged DNA. The regulation of the cell cycle is closely coupled with DNA repair mechanisms that are initiated in response to DNA damage, and most DNA repair mechanisms occur before the S phase (when DNA is replicated). Therefore, cells that contain DNA damage get arrested during G1 and G2 phases in order to provide time for the DNA to be repaired

and to avoid mutations getting fixed into the daughter cells of mitosis. This division delay in cells containing damaged DNA therefore means that micronucleated binucleate cells appear at a later time in the culture, in comparison to non-damaged cells. This therefore suggests that it will be beneficial to harvest binucleated cells at a later time point to ensure that DNA-damaged cells have been able to complete cytokinesis before harvest and subsequent imaging. Therefore, in future experiments following on from this project, a preliminary study should firstly be conducted to determine when the frequency of MN plateaus in binucleated cells cultured with the different test compounds. After this, the CBMN assay can be run and the binucleated cells should be harvested for imaging at the time point of when the number of MN stops increasing and levels off. This ensures that all of the DNA-damaged cells have been able to complete nuclear division. (Kirsch-Volders & Fenech, 2001).

On the other hand, seeing higher frequencies of MN scored in mononucleated cells is synergistic with the expected outcome from the perspective of training the deep learning neural network. For a network to be trained to identify mononucleated cells with MN, it must detect two circular bodies, with one being larger than the other by $1/3^{rd}$ to $1/16^{th}$. However, for the neural network to correctly score binucleated cells with MN, it must detect three circular entities, with two being approximately the same size and one smaller. Therefore, this shows that there is a greater number of outcomes that can still lead to the same phenotype score of binucleates with MN. This means that there is more chance for the algorithm to become confused and mis-score these MN BN events.

The underscoring of binucleated cells with MN is a significant limitation of the deep learning neural network as the frequency of MN in binucleated cells is an essential component of the dose response calculation. Therefore, work following on from this project should focus on the optimisation of the ground truth set of images to have a higher representation of binucleated with MN, including images that contain shadows and debris to try and prevent the algorithm from slipping up on images where the classification is less obvious. To optimise the ground truth, human scorers score a sample of images that the algorithm scored as 'other or unscorable' to determine whether the images classified as so, could instead be scored as an objective phenotype, this will therefore re-supplement the ground truth with images with a definite classification and this hence may aide the network in identifying binucleated cells with MN.

Previous publications have shown that the magnification of the objective lens and the resulting depth of focus can influence the accuracy of imaging MN. The magnification used in this project was x40 which results in a 4 µm depth of focus. On the other hand, x60 magnification produces a depth of focus of 2.5 µm. Therefore, using x40 magnification can allow for MN to appear brighter than the main nuclei if the MN are located in a different depth of focus than the main nuclei. Since the MN appear brighter, they can be identified and thus score more coherently which increases the accuracy of the calculated dose response of a chemical or compound (Rodrigues, 2018). Consequently, even though x40 magnification is a slightly lower magnification, it can detect MN that were dimmer in the image, which may have been missed at x60 magnification with the shorter depth of focus.

An additional result identified in this study was that the results from Swansea University's data set had lower resolution, however the accuracy and precision of the results are significantly higher than the results from Newcastle and Aberystwyth universities. A factor contributing to this is the transport the samples underwent from Swansea to each of these two universities. The samples contain DNA and RNA therefore, if the samples were not packaged appropriately, gene expression may be repressed or induced. This can lead to subsequent DNA degradation which can thus cause abnormal behaviour in the cell. For example, in environments of elevated heat and stress, the *Fas/FasL* signalling pathway is upregulated which is a pro-apoptotic pathway (Bouchier-Hayes et al.,

2010). Therefore, the cells will undergo uncontrolled cell death leading to lots of debris extracellularly. If this has occurred to the cells being used for the ground truth data set, the images will not be clear so manual and automated scoring will be difficult, thus the data decreases in accuracy and precision. To improve the accuracy and precision of ground truth images of cells that have travelled, DNA and RNA stabilizers such as EDTA can be added to the samples to reduce DNA and RNA degradation. Also, the samples should be in a controlled temperature environment. Previous studies have shown that a temperature of 4°C provides stability for cells being transported. Finally, once the cells have reached their destination, detection experiments of biomarkers should be performed. Expression of genes such as *GAPDH* and *interleukin-8* (IL-8) can be tested as it is known in what conditions these genes are induced or repressed, therefore this can allude to whether other genes of interest could have been up or downregulated during transportation (Malentacchi et al., 2016).

When dose responses were calculated for the data sets from each university, it was essential that statistical tests were carried out to check if the data was normally distributed and statistically significant in order to determine observed effect levels and benchmark responses. These statistical tests were conducted according to the protocol defined in (Johnson et al., 2014). If the data was N=3, found to be normally distributed and homogeneously varied, a Dunnett's test was used which is equivalent to a one-sided ANOVA. The Dunnett's test calculates variation between the samples with dose and control samples. Therefore, the data must be normally distributed to conduct this calculation. On the other hand, when the data was N=2, the Fisher's Exact test was used. The Fisher's Exact test is as statistical test of independence which determines if the proportion of a variable you are testing is different depending on the value of the other variable (McDonald, 2009). Of the seven compounds tested across the three laboratories in this project, only the dose response from carbendazim in the cytochalasin-B (binucleates) experiment was found to be statistically significant and in the mononucleates experiments, the dose responses for Aroclor, carbendazim and MMS were found to be statistically significant.

From these findings, it could be concluded that the lowest observed effect level (LOEL) for carbendazim is 0.8 µg/mL. This is because statistical significance in the Newcastle data was found in the dosed samples from 0.8 µg/mL upwards, and in the Swansea data, 0.8 µg/mL was the lowest dosed sample. However, since the 0.4 µg/mL dosed sample was not found to be significant in the Newcastle data, it could be determined that the no observed effect level (NOEL) was 0.4 µg/mL.

It was common in the results to see that the data had a relatively high degree of variance from the mean. This was found particularly in the data from the binucleates experiments. This could be due to several reasons. Firstly, errors may have arisen during the cell treatment process in the laboratory in one of the experiment repetitions. For example, there may have been irregularities in the amount of time each sample was left exposed to the compound. Another reason for high variation in the results is under or over scoring of MN events by the deep learning neural network. This is likely to be the main contributor to variance because binucleated cells with MN are under-represented in the ground truth used to previously train the neural network. Therefore, the frequency of MN events that have been mis-scored is increased because the algorithm has had less training on this phenotype. This is a limitation of automating the *in vitro* MN assay because MN are less abundant than binucleated and mononucleated cells without MN, therefore, these categories will always be scored more accurately and abundantly. To overcome this limitation, technology is being developed to train a deep learning neural network on a synthetic ground truth where each category is manually drawn on computer software. This means that the neural network will no longer be trained on images taken on the imaging flow cytometer. Therefore, there can be equal representation of the

cell phenotypes in the ground truth so there is less and eventually no bias when unseen images are scored using the neural network.

However, a key concern for the application of deep learning neural networks to the *in vitro* MN assay, that still exists with a synthetic ground truth, is that they have a reputation as a 'black box'. The 'black box' concept illustrates the lack of transparency and interpretability when deep learning neural networks transforms the input data into an output. For example, in this study, there is no quantifiable measure that can be analysed that explains how the algorithm classified an image with a particular phenotype. This therefore makes the algorithm very difficult to validate and optimise. Therefore, going forward, to make this method more robust, more work must be done to increase the transparency and interpretability of the neural network (Sheu, 2020). Developing the synthetic ground truth may help with this concern because the data that the algorithm is trained on will have decreased background noise, therefore the output that the algorithm predicts will be easier to interpret.

Pairwise testing is a technique widely used by researchers to test their data for statistical significance and observable effect levels. It is also relatively easy to carry out using programs such as GraphPad Prism, SPSS and mutait.org. However, there are limitations to pairwise testing that reduce its power as a tool. Pairwise testing is based completely on the study design, space between doses and sample sizes. Therefore, for example when determining the LOEL doses, only the experimental doses can be used, and this is not representative of what the true LOEL is of a compound. Extrapolation of the assay's results hence does not provide reliable data for assigning observable effect levels of a compound. Therefore, because of this, the LOEL value from the data set may not actually be a dose with no effect, and this can have serious implications when bringing the test chemical or compound into trials (OECD, 2010).

Due to these limitations to pairwise testing, it is advantageous to carry out additional statistical tests to provide a further in-depth analysis of the continuous data. An alternative approach is the benchmark dose analysis (BMD) approach which is widely accepted as a test used to determine benchmark responses (BMR) and PoD's. A factor that separates BMD from pairwise testing is that it uses confidence levels. Therefore, finding out degrees of confidence makes the results more reliable and valid. BMD is also very effective at identifying trends in the dose response curve, therefore, the BMR, BMDU and BMDL values are not necessarily the experimental doses. This therefore increases the accuracy and precision of these values. As demonstrated in the results of this project, BMD analysis allowed BMR, BMDU and BMDL values to be calculated even when the results were not found to be statistically significant in the pairwise testing (Tables 2 and 3.) This therefore allows results and conclusions to be drawn from acquired data despite not being normally distributed and or statistically significant. However, it should be highlighted also that BMD analysis has limitations to be aware of before using this technique. BMD has limited use to data from non-genotoxic compounds. Therefore, it is not possible to extrapolate the dose response curves for these compounds to determine if there is an upper dose limit when the compound starts to become genotoxic or cytotoxic. For example, quinoxaline and D-Mannitol are non-genotoxic, however since they are used clinically, it is beneficial to calculate if there is an upper dose where it starts to become toxic (Sand et al., 2008).

When applying imaging flow cytometry and deep learning neural network technology to the *in vitro* MN assay, it can be challenging to advance the specificity of the deep learning network to identifying MN events. This is because the parameters for MN classification will have to represent very subtle morphological changes of MN compared to parent nuclei, therefore this can in turn lead to over training of the deep learning model. Hence, as deep learning codes become more advanced and

computer power increases, developments can occur through a multiplexed MN assay. In a multiplexed MN assay, fluorescence stains and fluorescent antibodies can be used to detect different markers that can show if cytotoxicity or genotoxicity is occurring post compound or chemical exposure. For example, fluorescence can be used to detect the presence of apoptotic and necrotic bodies. This is advantageous as these entities are morphologically similar to MN so this would avoid mis-scoring. Additionally, MN have their own nuclear envelope so fluorescent antibodies specific to a protein of the nuclear envelope could be incorporated into the protocol during cell treatment and exposure and deep learning neural networks can be trained to identify the phenotype. MN's nuclear envelopes differ in structure compared to parent nuclei, for example MN nuclear envelopes are composed with significantly more emerin protein. Therefore, if emerin fluorescent antibodies are used, this could provide more conclusive results for the presence of MN (Maass et al., 2018). Another example of multiplex labelling used in the *in vitro* MN assay is the use of fluorescent antibodies specific to the kinetochore of chromosomes. This can allow the detection of chromosomes with centromeres and acentric chromosomes. This development can aide the identification of MN which contain these entities but can also contribute to proposing the mode of action of a chemical or compound. The presence of chromosomes with full centromeres is indicative of aneugenic modes of actions and the presence of acentric chromosomes can indicate clastogenic activity (Rodrigues et al., 2021).

# 5. Conclusion

In this project it has been demonstrated that the combination of imaging flow cytometry and deep learning neural network automated scoring mechanisms can produce accurate dose response results which are parallel to the ECVAM guidelines and akin to the gold standard. The developed deep learning neural network was applied to images of individual cells taken at three different laboratories in Swansea, Newcastle, and Aberystwyth universities. Even though the ground truth population of binucleated MN cells was very small, the deep learning neural network developed in this project was shown to achieve high accuracy levels in identifying mononucleated and binucleated MN cells across all three data sets. Applying the algorithm to images from different locations has streamlined the automation of the *in vitro* MN further as researchers are able to use the pre-created ground truth and deep learning neural network to analyse unseen samples. This project has shown a positive step forward in the modernizing and streamlining of the genotoxicity testing of new chemicals and compounds in pharmaceuticals, cosmetics, agricultural and other industries.

The deep learning neural network was found to correctly score binucleated and mononucleated MN cells consistently, when the algorithm assigned 70% confidence in the phenotypic score. Then using only images that the algorithm scored with 70% confidence, the calculated dose responses were parallel to the ECVAM guidelines.

Forming greater ground truth populations leads to an increase in the occurrence of rarer cell phenotypes in the ground truth, such as binucleated cells with MN, therefore this could increase the accuracy of the deep learning neural network and thus improve the accuracy of the calculated dose responses. In the Newcastle dose responses, only the dose response curve between 0.8 – 1.6 μg/mL of carbendazim was found to be statistically significant. Likewise, in the Swansea dose responses, only carbendazim and MMS dose responses were found to be statistically significant. The measured dose responses across all three data sets in the CBMN assay were found to be lower than dose responses previously recorded, therefore this suggests that the binucleated MN cell accuracy is not high enough. Increasing the number of images in the ground truth population will allow for a greater representation of binucleated MN cells in the ground truth which can be manually scored and subsequently used to train the deep learning neural network. This in turn will allow the identification of binucleated MN cells to be increased and hence improve the calculated dose responses.

Previous studies have shown that different chemicals and compounds can induce the production of MN with slight morphological changes, therefore this could cause a bottleneck in a deep learning neural network's ability to identify micro-nucleated cells as the ground truth may not be representative of the abnormal phenotype. Therefore, an exciting future study could be based on the development of a deep learning neural network trained on multiplex labels such as fluorescent antibodies specific to proteins on the nuclear envelope of MN such as the emerin protein. Another example of a multiplex label is fluorescent antibodies specific to the kinetochore of chromosome. These labels are both strong biomarkers for genetic damage so will be alternative effective means to test the genotoxicity of chemicals and compounds.

With further optimisation of the neural network to improve the network's ability to recognise binucleated cells with MN and potentially identify different multiplex labels in the future, this protocol could be established as a 21$^{st}$ century assay for testing the genotoxicity of chemicals and compounds across multiple laboratories, industries, and research fields.

# Glossary

**Artificial Intelligence:** A branch of computer science that allows computers and machines to mimic problem-solving and decision-making abilities of the human brain.

**Benchmark Dose:** The dose or concentration of a test chemical or compound that produces a predetermined change in an adverse response compared to the response in an unexposed subject.

**Benchmark Response:** The predetermined change in response that the benchmark dose induces. Examples of benchmark responses include 5% or 10% increase in micronuclei frequency, body weight or erythrocyte count.

**Cytochalasin-B:** Cell-permeable mycotoxin that inhibits cytokinesis (cytoplasmic division) by preventing the formation of actin filaments. It is commonly used in the *in vitro* MN assay as it disrupts cytokinesis whilst nuclear division continues, thus leading to the formation of binucleated cells.

**Deep Learning:** A branch of artificial intelligence that imitates the human brain by using neural networks to identify cellular phenotypes from image data sets.

**Ground Truth:** A set of images that have been manually scored by the researcher and assigned a particular phenotype. The neural network is trained on these images that compose the data set.

**Imaging Flow Cytometry:** A microscopy tool that combines flow cytometry with digital fluorescent microscopy. Samples are suspended in fluid and labelled with fluorescent markers. The fluorescent markers are excited by light which subsequently scatters to allow high-throughput data analysis to occur. Individual images of cells can be analysed by the researcher due to the flow cytometry aspect of the machine. This allows extra confidence and transparency to be attributed to the results of the data analysis.

**Machine Learning:** A branch of artificial intelligence that develops computer systems to learn, using algorithms and models to analyse patterns in a data set.

**MATLAB®:** Programming and coding software used to create the deep learning neural networks on which the ground truth data sets will be analysed in order to score unseen images in the *in vitro* MN assay and to thus calculate the dose response for test chemicals and compounds.

**Micronucleus:** Small DNA-containing nuclear structures, spatially isolated from the main nucleus that form when a whole or fragments of chromosomes are not incorporated into the main nucleus during mitosis. Usually $1/3^{rd}$ – $1/16^{th}$ the diameter of a regular nucleus, micronuclei are commonly used to assess the genotoxic potential of a test compound or chemical.

**Neural Network:** A series of algorithms that are all connected in order to recognise patterns in data sets. They mimic the function of the neuronal system in the human brain.

# References


Alberts, B., 2008. *Molecular biology of the cell*. 4th ed. New York: Garland Science.

Allemang, A., Thacker, R., DeMarco, R., Rodrigues, M., & Pfuhler, S. (2021). The 3D reconstructed skin micronucleus assay using imaging flow cytometry and deep learning: A proof-of-principle investigation. *Mutation Research/Genetic Toxicology And Environmental Mutagenesis*, *865*, 503314. https://doi.org/10.1016/j.mrgentox.2021.503314

Ames, B., Gurney, E., Miller, J., & Bartsch, H. (1972). Carcinogens as Frameshift Mutagens: Metabolites and Derivatives of 2-Acetylaminofluorene and Other Aromatic Amine Carcinogens. *Proceedings Of The National Academy Of Sciences*, *69*(11), 3128-3132. https://doi.org/10.1073/pnas.69.11.3128

Ames, B., Durston, W., Yamasaki, E., & Lee, F. (1973). Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proceedings Of The National Academy Of Sciences*, *70*(8), 2281-2285. https://doi.org/10.1073/pnas.70.8.2281

Ames, B., Shigenaga, M., & Hagen, T. (1993). Oxidants, antioxidants, and the degenerative diseases of aging. *Proceedings Of The National Academy Of Sciences*, *90*(17), 7915-7922. https://doi.org/10.1073/pnas.90.17.7915

Aranda, A., Bezunartea, J., Casales, E., Rodriguez-Madoz, J., Larrea, E., Prieto, J., & Smerdou, C. (2014). A quick and efficient method to generate mammalian stable cell lines based on a novel inducible alphavirus DNA/RNA layered system. *Cellular And Molecular Life Sciences*, *71*(23), 4637-4651. https://doi.org/10.1007/s00018-014-1631-2

Baan, R., Stewart, B., & Straif, K. (2019). *Tumour site concordance and mechanisms of carcinogenesis* (pp. 107-115). IARC Scientific Publications, 165

Beale, M., Hagan, M., & Demuth, H. (2020). *Deep Learning Toolbox User's Guide* (pp. 2 - 38). Natick: MathWorks.

Biolegend. (2021). *DRAQ5* (pp. 1-3). San Diego: Biolegend.

Bio-Rad. (2021). *Hoechst 33342 | Bio-Rad*. Bio-Rad. Retrieved 8 July 2021, from https://www.bio-rad-antibodies.com/cell-health-nuclear-staining-cell-pureblu-hoechst-33342.html.

Biostatus. (2021). *DRAQ5*. Biostatus.com. Retrieved 8 July 2021, from http://www.biostatus.com/DRAQ5/.

Bouchier-Hayes, L., McBride, S., van Geelen, C., Nance, S., Lewis, L., Pinkoski, M., & Beere, H. (2010). Fas ligand gene expression is directly regulated by stress-inducible heat shock transcription factor-1. *Cell Death & Differentiation*, *17*(6), 1034-1046. https://doi.org/10.1038/cdd.2010.4

Cheng, G., Sa, W., Cao, C., Guo, L., Hao, H., & Liu, Z. et al. (2016). Quinoxaline 1,4-di-N-Oxides: Biological Activities and Mechanisms of Actions. *Frontiers In Pharmacology*, *7*. https://doi.org/10.3389/fphar.2016.00064

Committee on Mutagenicity of Chemicals in Food, Consumer Products and the Environment (COM) (2011) Guidance on a strategy for genotoxicity testing of chemical substances.

Cruz, J., Minoja, G., & Okuchi, K. (2001). Improving Clinical Outcomes from Acute Subdural Hematomas with the Emergency Preoperative Administration of High Doses of Mannitol: A

Randomized Trial. *Neurosurgery*, *49*(4), 864-871. https://doi.org/10.1097/00006123-200110000-00016

De Bont, R. (2004). Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis*, *19*(3), 169-185. https://doi.org/10.1093/mutage/geh025

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers In Artificial Intelligence*, *3*. https://doi.org/10.3389/frai.2020.00004

Enzo Life Sciences. (2021). *Vinblastine . sulfate - ALX-350-257 - Enzo Life Sciences*. Enzolifesciences.com. Retrieved 5 July 2021, from https://www.enzolifesciences.com/ALX-350-257/vinblastine-.-sulfate/.

Farrell, A. (2014). *Encyclopedia of Fish Physiology* (pp. 2069-2077). Elsevier Science.

Fenech, M. (1993). The cytokinesis-block micronucleus technique and its application to genotoxicity studies in human populations. *Environmental Health Perspectives*, *101*(suppl 3), 101-107. https://doi.org/10.1289/ehp.93101s3101

Fenech, M. (1997). The advantages and disadvantages of the cytokinesis-block micronucleus method. *Mutation Research/Genetic Toxicology And Environmental Mutagenesis*, *392*(1-2), 11-18. https://doi.org/10.1016/s0165-1218(97)00041-4

Fenech, M. (2000). The in vitro micronucleus technique. *Mutation Research/Fundamental And Molecular Mechanisms Of Mutagenesis*, *455*(1-2), 81-95. https://doi.org/10.1016/s0027-5107(00)00065-8

Fisher Scientific. (2021). *Aroclor 1254 in n-Hexane 10ug/mL, Fisher Chemical*. Fisher Scientific. Retrieved 5 July 2021, from https://www.fishersci.no/shop/products/metabolite-aroclor-1254-n-hexane-10-g-ml/12969533.

Griffiths, A., Suzuki, D., Miller, J., Lewontin, R., & Gelbart, W. (2000). *An introduction to genetic analysis* (7th ed.). W.H. Freeman.

Hake, S., & Allis, C. (2006). Histone H3 variants and their potential role in indexing mammalian genomes: The "H3 barcode hypothesis". *Proceedings Of The National Academy Of Sciences*, *103*(17), 6428-6435. https://doi.org/10.1073/pnas.0600803103

Hanahan, D., & Weinberg, R. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646-674. https://doi.org/10.1016/j.cell.2011.02.013

International Agency for Research on Cancer. (2012). *A review of human carcinogens*. International Agency for Research on Cancer.

International Programme on Chemical Safety. (1993). *Carbendazim*. Geneva: World Health Organization.

Johnson, G., Soeteman-Hernández, L., Gollapudi, B., Bodger, O., Dearfield, K., & Heflich, R. et al. (2014). Derivation of point of departure (PoD) estimates in genetic toxicology studies and their potential applications in risk assessment. *Environmental And Molecular Mutagenesis*, *55*(8), 609-623. https://doi.org/10.1002/em.21870

Johnson, G (2020). Overview of Genotoxicity in Pharmacology. PM 316 Genetic Toxicology.

Kirkland, D., Kasper, P., Martus, H., Müller, L., van Benthem, J., Madia, F., & Corvi, R. (2016). Updated recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests. *Mutation Research/Genetic Toxicology And Environmental Mutagenesis*, *795*, 7-30. https://doi.org/10.1016/j.mrgentox.2015.10.006

Kirsch-Volders, M., & Fenech, M. (2001). Inclusion of micronuclei in non-divided mononuclear lymphocytes and necrosis/apoptosis may provide a more comprehensive cytokinesis block micronucleus assay for biomonitoring purposes. *Mutagenesis*, *16*(1), 51-58. https://doi.org/10.1093/mutage/16.1.51

Loeb, K., & Loeb, L. (2000). Significance of multiple mutations in cancer. *Carcinogenesis*, *21*(3), 379-385. https://doi.org/10.1093/carcin/21.3.379

Lovell, D. P., Fellows, M., Marchetti, F., Christiansen, J., Elhajouji, A., Hashimoto, K., et al. (2018). Analysis of negative historical control group data from the in vitro micronucleus assay using TK6 cells. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 825,* 40-50. doi:10.1016/j.mrgentox.2017.10.006Lu, S., Liao, J., Kuo, M., Wang, S., Hwang, J., & Ueng, T. (2004). Endocrine-disrupting activity in carbendazim-induced reproductive and developmental toxicity in rats. *Journal Of Toxicology And Environmental Health, Part A*, *67*(19), 1501-1515. https://doi.org/10.1080/15287390490486833 Lundin, C., North, M., Erixon, K., Walters, K., Jenssen, D., Goldman, A., & Helleday, T. (2005). Methyl methanesulfonate (MMS) produces heat-labile DNA damage but no detectable in vivo DNA double-strand breaks. *Nucleic Acids Research*, *33*(12), 3799-3811. https://doi.org/10.1093/nar/gki681

Luzhna, L., Kathiria, P., & Kovalchuk, O. (2013). Micronuclei in genotoxicity assessment: from genetics to epigenetics and beyond. *Frontiers In Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00131

Maass, K., Rosing, F., Ronchi, P., Willmund, K., Devens, F., & Hergt, M. et al. (2018). Altered nuclear envelope structure and proteasome function of micronuclei. *Experimental Cell Research*, *371*(2), 353-363. https://doi.org/10.1016/j.yexcr.2018.08.029

MacGregor, J., Casciano, D., & Müller, L. (2000). Strategies and testing methods for identifying mutagenic risks. *Mutation Research/Fundamental And Molecular Mechanisms Of Mutagenesis*, *455*(1-2), 3-20. https://doi.org/10.1016/s0027-5107(00)00116-0

Mah, L., El-Osta, A., & Karagiannis, T. (2010). γH2AX: a sensitive molecular marker of DNA damage and repair. *Leukemia*, *24*(4), 679-686. https://doi.org/10.1038/leu.2010.6

Malentacchi, F., Pizzamiglio, S., Wyrich, R., Verderio, P., Ciniselli, C., Pazzagli, M., & Gelmini, S. (2016). Effects of Transport and Storage Conditions on Gene Expression in Blood Samples. *Biopreservation And Biobanking*, *14*(2), 122-128. https://doi.org/10.1089/bio.2015.0037

Malling, H. (1966). Mutagenicity of two potent carcinogens, dimethylnitrosamine and diethylnitrosamine, in Neurospora crassa. *Mutation Research/Fundamental And Molecular Mechanisms Of Mutagenesis*, *3*(6), 537-540. https://doi.org/10.1016/0027-5107(66)90078-9

Malling, H. (1971). Dimethylnitrosamine: formation of mutagenic compounds by interaction with mouse liver microsomes☆. *Mutation Research/Fundamental And Molecular Mechanisms Of Mutagenesis*, *13*(4), 425-429. https://doi.org/10.1016/0027-5107(71)90054-6

2022. *MATLAB Operators and Special Characters*. [ebook] MathWorks. Available at: <https://uk.mathworks.com/help/matlab/matlab_prog/matlab-operators-and-special-characters.html> [Accessed 6 March 2022].

McDonald, J. (2009). *Handbook of biological statistics* (3rd ed., pp. 77-85). Sparky House Publishing.

MedChemExpress. (2021). *D-Mannitol (Mannitol) | Endogenous Metabolite | MedChemExpress*. MedchemExpress.com. Retrieved 5 July 2021, from https://www.medchemexpress.com/D-Mannitol.html.

Merck. (2021). *Methyl methanesulfonate - Methanesulfonic acid methyl ester, Methyl methanesulfonate*. Sigmaaldrich.com. Retrieved 5 July 2021, from https://www.sigmaaldrich.com/GB/en/substance/methylmethanesulfonate1101366273.

Merck. (2021). *Quinoxaline ≥95.0% | Sigma-Aldrich*. Sigmaaldrich.com. Retrieved 12 July 2021, from https://www.sigmaaldrich.com/GB/en/product/aldrich/22710.

Mhaidat, N., Alzoubi, K., Khabour, O., Alawneh, K., Raffee, L., & Alsatari, E. et al. (2016). Assessment of genotoxicity of vincristine, vinblastine and vinorelbine in human cultured lymphocytes: a comparative study. *Balkan Journal Of Medical Genetics*, *19*(1), 13-20. https://doi.org/10.1515/bjmg-2016-0002

Müller, L., Kikuchi, Y., Probst, G., Schechtman, L., Shimada, H., Sofuni, T., & Tweats, D. (1999). ICH-Harmonised guidances on genotoxicity testing of pharmaceuticals: evolution, reasoning and impact. *Mutation Research/Reviews In Mutation Research*, *436*(3), 195-225. https://doi.org/10.1016/s1383-5742(99)00004-6

Nath, J., & Krishna, G. (1998). Safety Screening of Drugs in Cancer Therapy. *Acta Haematologica*, *99*(3), 138-147. https://doi.org/10.1159/000040828

OECD. (2010). *OECD Guidance Document for the Design and Conduct of Chronic Toxicity and Carcinogenicity Studies, Supporting Tg 451, 452 And 453* (pp. 94 - 118). OECD.

OECD. (2016) OECD TG487 Guideline for the Testing of Chemicals, In Vitro Mammalian Cell Micronucleus Test. Organisation for Economic Cooperation and OECD, Paris.

Padalino, G., El-Sakkary, N., Liu, L., Liu, C., Harte, D., & Barnes, R. et al. (2021). Anti-schistosomal activities of quinoxaline-containing compounds: From hit identification to lead optimisation. *European Journal Of Medicinal Chemistry*, *226*, 113823. https://doi.org/10.1016/j.ejmech.2021.113823

Palero, F., & Crandall, K. (2009). Phylogenetic Inference Using Molecular Data. *Decapod Crustacean Phylogenetics*, 67-88. https://doi.org/10.1201/9781420092592-c5

Parsons, B. (2008). Many different tumor types have polyclonal tumor origin: Evidence and implications. *Mutation Research/Reviews In Mutation Research*, *659*(3), 232-247. https://doi.org/10.1016/j.mrrev.2008.05.004

Peltomäki, P. (2012). Mutations and epimutations in the origin of cancer. *Experimental Cell Research*, *318*(4), 299-310. https://doi.org/10.1016/j.yexcr.2011.12.001

Perri, F., Pisconti, S., & Della Vittoria Scarpati, G. (2016). P53 mutations and cancer: a tight linkage. *Annals Of Translational Medicine*, *4*(24), 522-522. https://doi.org/10.21037/atm.2016.12.40

Pfuhler, S., Fautz, R., Ouedraogo, G., Latil, A., Kenny, J., & Moore, C. et al. (2014). The Cosmetics Europe strategy for animal-free genotoxicity testing: Project status up-date. *Toxicology In Vitro*, *28*(1), 18-23. https://doi.org/10.1016/j.tiv.2013.06.004

Pray, L. (2008). DNA Replication and Causes of Mutation. *Nature Education*, *1*(1), 214.

Rees, B., Tate, M., Lynch, A., Thornton, C., Jenkins, G., Walmsley, R., & Johnson, G. (2017). Development of an in vitro PIG-A gene mutation assay in human cells. *Mutagenesis*. https://doi.org/10.1093/mutage/gew059

Ren, N., Atyah, M., Chen, W., & Zhou, C. (2017). The various aspects of genetic and epigenetic toxicology: testing methods and clinical applications. *Journal Of Translational Medicine*, *15*(1). https://doi.org/10.1186/s12967-017-1218-4

Rodrigues, M. (2018). Automation of the in vitro micronucleus assay using the Imagestream® imaging flow cytometer. *Cytometry Part A*, *93*(7), 706-726. https://doi.org/10.1002/cyto.a.23493

Rodrigues, M., Probst, C., Zayats, A., Davidson, B., Riedel, M., Li, Y., & Venkatachalam, V. (2021). The in vitro micronucleus assay using imaging flow cytometry and deep learning. *Npj Systems Biology And Applications*, *7*(1). https://doi.org/10.1038/s41540-021-00179-5

Sand, S., Victorin, K., & Filipsson, A. (2008). The current state of knowledge on the use of the benchmark dose concept in risk assessment. *Journal Of Applied Toxicology*, *28*(4), 405-421. https://doi.org/10.1002/jat.1298

Seager, A., Shah, U., Brusehafer, K., Wills, J., Manshian, B., & Chapman, K. et al. (2014). Recommendations, evaluation and validation of a semi-automated, fluorescent-based scoring protocol for micronucleus testing in human cells. *Mutagenesis*, *29*(3), 155-164. https://doi.org/10.1093/mutage/geu008

Sheu, Y. (2020). Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers In Psychiatry*, *11*. https://doi.org/10.3389/fpsyt.2020.551299

Smith, G., & Grisham, J. (1983). Cytotoxicity of monofunctional alkylating agents methyl methanesulfonate and methyl-N'-nitro-N-nitrosoguanidine have different mechanisms of toxicity for 10T12 cells. *Mutation Research/Fundamental And Molecular Mechanisms Of Mutagenesis*, *111*(3), 405-417. https://doi.org/10.1016/0027-5107(83)90036-2

Subramanian, D., Huang, J., Sevugan, M., Robinson, R., Balasubramanian, M., & Tang, X. (2013). Insight into Actin Organization and Function in Cytokinesis from Analysis of Fission Yeast Mutants. *Genetics*, *194*(2), 435-446. https://doi.org/10.1534/genetics.113.149716

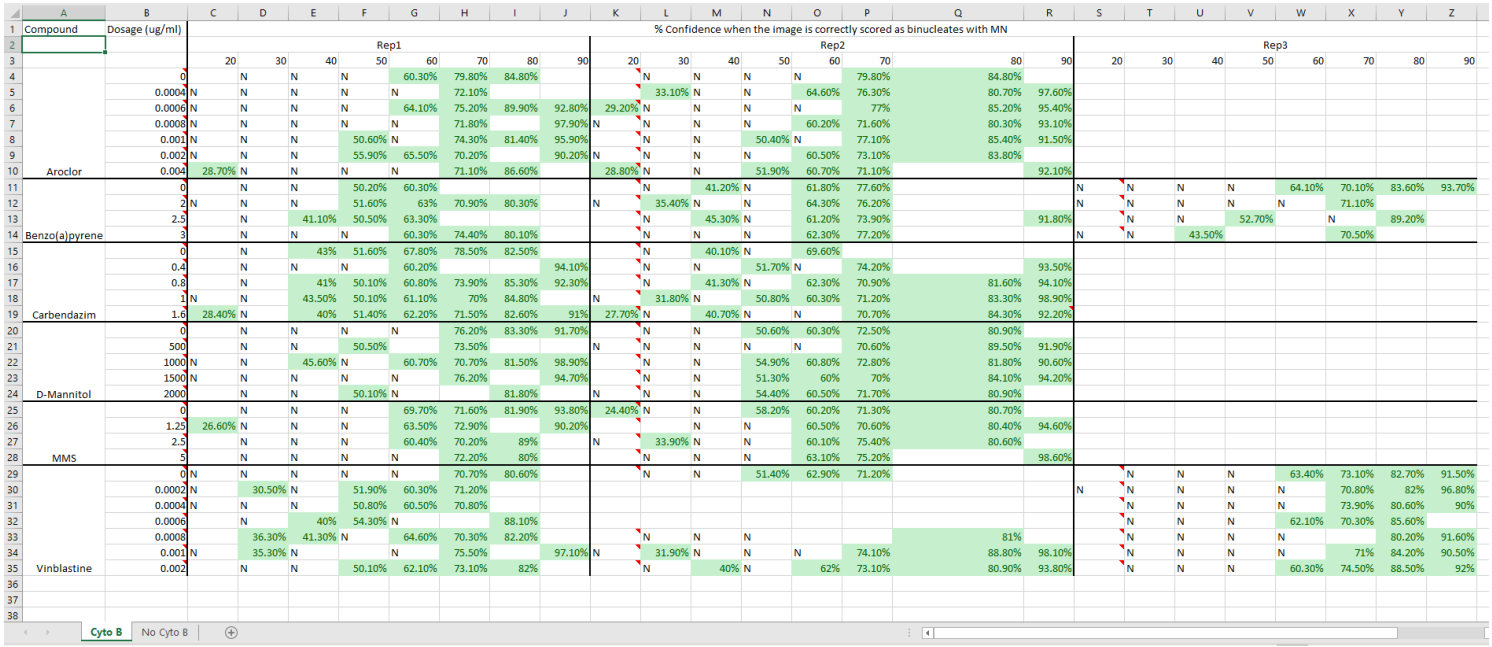
ThermoFisher Scientific. (2021). *Hoescht 33342, Trihydrochloride, Trihydrate - 10 mg/mL Solution in Water*. Retrieved 8 July 2021, from https://www.thermofisher.com/order/catalog/product/H3570#/H3570.

Toronto Research Chemicals. (2021). *Benzopyrene B205802*. Toronto Research Chemicals. Retrieved 5 July 2021, from https://www.trc-canada.com/product-detail/?B205802.

Verma, J., Rees, B., Wilde, E., Thornton, C., Jenkins, G., Doak, S., & Johnson, G. (2016). Evaluation of the automated MicroFlow® and Metafer™ platforms for high-throughput micronucleus scoring and dose response analysis in human lymphoblastoid TK6 cells. *Archives Of Toxicology*, *91*(7), 2689-2698. https://doi.org/10.1007/s00204-016-1903-8

Wills, J., Verma, J., Rees, B., Harte, D., Haxhiraj, Q., & Barnes, C. et al. (2021). Inter-laboratory automation of the in vitro micronucleus assay using imaging flow cytometry and deep learning. *Archives Of Toxicology*, *95*(9), 3101-3115. https://doi.org/10.1007/s00204-021-03113-0

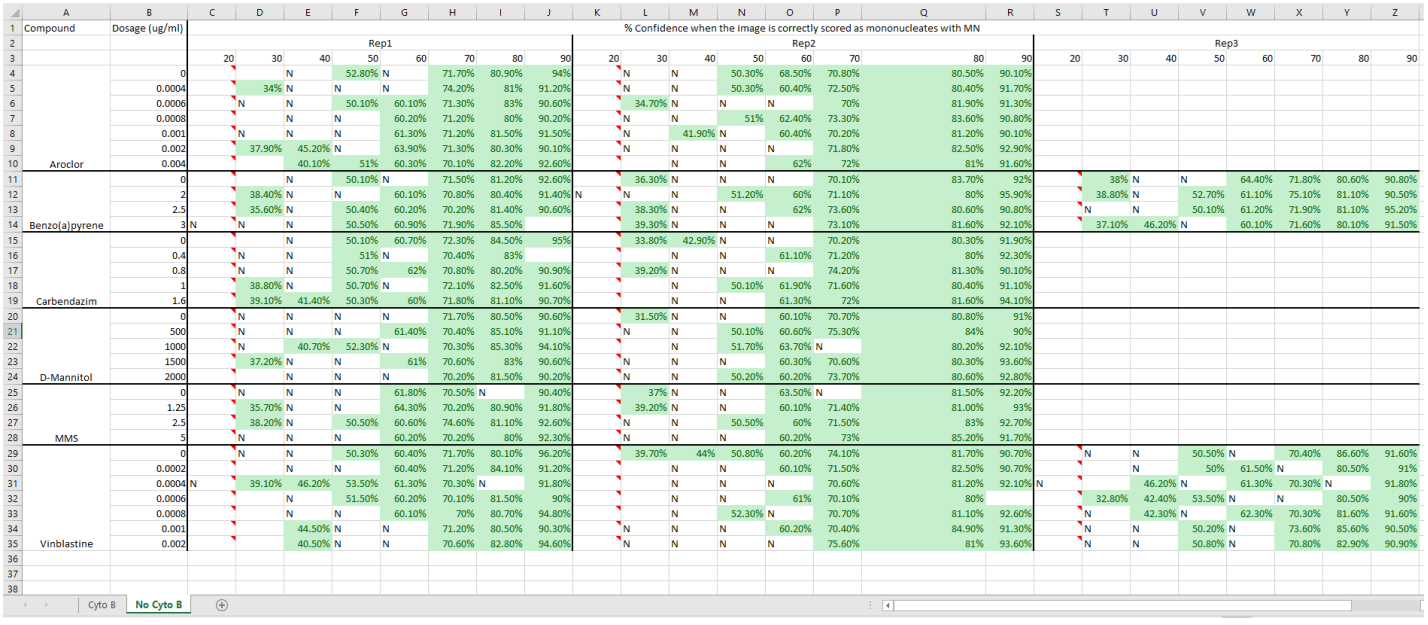
# Appendix 1



| Compound | Dosage (ug/ml) | % Confidence when the image is correctly scored as binucleates with MN | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rep1 | | | | | | | | | Rep2 | | | | | | | | Rep3 | | | | | | | | | | |
| | | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | | | | |
| | 0 | | N | N | N | 60.30% | 79.80% | 84.80% | | | N | N | N | N | 79.80% | 84.80% | | | | | | | | | | | | | |
| | 0.0004 | N | N | N | N | N | 72.10% | | | | 33.10% | N | N | 64.60% | 76.30% | 80.70% | 97.60% | | | | | | | | | | | | |
| | 0.0006 | N | N | N | N | 64.10% | 75.20% | 89.90% | 92.80% | 29.20% | N | N | N | N | 77% | 85.20% | 95.40% | | | | | | | | | | | | |
| | 0.0008 | N | N | N | N | N | 71.80% | | 97.90% | N | N | N | N | 60.20% | 71.60% | 80.30% | 93.10% | | | | | | | | | | | | |
| | 0.001 | N | N | N | 50.60% | N | 74.30% | 81.40% | 95.90% | | N | N | 50.40% | N | 77.10% | 85.40% | 91.50% | | | | | | | | | | | | |
| | 0.002 | N | N | N | 55.90% | 65.50% | 70.20% | | 90.20% | N | N | N | N | 60.50% | 73.10% | 83.80% | | | | | | | | | | | | | |
| Aroclor | 0.004 | 28.70% | N | N | N | N | 71.10% | 86.60% | | 28.80% | N | N | 51.90% | 60.70% | 71.10% | | 92.10% | | | | | | | | | | | | |
| | 0 | | N | N | 50.20% | 60.30% | | | | | N | 41.20% | N | 61.80% | 77.60% | | | N | N | N | N | 64.10% | 70.10% | 83.60% | 93.70% | | | | |
| | 2 | N | N | N | 51.60% | 63% | 70.90% | 80.30% | | N | 35.40% | N | N | 64.30% | 76.20% | | | N | N | N | N | N | 71.10% | | | | | | | |
| | 2.5 | | N | 41.10% | 50.50% | 63.30% | | | | | N | 45.30% | N | 61.20% | 73.90% | | 91.80% | | N | N | 52.70% | | N | 89.20% | | | | | |
| Benzo(a)pyrene | 3 | | N | N | N | 60.30% | 74.40% | 80.10% | | | N | N | N | 62.30% | 77.20% | | | N | N | 43.50% | | | 70.50% | | | | | | | |
| | 0 | | N | 43% | 51.60% | 67.80% | 78.50% | 82.50% | | | N | 40.10% | N | 69.60% | | | | | | | | | | | | | | | |
| | 0.4 | | N | N | N | 60.20% | | | 94.10% | | N | N | 51.70% | N | 74.20% | | 93.50% | | | | | | | | | | | | |
| | 0.8 | | N | 41% | 50.10% | 60.80% | 73.90% | 85.30% | 92.30% | | N | 41.30% | N | 62.30% | 70.90% | 81.60% | 94.10% | | | | | | | | | | | | |
| | 1 | N | N | 43.50% | 50.10% | 61.10% | 70% | 84.80% | | N | 31.80% | N | 50.80% | 60.30% | 71.20% | 83.30% | 98.90% | | | | | | | | | | | | |
| Carbendazim | 1.6 | 28.40% | N | 40% | 51.40% | 62.20% | 71.50% | 82.60% | 91% | 27.70% | N | 40.70% | N | N | 70.70% | 84.30% | 92.20% | | | | | | | | | | | | |
| | 0 | | N | N | N | N | 76.20% | 83.30% | 91.70% | | N | N | 50.60% | 60.30% | 72.50% | 80.90% | | | | | | | | | | | | | |
| | 500 | | N | N | 50.50% | | 73.50% | | | N | N | N | N | N | 70.60% | 89.50% | 91.90% | | | | | | | | | | | | |
| | 1000 | N | N | 45.60% | N | 60.70% | 70.70% | 81.50% | 98.90% | | N | N | 54.90% | 60.80% | 72.80% | 81.80% | 90.60% | | | | | | | | | | | | |
| | 1500 | N | N | N | N | N | 76.20% | | 94.70% | | N | N | 51.30% | 60% | 70% | 84.10% | 94.20% | | | | | | | | | | | | |
| D-Mannitol | 2000 | | N | N | 50.10% | N | | 81.80% | | N | N | N | 54.40% | 60.50% | 71.70% | 80.90% | | | | | | | | | | | | | |
| | 0 | | N | N | N | 69.70% | 71.60% | 81.90% | 93.80% | 24.40% | N | N | 58.20% | 60.20% | 71.30% | 80.70% | | | | | | | | | | | | | |
| | 1.25 | 26.60% | N | N | N | 63.50% | 72.90% | | 90.20% | | | N | N | 60.50% | 70.60% | 80.40% | 94.60% | | | | | | | | | | | | |
| | 2.5 | | N | N | N | 60.40% | 70.20% | 89% | | N | 33.90% | N | N | 60.10% | 75.40% | 80.60% | | | | | | | | | | | | | |
| MMS | 5 | | N | N | N | N | 72.20% | 80% | | | N | N | N | 63.10% | 75.20% | | 98.60% | | | | | | | | | | | | |
| | 0 | N | N | N | N | N | 70.70% | 80.60% | | | N | N | 51.40% | 62.90% | 71.20% | | | | N | N | N | 63.40% | 73.10% | 82.70% | 91.50% | | | | |
| | 0.0002 | N | 30.50% | N | 51.90% | 60.30% | 71.20% | | | | | | | | | | | N | N | N | N | N | 70.80% | 82% | 96.80% | | | | |
| | 0.0004 | N | N | N | 50.80% | 60.50% | 70.80% | | | | | | | | | | | | N | N | N | N | 73.90% | 80.60% | 90% | | | | |
| | 0.0006 | | N | 40% | 54.30% | N | | 88.10% | | | | | | | | | | | N | N | N | 62.10% | 70.30% | 85.60% | | | | | |
| | 0.0008 | | 36.30% | 41.30% | N | 64.60% | 70.30% | 82.20% | | | N | N | N | | | 81% | | | N | N | N | N | | 80.20% | 91.60% | | | | |
| | 0.001 | N | 35.30% | N | | N | 75.50% | | 97.10% | N | 31.90% | N | N | N | 74.10% | 88.80% | 98.10% | | N | N | N | N | 71% | 84.20% | 90.50% | | | | |
| Vinblastine | 0.002 | | N | N | 50.10% | 62.10% | 73.10% | 82% | | | N | 40% | N | 62% | 73.10% | 80.90% | 93.80% | | N | N | N | 60.30% | 74.50% | 88.50% | 92% | | | | |

Analysis of the percentage confidence of the deep learning algorithm in the CBMN assay. Green indicates at what percentage confidence levels the algorithm consistently scores the images correctly.

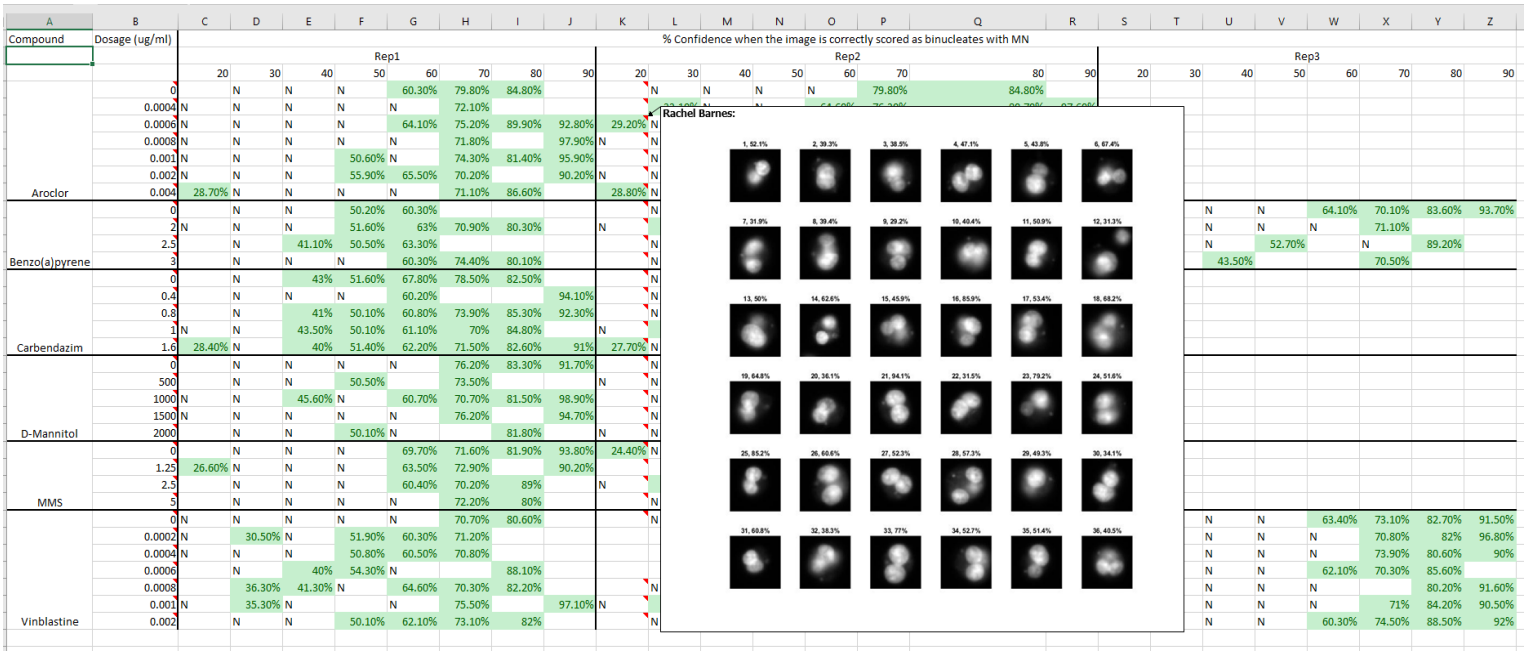
# Appendix 2



| Compound | Dosage (ug/ml) | % Confidence when the image is correctly scored as mononucleates with MN | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Rep1 | | | | | | | | | Rep2 | | | | | | | | Rep3 | | | | | | | | | | |
| | | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | | | | |
| | 0 | | N | | 52.80% | N | 71.70% | 80.90% | 94% | | N | N | 50.30% | 68.50% | 70.80% | 80.50% | 90.10% | | | | | | | | | | | | |
| | 0.0004 | | 34% | N | N | N | 74.20% | 81% | 91.20% | | N | N | 50.30% | 60.40% | 72.50% | 80.40% | 91.70% | | | | | | | | | | | | |
| | 0.0006 | | N | N | 50.10% | 60.10% | 71.30% | 83% | 90.60% | | 34.70% | N | N | N | 70% | 81.90% | 91.30% | | | | | | | | | | | | |
| | 0.0008 | | | N | N | 60.20% | 71.20% | 80% | 90.20% | | N | N | 51% | 62.40% | 73.30% | 83.60% | 90.80% | | | | | | | | | | | | |
| | 0.001 | | N | N | N | 61.30% | 71.20% | 81.50% | 91.50% | | N | 41.90% | N | 60.40% | 70.20% | 81.20% | 90.10% | | | | | | | | | | | | |
| | 0.002 | | 37.90% | 45.20% | N | 63.90% | 71.30% | 80.30% | 90.10% | | N | N | N | N | 71.80% | 82.50% | 92.90% | | | | | | | | | | | | |
| Aroclor | 0.004 | | | 40.10% | 51% | 60.30% | 70.10% | 82.20% | 92.60% | | | N | N | 62% | 72% | 81% | 91.60% | | | | | | | | | | | | |
| | 0 | | | N | 50.10% | N | 71.50% | 81.20% | 92.60% | | 36.30% | N | N | N | 70.10% | 83.70% | 92% | | 38% | N | N | 64.40% | 71.80% | 80.60% | 90.80% | | | | |
| | 2 | | 38.40% | N | N | 60.10% | 70.80% | 80.40% | 91.40% | N | N | N | 51.20% | 60% | 71.10% | 80% | 95.90% | | 38.80% | N | 52.70% | 61.10% | 75.10% | 81.10% | 90.50% | | | | |
| | 2.5 | | 35.60% | N | 50.40% | 60.20% | 70.20% | 81.40% | 90.60% | | 38.30% | N | N | 62% | 73.60% | 80.60% | 90.80% | N | N | 50.10% | 61.20% | 71.90% | 81.10% | 95.20% | | | | |
| Benzo(a)pyrene | 3 | N | N | N | 50.50% | 60.90% | 71.90% | 85.50% | | | 39.30% | N | N | N | 73.10% | 81.60% | 92.10% | | 37.10% | 46.20% | N | 60.10% | 71.60% | 80.10% | 91.50% | | | | |
| | 0 | | | N | 50.10% | 60.70% | 72.30% | 84.50% | 95% | | 33.80% | 42.90% | N | N | 70.20% | 80.30% | 91.90% | | | | | | | | | | | | |
| | 0.4 | | N | N | 51% | N | 70.40% | 83% | | | | N | N | 61.10% | 71.20% | 80% | 92.30% | | | | | | | | | | | | |
| | 0.8 | | N | N | 50.70% | 62% | 70.80% | 80.20% | 90.90% | | 39.20% | N | N | N | 74.20% | 81.30% | 90.10% | | | | | | | | | | | | |
| | 1 | | 38.80% | N | 50.70% | N | 72.10% | 82.50% | 91.60% | | | N | 50.10% | 61.90% | 71.60% | 80.40% | 91.10% | | | | | | | | | | | | |
| Carbendazim | 1.6 | | 39.10% | 41.40% | 50.30% | 60% | 71.80% | 81.10% | 90.70% | | | N | N | 61.30% | 72% | 81.60% | 94.10% | | | | | | | | | | | | |
| | 0 | | N | N | N | N | 71.70% | 80.50% | 90.60% | | 31.50% | N | N | 60.10% | 70.70% | 80.80% | 91% | | | | | | | | | | | | |
| | 500 | | N | N | N | 61.40% | 70.40% | 85.10% | 91.10% | | N | N | 50.10% | 60.60% | 75.30% | 84% | 90% | | | | | | | | | | | | |
| | 1000 | | N | 40.70% | 52.30% | N | 70.30% | 85.30% | 94.10% | | | N | 51.70% | 63.70% | N | 80.20% | 92.10% | | | | | | | | | | | | |
| | 1500 | | 37.20% | N | N | 61% | 70.60% | 83% | 90.60% | | N | N | N | 60.30% | 70.60% | 80.30% | 93.60% | | | | | | | | | | | | |
| D-Mannitol | 2000 | | | N | N | N | 70.20% | 81.50% | 90.20% | | N | N | 50.20% | 60.20% | 73.70% | 80.60% | 92.80% | | | | | | | | | | | | |
| | 0 | | N | N | N | 61.80% | 70.50% | N | 90.40% | | 37% | N | N | 63.50% | N | 81.50% | 92.20% | | | | | | | | | | | | |
| | 1.25 | | 35.70% | N | N | 64.30% | 70.20% | 80.90% | 91.80% | | 39.20% | N | N | 60.10% | 71.40% | 81.00% | 93% | | | | | | | | | | | | |
| | 2.5 | | 38.20% | N | 50.50% | 60.60% | 74.60% | 81.10% | 92.60% | | N | N | 50.50% | 60% | 71.50% | 83% | 92.70% | | | | | | | | | | | | |
| MMS | 5 | | N | N | N | 60.20% | 70.20% | 80% | 92.30% | | N | N | N | 60.20% | 73% | 85.20% | 91.70% | | | | | | | | | | | | |
| | 0 | | N | N | 50.30% | 60.40% | 71.70% | 80.10% | 96.20% | | 39.70% | 44% | 50.80% | 60.20% | 74.10% | 81.70% | 90.70% | | N | N | 50.50% | N | 70.40% | 86.60% | 91.60% | | | | |
| | 0.0002 | | | N | N | 60.40% | 71.20% | 84.10% | 91.20% | | | N | N | 60.10% | 71.50% | 82.50% | 90.70% | | | N | 50% | 61.50% | N | 80.50% | 91% | | | | |
| | 0.0004 | N | 39.10% | 46.20% | 53.50% | 61.30% | 70.30% | N | 91.80% | | | N | N | N | 70.60% | 81.20% | 92.10% | N | | 46.20% | N | 61.30% | 70.30% | N | 91.80% | | | | |
| | 0.0006 | | | N | 51.50% | 60.20% | 70.10% | 81.50% | 90% | | | N | N | 61% | 70.10% | 80% | | | 32.80% | 42.40% | 53.50% | N | N | 80.50% | 90% | | | | |
| | 0.0008 | | | N | N | 60.10% | 70% | 80.70% | 94.80% | | | N | 52.30% | N | 70.70% | 81.10% | 92.60% | | N | 42.30% | N | 62.30% | 70.30% | 81.60% | 91.60% | | | | |
| | 0.001 | | | 44.50% | N | N | 71.20% | 80.50% | 90.30% | | N | N | N | 60.20% | 70.40% | 84.90% | 91.30% | | N | N | 50.20% | N | 73.60% | 85.60% | 90.50% | | | | |
| Vinblastine | 0.002 | | | 40.50% | N | N | 70.60% | 82.80% | 94.60% | | N | N | N | N | 75.60% | 81% | 93.60% | | N | N | 50.80% | N | 70.80% | 82.90% | 90.90% | | | | |

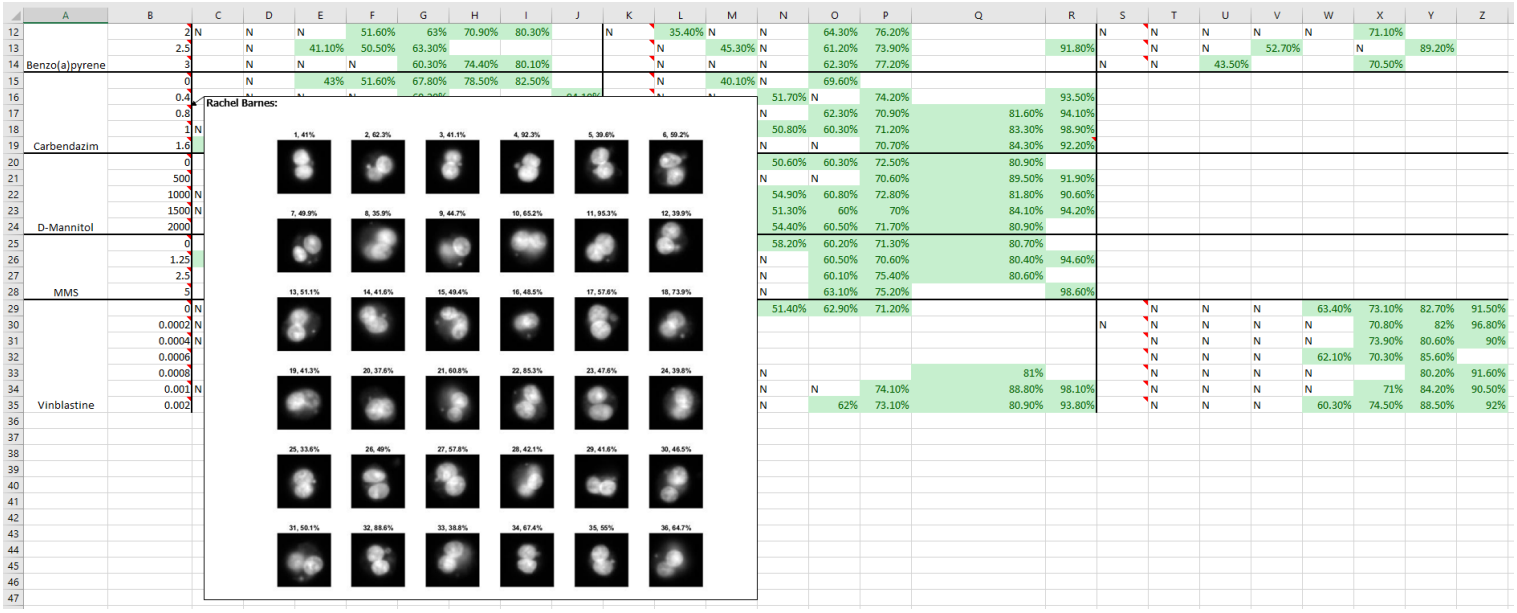
Analysis of the percentage confidence of the deep learning algorithm in the non-cyto b assay. Green indicates at what percentage confidence levels the algorithm consistently scores the images correctly.

# Appendix 3




Sample of 36 images that the neural network scored as binucleated with MN from the cytochalasin B experiment with 0.0006 µg/mL Aroclor, second repetition.

# Appendix 4



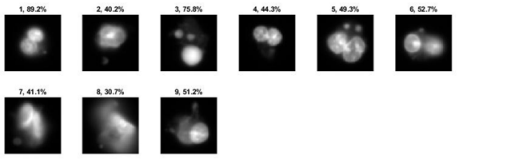

Sample of 36 images that the neural network scored as binucleated with MN from the cytochalasin B experiment with 0.8 µg/mL Carbendazim, first repetition.

# Appendix 5

% Confidence when the image is correctly scored as binucleates with MN

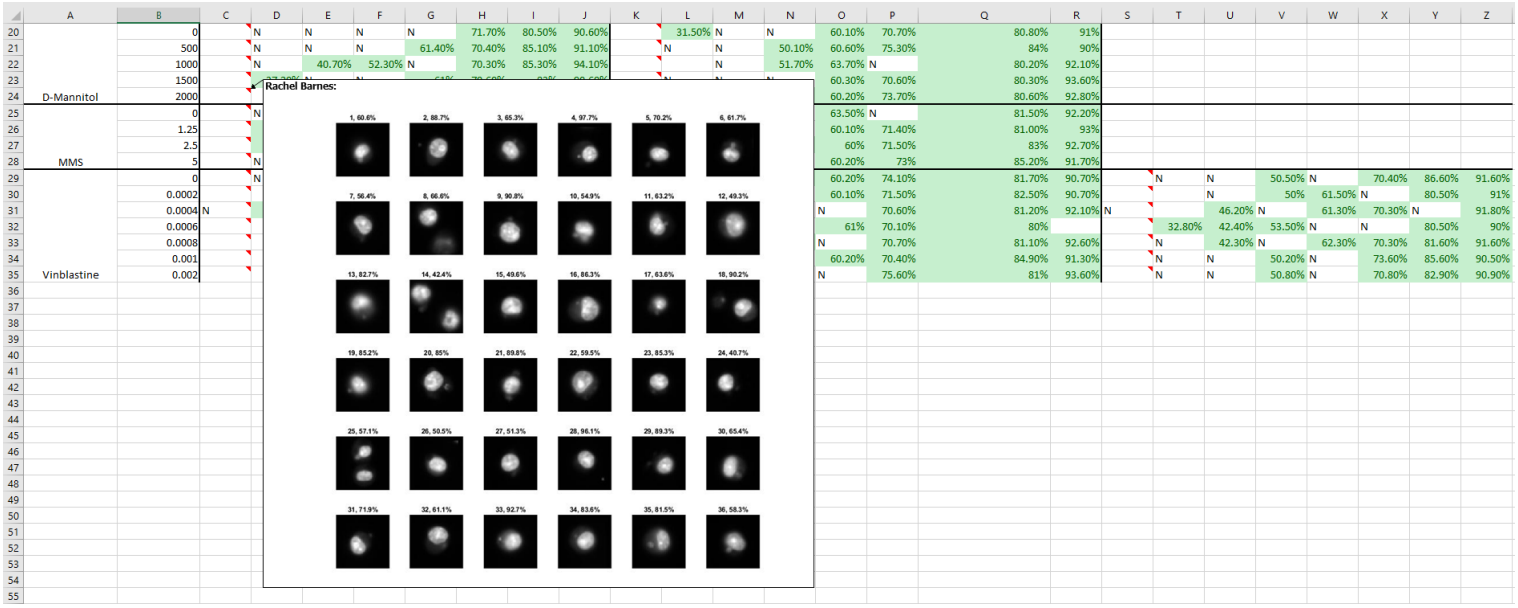
| Rep1 | | | | | | | Rep2 | | | | | | | | Rep3 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 40 | 50 | 60 | 70 | 80 | 90 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
| N | N | N | 60.30% | 79.80% | 84.80% | | | N | N | N | N | 79.80% | 84.80% | | | | | | | | | |
| N | N | N | N | 72.10% | | | | 33.10% | N | N | 64.60% | 76.30% | 80.70% | 97.60% | | | | | | | | |
| N | N | N | 64.10% | 75.20% | 89.90% | 92.80% | 29.20% | N | N | N | N | 77% | 85.20% | 95.40% | | | | | | | | |
| N | N | N | N | 71.80% | | 97.90% | N | N | N | N | 60.20% | 71.60% | 80.30% | 93.10% | | | | | | | | |
| N | N | 50.60% | N | 74.30% | 81.40% | 95.90% | | N | N | 50.40% | N | 77.10% | 85.40% | 91.50% | | | | | | | | |
| N | N | 55.90% | 65.50% | 70.20% | | 90.20% | N | N | N | N | 60.50% | 73.10% | 83.80% | | | | | | | | | |
| N | N | N | N | 71.10% | 86.60% | | 28.80% | N | N | 51.90% | 60.70% | 71.10% | | 92.10% | | | | | | | | |
| N | N | 50.20% | 60.30% | | | | | N | 41.20% | N | 61.80% | 77.60% | | | N | N | N | N | 64.10% | 70.10% | 83.60% | 93.70% |
| N | N | 51.60% | 63% | 70.90% | 80.30% | | N | 35.40% | N | N | 64.30% | 76.20% | | | N | N | | | | | | |
| N | 41.10% | 50.50% | 63.30% | | | | | N | 45.30% | N | 61.20% | 73.90% | | 91.80% | | N | | | | | | |
| N | N | N | 60.30% | 74.40% | 80.10% | | | N | N | N | 62.30% | 77.20% | | | N | N | | | | | | |
| N | 43% | 51.60% | 67.80% | 78.50% | 82.50% | | | N | 40.10% | N | 69.60% | | | | | | | | | | | |
| N | N | N | 60.20% | | | 94.10% | | N | N | 51.70% | N | 74.20% | | 93.50% | | | | | | | | |
| N | 41% | 50.10% | 60.80% | 73.90% | 85.30% | 92.30% | | N | 41.30% | N | 62.30% | 70.90% | 81.60% | 94.10% | | | | | | | | |
| N | 43.50% | 50.10% | 61.10% | 70% | 84.80% | | N | 31.80% | N | 50.80% | 60.30% | 71.20% | 83.30% | 98.90% | | | | | | | | |
| N | 40% | 51.40% | 62.20% | 71.50% | 82.60% | 91% | 27.70% | N | 40.70% | N | N | 70.70% | 84.30% | 92.20% | | | | | | | | |
| N | N | N | N | 76.20% | 83.30% | 91.70% | | N | N | 50.60% | 60.30% | 72.50% | 80.90% | | | | | | | | | |
| N | N | 50.50% | | 73.50% | | | N | N | N | N | N | 70.60% | 89.50% | 91.90% | | | | | | | | |
| N | 45.60% | N | 60.70% | 70.70% | 81.50% | 98.90% | | N | N | 54.90% | 60.80% | 72.80% | 81.80% | 90.60% | | | | | | | | |
| N | N | N | N | 76.20% | | 94.70% | | N | N | 51.30% | 60% | 70% | 84.10% | 94.20% | | | | | | | | |
| N | N | 50.10% | N | | 81.80% | | N | N | N | 54.40% | 60.50% | 71.70% | 80.90% | | | | | | | | | |
| N | N | N | 69.70% | 71.60% | 81.90% | 93.80% | 24.40% | N | N | 58.20% | 60.20% | 71.30% | 80.70% | | | | | | | | | |
| N | N | N | 63.50% | 72.90% | | 90.20% | | | N | N | 60.50% | 70.60% | 80.40% | 94.60% | | | | | | | | |
| N | N | N | 60.40% | 70.20% | 89% | | N | 33.90% | N | N | 60.10% | 75.40% | 80.60% | | | | | | | | | |
| N | N | N | N | 72.20% | 80% | | | N | N | N | 63.10% | 75.20% | | 98.60% | | | | | | | | |
| N | N | N | N | 70.70% | 80.60% | | | N | N | 51.40% | 62.90% | 71.20% | | | | N | | | | | | |
| 30.50% | N | 51.90% | 60.30% | 71.20% | | | | | | | | | | | N | N | | | | | | |
| N | N | 50.80% | 60.50% | 70.80% | | | | | | | | | | | | N | N | N | N | 73.90% | 80.60% | 90% |
| N | 40% | 54.30% | N | | 88.10% | | | | | | | | | | | N | N | N | 62.10% | 70.30% | 85.60% | |
| 36.30% | 41.30% | N | 64.60% | 70.30% | 82.20% | | | N | N | N | | | 81% | | | N | N | N | N | | 80.20% | 91.60% |
| 35.30% | N | | N | 75.50% | | 97.10% | N | 31.90% | N | N | N | 74.10% | 88.80% | 98.10% | | N | N | N | N | 71% | 84.20% | 90.50% |
| N | N | 50.10% | 62.10% | 73.10% | 82% | | | N | 40% | N | 62% | 73.10% | 80.90% | 93.80% | | N | N | N | 60.30% | 74.50% | 88.50% | 92% |

Rachel Barnes:



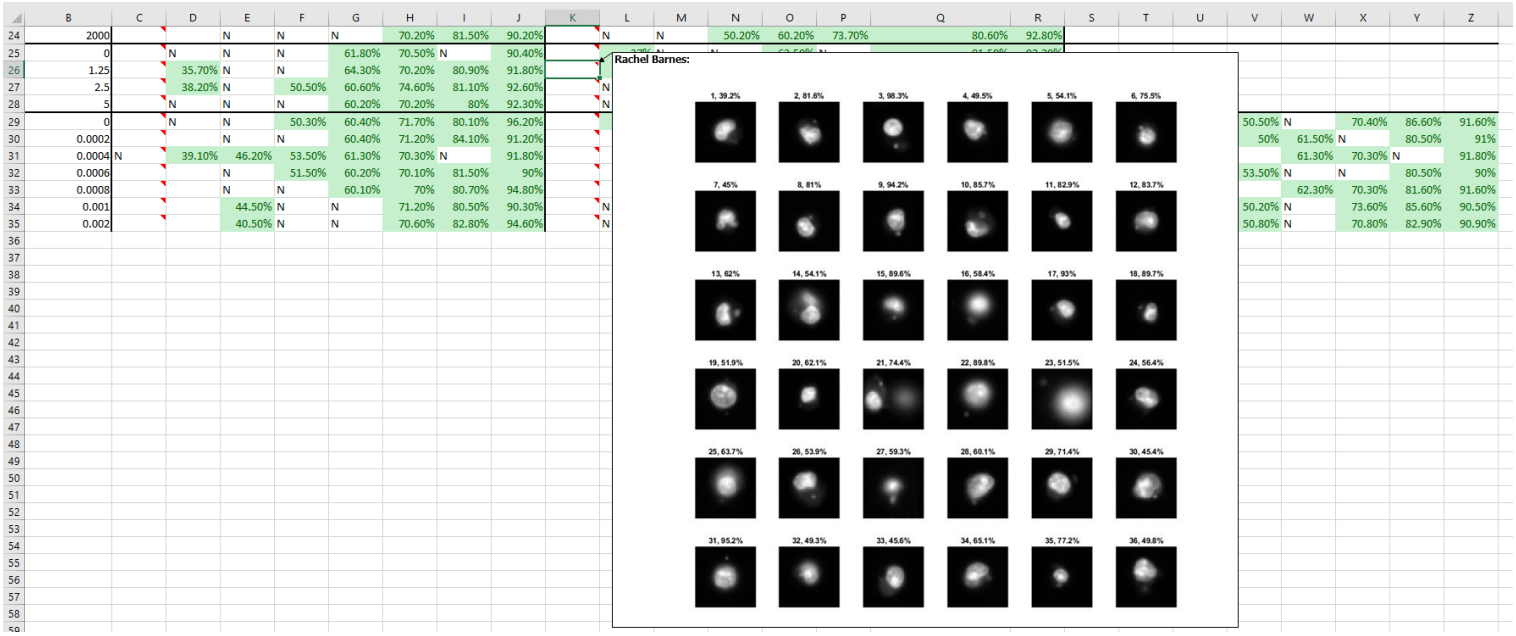

Sample of 9 images that the neural network scored as binucleated with MN from the cytochalasin B experiment with 2.5 µg/mL Benzo(a)pyrene, third repetition.

# Appendix 6



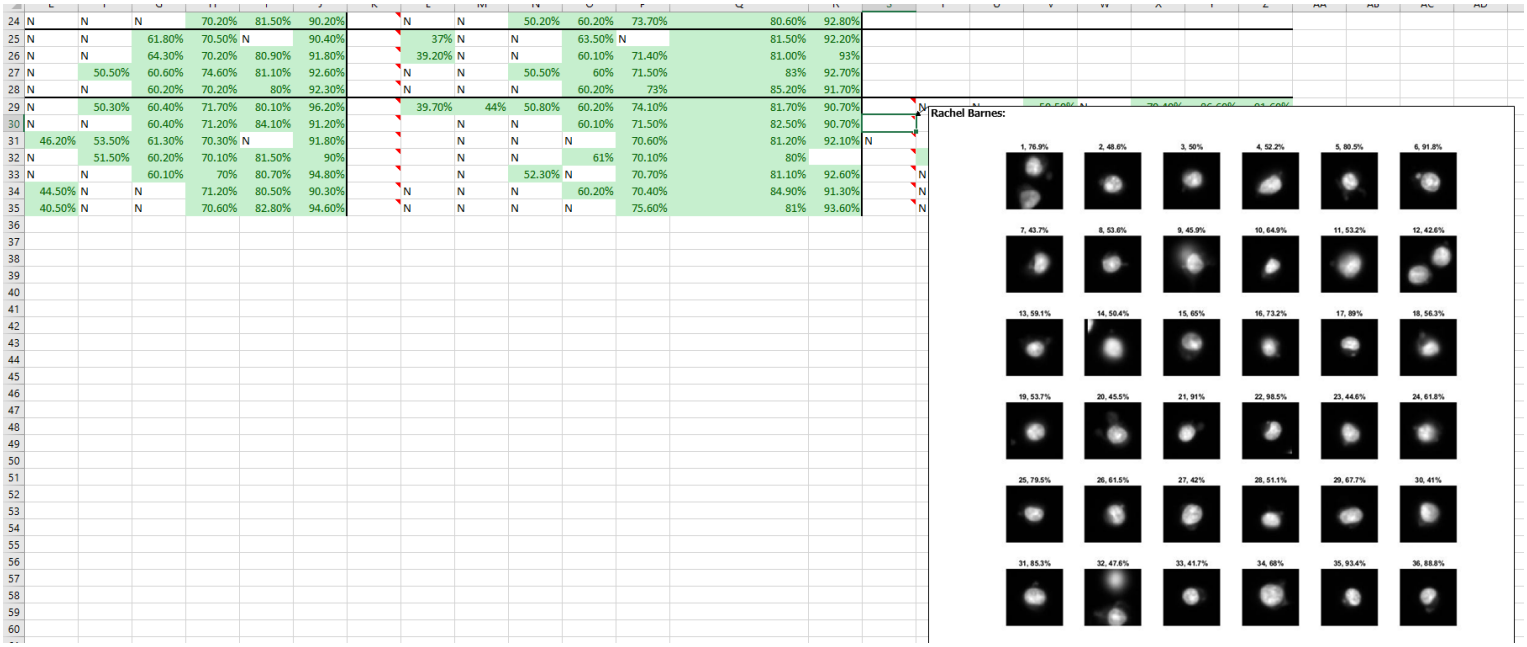

Sample of 36 images that the neural network scored as mononucleated with MN from the mononucleate experiment with 2000 µg/mL D-Mannitol, first repetition. This appendix includes an exception to the 70% cut-off rule. The third image in the sample was classified by the neural network as a binucleated cell with an MN with 75.8% confidence however, after manual visualisation, the image is manually scored as a mononucleated cell with two MN, therefore the algorithm has incorrectly scored the image. The eighth image also provides an example of where misclassifications of phenotypes occur due to the presence of debris in the image. In this image, the cells appear very faint and fuzzy therefore it is very difficult to manually score the phenotype of the image.

# Appendix 7




Sample of 36 images that the neural network scored as mononucleated with MN from the mononucleate experiment with 1.25 µg/mL MMS, second repetition.

# Appendix 8




Sample of 36 images that the neural network scored as mononucleated with MN from the mononucleate experiment with 0.0002 µg/mL vinblastine, third repetition.

This appendix includes another exception to the 70% cut-off rule. The first image in the sample was classified by the neural network as a mononucleated cell with an MN with 76.9% confidence however, after manual visualisation, the image is manually scored as a binucleated cell with an MN, therefore the algorithm has incorrectly scored the image.