THE UNIVERSITY *of* EDINBURGH

Discovery of Novel Microglial Homeostasis Modulators Through Machine Learning

Allen Shaw

THE UNIVERSITY *of* EDINBURGH

A thesis presented for the degree of Master of Science by Research

Centre for Discovery Brain Sciences
University of Edinburgh

December 2024

Declaration

I declare that all work presented in this thesis is my own, except where otherwise indicated. This work has not been submitted for any other degree or professional qualification.

– Allen Shaw
December 12, 2024

Acknowledgements

I would like to acknowledge Dr. Jing Qiu for proposing this project, Dr. Roderick Carter for conducting the original screening that provided the data for my model, Tom Leah for teaching me all the lab techniques, and Dr. Daga Panas for her constant guidance and support throughout my machine-learning journey, starting from when I knew nothing about ML.

Table of Contents

Abstract

Identifying molecules capable of reducing microglial inflammation has been a major goal in neurodegeneration research, as dysregulated inflammation is a hallmark of most neurodegenerative diseases, and microglia, the brain's tissue-resident macrophages, play a large role in initiating this inflammation. The traditional approach of drug discovery through screening thousands of compounds is both costly and time-consuming. Therefore, inspired by a study by Smer-Barreto et al. (2023), we utilized data from a previous drug screening conducted by our lab to develop a machine-learning model that can identify new candidate drugs from online databases. Using this approach, we identified 36 promising compounds that may have anti-inflammatory effects on microglia and are performing experimental validations on them. Should the lab results return positive, this proof-of-concept study will demonstrate the validity of machine learning-assisted drug screening in inflammation research and facilitate the development of more efficient screening methods. Furthermore, the validated hits will be added to the repertoire of neurodegenerative therapeutics and help us study the mechanisms governing microglial homeostasis.

Layman Summary

Microglial cells, commonly known as the brain's immune sentinels, play a key role in worsening neurodegenerative diseases when they trigger excessive inflammation. Reducing this inflammation is a major focus of research, but finding effective drugs is typically slow and expensive. Inspired by recent advances, our team performed a drug screening and used its data to train a computer program (a process known as machine learning) to predict which compounds might work as anti-inflammatory drugs for microglia. This program identified 36 potential candidates, which we are now testing in the lab. If these tests find useful drugs, it will show how machine learning can speed up drug discovery in this context, paving the way for faster, more efficient methods to find treatments for brain diseases. It could also expand our understanding of how to keep microglial cells healthy and uncover new therapeutic options for conditions like Alzheimer's and Parkinson's.

1 Introduction

1.1 Project Aim

This project aimed to contribute to our lab's overarching goal of discovering the mechanisms of, and the means to mitigate, neuroinflammation. Specifically, this project was designed to 1) discover novel compounds that can encourage microglial homeostasis and 2) test a novel machine-learning-assisted drug discovery approach in the context of neuroinflammation.

To provide background for this project, I will illustrate the significance of microglia and explain the use of machine learning (ML) in drug discovery in this introduction.

1.2 Microglia in Health and Disease

1.2.1 Microglia Overview

Most organs in the body contain populations of tissue-resident macrophages (Y. Wu & Hirschi, 2021), and microglia are the tissue-resident macrophages of the central nervous system (CNS)—more specifically, of the brain parenchyma and spinal cord (Ginhoux et al., 2010). Thus, sharing the same myeloid origin with macrophages, microglia are best known for their role as immune sentinels, although they have also evolved additional functions after their prolonged residence in the CNS. In homeostatic conditions, microglia have relatively small cell bodies and send out highly branched and motile processes to survey their microenvironment for any changes (Nimmerjahn et al., 2005). Healthy microglia with this shape are typically referred to as ramified. Like other tissue-resident macrophages, microglia act as the "first responder" in the event of a pathogen invasion. In addition to external threats, they can also detect toxins and molecules (such as ATP) secreted by damaged cells in cases like traumatic brain injury or neurodegeneration (Davalos et al., 2005; Hanisch, 2002). Upon detecting these "danger-associated molecular patterns" (DAMPs) or "pathogen-associated molecular patterns" (PAMPs) microglia take on a rounder, more "ameboid" shape, which is generally taken as an indication of entering an "active" state, ready to engulf any threats (Augusto-Oliveira et al., 2019).

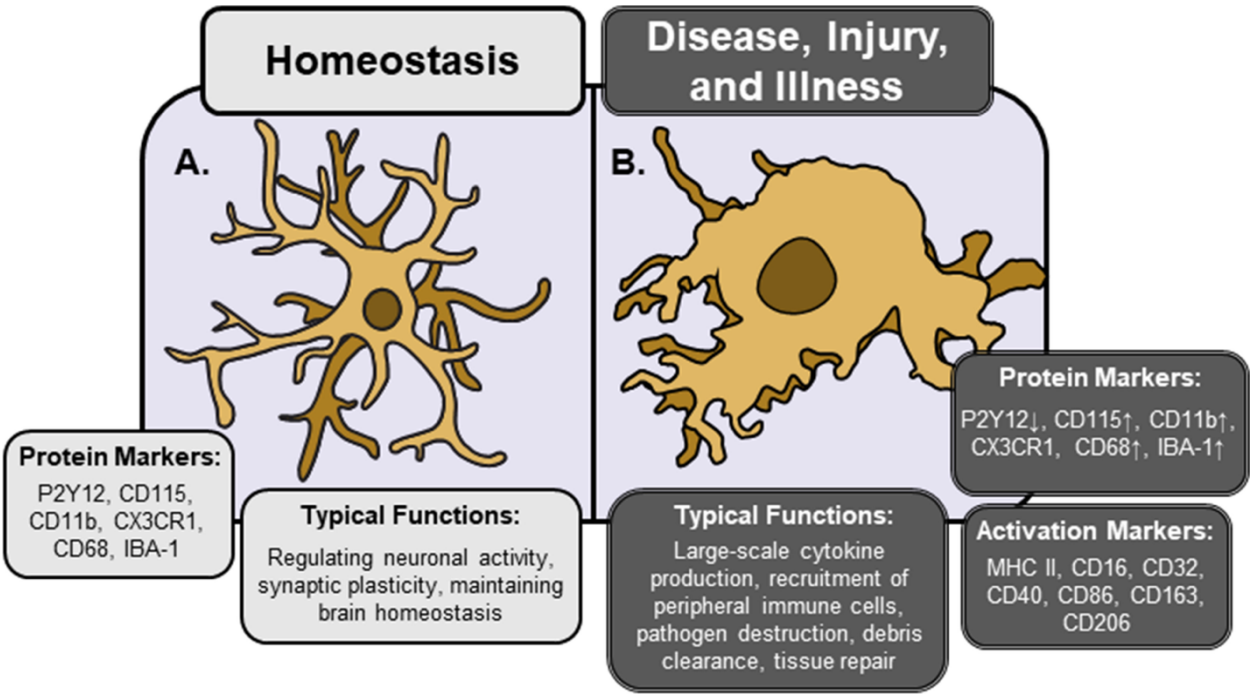


Fig. 1 (A) Homeostatic, ramified morphology. (B) Activated, amoeboid morphology (Woodburn et al., 2021)

For a long time, microglia, along with other glial cells, were thought to merely provide auxiliary functions for supporting neurons, the star of the brain responsible for all cognitive tasks. However, it is now widely recognized that, in addition to immune surveillance, microglia also play an active role in neurodevelopment and shaping brain circuits, thus being involved in learning and memory (Wolf et al., 2017). Therefore, when microglia become dysfunctional due to exaggerated inflammation, not only do they harm the tissue by inducing further inflammation in neighboring cells, but they also harm the tissue by neglecting their normal physiological functions. To illustrate why that is problematic, I will describe below some of the key functions performed by homeostatic microglia.

1.2.2 Microglia in Health

Microglia in Development

During the prenatal and postnatal stages of development, the brain undergoes a period of rapid neurogenesis. If left unchecked, the overproduction of neurons can disrupt proper neurodevelopment. In 2013, Cunningham et al. described how microglia serve as a mechanism for limiting cell production. In the developing brain, most of the cortical neurons are generated in two proliferative zones: the ventricular zone (VZ) and subventricular zone (SVZ). The group found that microglia selectively colonize these proliferative zones and phagocytose neural precursor cells. They also confirmed that phagocytosed cells did not display apoptotic signals but were viable cells. The authors suggested that pathological conditions during pregnancy may stimulate or inhibit microglia activity, leading to abnormal

development—a hypothesis supported by how maternal inflammation seems to be correlated with schizophrenia (Brown et al., 2004) and autism (Hagberg et al., 2012).

In addition to neurons, the brain also overproduces synapses during development, forming far more connections than are needed in adulthood. These excess synapses are gradually removed during maturation, a process known as synaptic pruning (Sakai, 2020). Synaptic pruning operates on a "use it or lose it" principle: circuits that are frequently activated become stabilized, while unused connections are pruned away. This pattern of strengthening and weakening is also called Hebbian plasticity (Faust et al., 2021). In a landmark study published in 2011, Paolicelli et al. discovered that microglia engulf synaptic elements in the healthy, developing mouse brain, particularly during the first few weeks after birth. When this process was disrupted by knocking out the fractalkine receptor, known to regulate microglia migration and interaction with neurons, they observed excessive immature synapses. Before this discovery, the role of microglia in healthy brains was poorly understood. Now, it is recognized that microglia play an important role in normal brain circuit maturation and regulating synaptic function.

When synaptic pruning goes awry, many neurodevelopmental disorders can occur. The disorders most associated with abnormal pruning are autism spectrum disorders and schizophrenia: autism is associated with insufficient pruning, while schizophrenia is associated with excessive pruning (Sakai, 2020). These abnormal pruning may be a result of microglial activation—when they become primed for immune functions rather than their usual homeostatic roles. Such activation could be triggered by various factors, including genetic predisposition or infection. Supporting this view, a postmortem study found increased microglial activation and proliferation, one of the activated phenotypes, in autism cases (Morgan et al., 2010).

Microglia in the Adult Brain

Role in immune functions

Microglia continue to play a vital role in ensuring optimal neural functioning past the development stage in a mature brain. As mentioned in **Microglia Overview**, their most famous role is in immune surveillance. The homeostatic microglia, with their ramified morphology (**Fig. 1**), extend long, motile processes equipped with pattern recognition receptors to monitor the environment. Upon detecting molecules that indicate threat (i.e., DAMPs or PAMPs), microglia take on their amoeboid shape and migrate towards the threat to perform their immune functions. Due to this observation, microglia have traditionally been classified as having two functional states: “resting” and “activated” (Paolicelli et al., 2022). However, this terminology is increasingly viewed as outdated. Firstly, the so called “resting” microglia are highly active. Even in the homeostatic state, they perform essential housekeeping tasks such as the removal of cellular debris and apoptotic cells through phagocytosis. This constant cleanup activity is crucial for maintaining a healthy neural environment and preventing inflammation (Neumann et al., 2009). Additionally, they mitigate excitotoxicity by migrating toward hyperactive neurons and clearing excess

glutamate, which prevents cellular damage and reduces the risk of seizures (Badimon et al., 2020; Kato et al., 2016).

On the other hand, "activated" microglia have been associated with deleterious neuroinflammation, even though their activities are highly variable and can result in both tissue repair and damage (Paolicelli et al., 2022). Acute inflammation in response to minor injuries or infections is necessarily for health, while chronic inflammation in pathological conditions can become maladaptive (as will be discussed in **Microglia in Disease**). Thus, newer terms have been borrowed from macrophage studies: M1 for the pro-inflammatory state and M2 for the anti-inflammatory state. However, a 2022 perspective paper, crediting over 130 researchers, asserted that this nomenclature is still overly dichotomous and does not capture the dynamic and multidimensional states of microglia (Paolicelli et al.).

Nonetheless, for the purpose of this thesis, I will use "homeostatic" to describe microglia performing routine functions, such as synaptic pruning, and "activated" to describe microglia responding to perceived danger, primarily in the maladaptive sense when I discuss neurodegenerative conditions.

Role in neuroplasticity

There is increasing recognition in microglia's role outside of immune defense, such as in neuroplasticity—a broad term that describes any changes to the brain's neural circuitry, but particularly changes related to learning and memory.

After the extensive period of synaptic pruning during early development, microglia's role in synaptic plasticity shifts towards supporting synapse formation in adulthood. In response to neuronal activity, they promote synapse formation by releasing brain-derived neurotrophic factor (BDNF). Depleting microglia or microglial BDNF in adult mice results in reduced learning-induced synapse formation and impaired performance on learning tasks (Parkhurst et al., 2013). However, it should be noted that the majority of the BDNF in the brain is still released by neurons (Parkhurst et al., 2013).

While microglia's pruning function is predominantly associated with the developmental period, in regions like the subventricular zone (SVZ) and dentate gyrus (DG), where new neurons are continuously produced, microglia still play a role in pruning the synapses of these adult-born neurons and supporting their maturation (Kurematsu et al., 2022).

Microglia's role in learning is further implemented by their interactions with the extracellular matrix (ECM). ECM, particularly in the form of perineuronal nets (PNNs), plays a major role in promoting the maturation of inhibitory neurons and dampening neuroplasticity (Pizzorusso et al., 2002). Recent studies have revealed that microglia can facilitate activity-dependent changes in certain hippocampal neurons by engulfing the ECM components, providing an additional mechanism for supporting synapse formation (Nguyen et al., 2020; Strackeljan et al., 2021).

Therefore, microglia perform a variety of essential functions that support neuronal health and plasticity. However, these essential functions are cast aside when microglia participate in chronic neuroinflammation.

1.2.3 Microglia in Disease

Although microglia play vital roles in maintaining brain health during homeostasis, their functions become increasingly detrimental when they transition into a chronically activated state during pathological conditions, such as neurodegenerative diseases, traumatic brain injury, stroke, or chronic psychiatric disorders. While the effects of their activation in these situations are complex and multifaceted, it is generally agreed that their net effects contribute to disease progression (Wolf et al., 2017).

Neurodegenerative Diseases

Traditionally, people did not consider microglia to be important in neurodegenerative diseases. Thus, researchers were surprised when genome-wide association studies (GWAS) revealed many disease-associated genes are related to microglia function, especially in Alzheimer's and Parkinson's disease (Bohlen et al., 2019).

Alzheimer's Disease

Alzheimer's disease (AD) is the most common form of dementia, which is the leading cause of death in England and Wales (Lane et al., 2018). The biological hallmarks of AD include an accumulation of misfolded and aggregated proteins, specifically amyloid-β (Aβ) plaques and neurofibrillary tangles (NFTs) (Scheltens et al., 2021).

a) Microglia & Aβ plaques

GWAS studies have shown that many AD risk genes are highly expressed in microglia, and a well-known risk gene is TREM2 (Bohlen et al., 2019). The TREM2 receptor helps microglia detect and phagocytose Aβ. However, this seems to be a double-edged sword in AD progression. Using a mouse line engineered to overproduce Aβ, a 2019 study found that knocking out TREM2 increased amyloid seeding, consistent with the gene's established role in microglial clearance of plaques (Parhizkar et al.). However, in the TREM2 knockout mice, the Aβ plaques also had less deposits of a harmful protein called ApoE, another major risk gene for AD. While ApoE likely has a protective role in sequestering Aβ, its excessive deposition around Aβ plaques also contributes to larger and more stable plaques. These larger plagues attract microglia without being degraded, leading to chronic inflammation. On the other hand, microglia without TREM2 do not cluster around these insoluble plaques, which reduces inflammation. This suggests that while microglial activation in early-stage AD helps clear Aβ, the same process promotes chronic inflammation and exacerbates neurotoxicity in late-stage AD.

Note, while studies generally consider the microglial clearance of Aβ to be beneficial, there is in fact an ongoing debate regarding Aβ plagues' role in AD—some hypotheses suggest that

they are toxic and induce oxidative stress, others argue that plaques may be a secondary or protective response to underlying pathology (Lane et al., 2018; Scheltens et al., 2021). Despite the controversy, most therapeutic attempts so far have focused on clearing Aβ plaques, and the microglial clearance of Aβ has also been a focus of research (Liu et al., 2019).

b) Microglia & NFTs

Neurofibrillary tangles (NFTs) are another of AD and are composed of intracellular aggregates of hyperphosphorylated tau protein. As tau aggregates are recognized by microglia's DAMP receptors, they amplify microglial activation, causing them to release pro-inflammatory cytokines such as IL-1β, IL-6, and TNF-α (Wolf et al., 2017). Furthermore, as microglia have limited capacity in neutralizing tau, they release tau seeds into the extracellular space, contributing to the propagation of tau pathology and further microglial activation across different brain regions (Hopp et al., 2018).

Parkinson's Disease

Parkinson's disease (PD) is the second most common neurodegenerative disorder, characterized by the progressive loss of dopaminergic neurons and the accumulation of α-synuclein aggregates, also known as Lewy bodies. While the exact cause of neuronal death in PD remains uncertain, growing evidence suggests an early protective and later deleterious role of microglia in the disease progression (Stefanova, 2022).

In the initial stages of PD, microglia detect and engulf misfolded α-synuclein aggregates, aiding in their clearance and preventing potential neuronal damage (Stefanova, 2022). However, persistent exposure to α-synuclein aggregates overwhelms the microglial clearance mechanisms, leading to a prolonged inflammatory response and impaired phagocytic ability. The inflammatory response is further triggered by dopaminergic cell death, and the resulting pro-inflammatory cytokines and reactive oxygen species (ROS) trigger yet more cell death, creating a positive feedback loop (Lv et al., 2023; Stefanova, 2022).

Additionally, similar to how they propagate tau proteins in AD, microglia can also contribute to the spread of α-synuclein. Prolonged inflammation decreases microglia's phagocytic ability, leading to incomplete degradation of α-synuclein, which could then be secreted via microglia-produced exosomes and be taken up by neurons (Lv et al., 2023).

Therefore, in both AD and PD, sustained inflammatory phenotype and the loss of their normal phagocytic ability result in microglia contributing to disease progression. Furthermore, these conditions typically manifest in later stages of life, when microglia are more prone to exaggerated responses (Lv et al., 2023).

Beyond Neurodegenerative Diseases

The 'double-edged sword' nature of microglia is a common theme in various conditions, including traumatic brain injury, epilepsy, and multiple sclerosis, with stroke serving as a well-studied example (Augusto-Oliveira et al., 2019; Bohlen et al., 2019).

In an ischemic stroke, microglia are traditionally known to promote pathogenesis. Microglia quickly adopt an activated phenotype in the early stages of ischemic stroke, releasing pro-inflammatory cytokines and reactive oxygen species (ROS) that exacerbate neuronal damage. Moreover, the activated microglia secrete matrix metalloproteinases that compromises the blood-brain barrier (BBB) integrity. The compromised BBB allows for the infiltration of peripheral leukocytes, which further exacerbate cellular damage and inflammation (Iadecola & Anrather, 2011; A. R. Patel et al., 2013).

However, more recent studies have also shown microglia exerting neuroprotective effects in stroke by producing anti-inflammatory cytokines such as IL-10 and transforming growth factor-β (TGF-β). Additionally, they phagocytose detritus at the injury site to help with healing as well as release vascular endothelial growth factor (VEGF), which promotes angiogenesis (Haupt et al., 2024). Consistent with these observations, a study found that depleting microglia in a stroke increased tissue damage, neuroinflammation, and peripheral leukocytes infiltration (Marino Lee et al., 2021). However, the opposite result has been reported as well (Li et al., 2021).

These conflicting results might be reconciled by the criticism on the “resting” (M2) vs “activated” (M1) dichotomy. A single-cell RNA sequencing study found heterogeneity among the microglia after ischemic stroke, none of them fully expressing either M1 or M2 marker genes (Guo et al., 2021). The two conflicting microglia depletion studies cited above targeted different cell-surface receptors in their microglial ablation, hence it is likely that they each removed different microglial subpopulations.

In conclusion, whether it is neurodegeneration or neural injury, microglia have the potential to confer protective effects, but the overwhelming pro-inflammatory triggers tend to cause them to adopt a maladaptive phenotype. Therefore, therapeutics capable of restoring homeostatic functions to microglia could provide benefits in a wide variety of pathologies.

1.2.4 Non-Cell-Autonomous Regulation of Microglial Homeostasis & TGFβ-2

Healthy microglia express a set of established “homeostatic markers”, such as *Cx3cr1*, *P2ry12*, and *Tmem1*. Researchers found that microglia isolated and grown in culture downregulate these homeostatic markers and upregulate disease-associated genes, becoming similar to activated microglia (Cadiz et al., 2022). One major reason behind this phenomenon is that microglia require continuous inputs from neurons and astrocytes to maintain their homeostasis. And, in 2021, our lab showed that this non-cell-autonomous regulation of microglia is primarily mediated by the secretion of Transforming Growth Factor

Beta 2 (TGF-β2) (Baxter et al., 2021). TGF-β2 is also known to perform various other roles, such as neurodevelopment, stem cell regulation, wound healing, as well as the proper development of multiple organs, including the heart, lungs, kidneys, and skeletal system (Wang et al., 2023). This broad spectrum of effects makes it a poor drug candidate. In addition, mid-sized hydrophilic proteins like TGF-β2 are typically degraded by the digestive system, have poor intestinal absorption, and cannot cross the blood-brain barrier. Nonetheless, its consistent anti-inflammatory effects on microglial cultures made it a suitable choice as the positive control in our drug screening and validation experiments.

1.3 ML-Assisted Drug Discovery

Drug discovery is a notoriously lengthy and laborious process, but the recent advances in computing power, large-scale data, and artificial intelligence (AI) offer the potential to accelerate the process. The application of machine learning (ML), a subset of AI techniques, in drug discovery is a relatively new but rapidly growing area of study. While computational approaches to drug discovery, such as molecular docking and quantitative structure-activity have existed for much longer, the integration of advanced ML techniques became prominent only in the last 10–15 years (Deng et al., 2022).

Machine learning, in its essence, is the process of enabling computers to extract patterns from examples. The pattern could be as simple as a linear relationship (e.g., mass vs volume), or a complex pattern that can only be observed when simultaneously considering a multitude of variables, which is the case in predicting the effects of chemical compounds. The goal of ML research in this area is to be able to learn hidden relationships between known drugs in order to discover novel drugs that share the hidden patterns.

We were inspired by a recent paper published by Smer-Barreto et al. (2023), who had success in using ML to discover novel senolytic drugs based on the data on known senolytics. One unique aspect about their study was that their training data came entirely from the literature: their positive observations were known senolytics found in past studies, while their negative observations were compounds from a chemical library with no reported senolytic effects. This approach, while efficient, ran the risk of containing false negatives in the data. However, the group reasoned that since senolytics are rare, the risk is low. They only had 58 observations with the desired label, which is unusual for ML applications. The actual data used was the physiochemical descriptors of the compounds, which were relatively easy to calculate using open-source cheminformatic packages. After virtually screening 4,340 compounds, the group selected 21 compounds for experimental validation and discovered 3 novel senolytics, thereby achieving a 14.3% working hit rate. This success demonstrated the potential of ML to efficiently reduce screening costs. Therefore, we wanted to test if the same approach could be used in searching for compounds that promote microglial homeostasis.

Another paper by Wong et al. (2023) employed a message-passing graph neural networks (GNNs) approach to identify senolytics. They performed an initial experimental screening on

2,352 compounds, which yielded 107 hits with senolytic activity; subsequently conducted a virtual screening of 804,959 compounds; selected 216 for experimental validation; and discovered 25 hits—achieving an 11.6% working hit rate, a sixfold increase.

When deciding which approach to follow—the conventional ML approach or the more novel GNNs—we chose the former due to 1) its relatively simpler usage, 2) the lack of correct predictions in our initial attempts with GNNs on the training data, and 3) the accessibility of Smer-Barreto et al., who are also based in Edinburgh. Additionally deep neural networks are generally considered as “black boxes”—that is, decision-making processes of the models are difficult be interpreted by humans. In drug discovery, it is desirable to learn generalizable principles about the structure-effect relationship, so more interpretable models are often preferred (Z. Wu et al., 2023). However, this caveat would be outweighed if GNNs were to significantly outperform other models.

2 Methods

2.1 Methods Overview

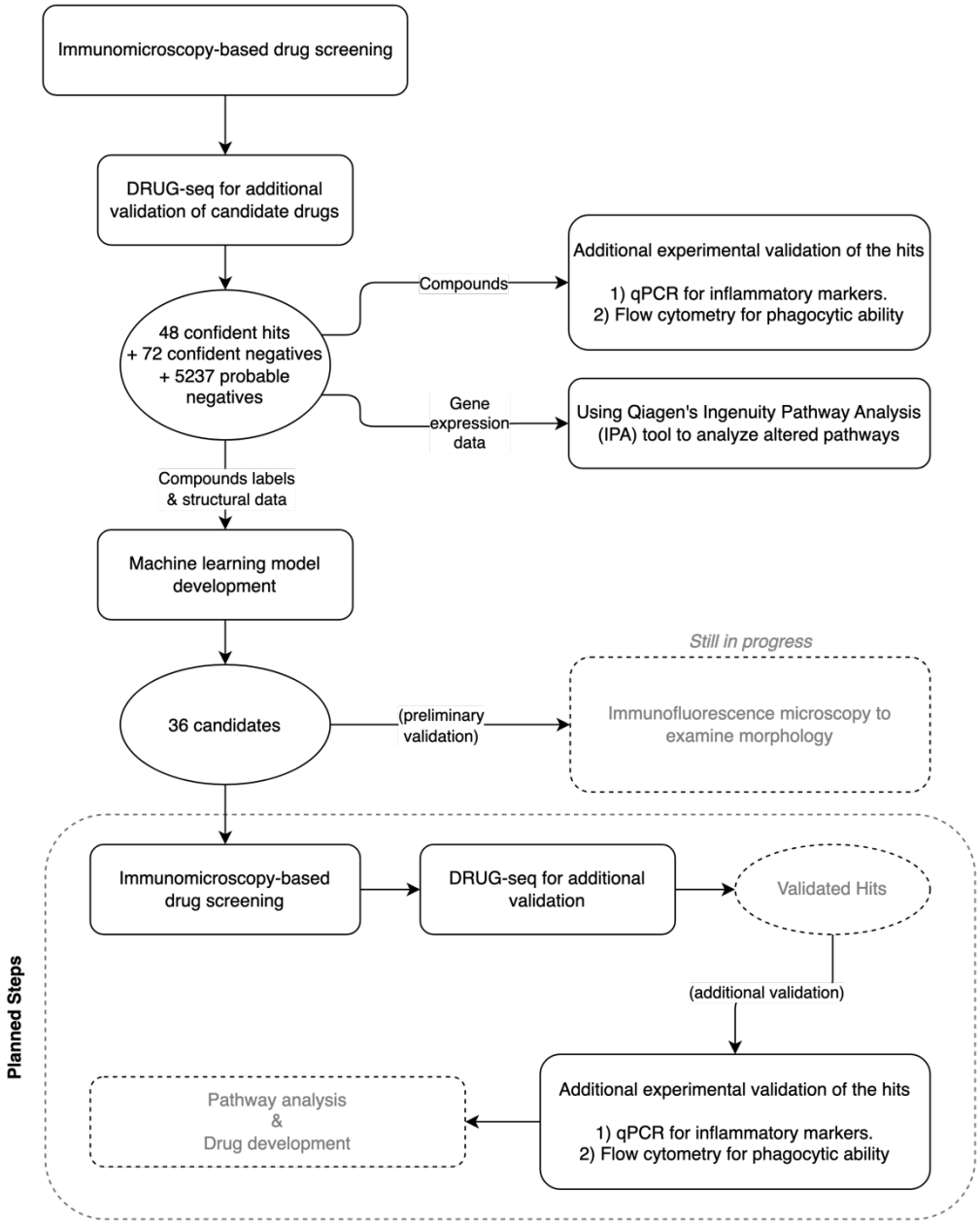


Fig. 2 Overview of the performed and planned research steps

Fig. 2 shows the steps we have performed and planned. In brief, we performed an initial screening that examined each compound's effect on microglia by examining the microglial morphology with immunofluorescence microscopy. The screening process is imprecise, however, so the hits were subjected to additional screening via DRUG-seq, which outputted transcriptomics data and allowed us to identify 48 positives and 72 negatives with relative

confidence. The 48 positives were further tested for their effects on microglial inflammation and phagocytic ability. Simultaneously, the machine learning (ML) was underway. 36 candidates were selected from the ML model and will undergo the same validation procedures as the original hits had. (Note, the 36 candidates have an arrow labeled "preliminary validation" to the box that also has an arrow labeled "additional validation" pointed to because the two use the same procedures—qPCR and Flow cytometry. However,

2.2 Machine Learning

2.2.1 Overview

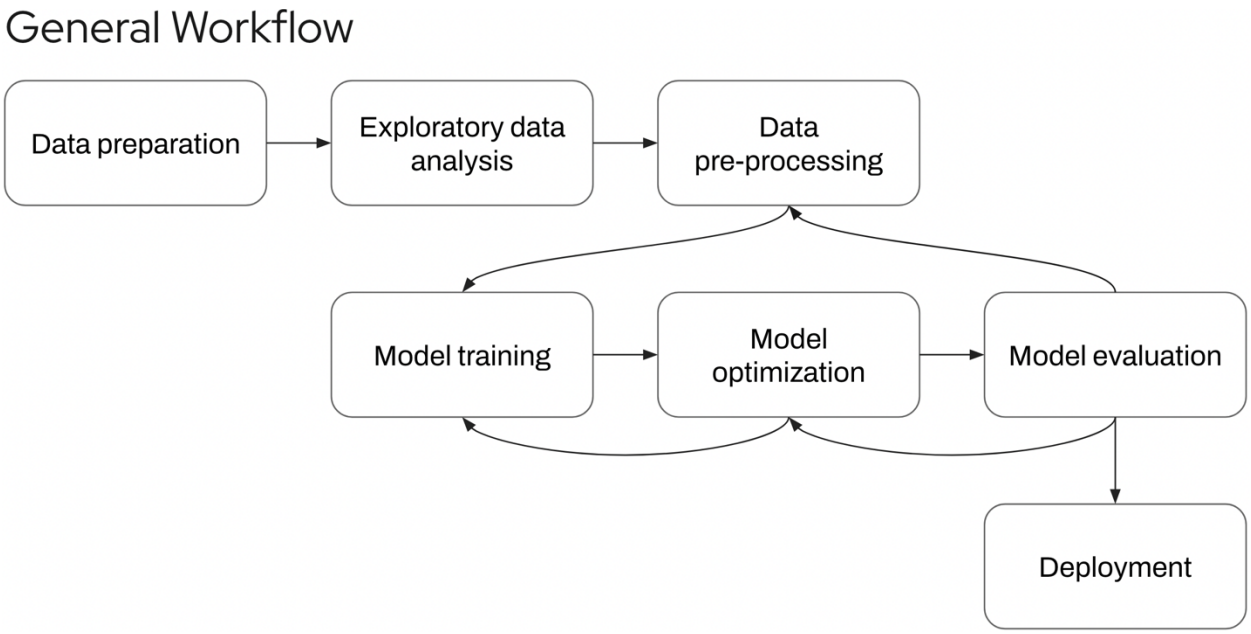


Fig. 3 General workflow of ML projects

Fig. 3 shows the general workflow of the ML projects, which we attempted to follow. However, due to both the nature of machine learning projects and the experimental nature of this project, where many trials and errors were involved in determining what worked best, the workflow was not linear but rather iterative and required backtracking at times. Regardless, this chart still represents the general progression.

2.2.2 Data Preparation

Data Source

We performed a drug screening on a total of 6651 compounds and obtained 6388 usable results, of which 5357 were unique compounds (see *Initial Drug Screening* for the screening details). A subset of promising compounds underwent further screening, from which we identified 48 hits (positives). All other compounds were labelled as negatives. However, due

to the difficulty of the high throughput screening, the labels were not conclusive, especially for the compounds that did not go through a second round of screening. We are more certain of the positive labels, all of which were screened twice, than we are of the negative labels. However, we also identified 72 compounds that were confidently negative from the second screening and will be referred to as "validated data" together with the 48 positives.

Featurization and Data Cleaning

To represent the compounds in a machine-readable format, we recorded the structure of each compound as a Simplified Molecular Input Line Entry System (SMILES) string, the most popular molecular string representation used in cheminformatics since the late 1980s (Krenn et al., 2022; Weininger et al., 1988). There are many ways to capture information about a molecule as features for use in machine learning, and we decided to use computed chemical descriptors, which encompass a broad range of numerical values derived from molecular structures, such as physicochemical properties (e.g., logP, molecular weight) and electronic characteristics (e.g., partial charges). To obtain these features from the SMILES strings, we used RDKit (Version 2023.9.1), an open-source cheminformatics toolkit, which computed 211 chemical descriptors for each compound. RDKit failed to compute certain descriptors for several compounds. As a result, we faced the choice of either filtering "Not a Number"s (NaNs) by row and exclude 225 compounds (one of which was a hit) or filtering NaNs by column and exclude 12 features. We decided to exclude the columns containing invalid data because 1) we already have very few hits in our dataset, so losing even one would be consequential, 2) a univariate analysis suggested that each feature only weakly correlated with the target value, and 3) a principal component analysis shows that only 99 to 111 number of components (for MinMax and Standard Scaler, respectively) are needed to explain 99% of the cumulative variance (**Fig. 4**). Note, as there were not enough information to determine which scaler best applies to our data at this point, we included the results from both in our preliminary analyses. We also discovered an error in RDKit that duplicated a descriptor called "SPS". Thus, after removing the duplicated and NaN-containing features, we had 198 features at this stage.

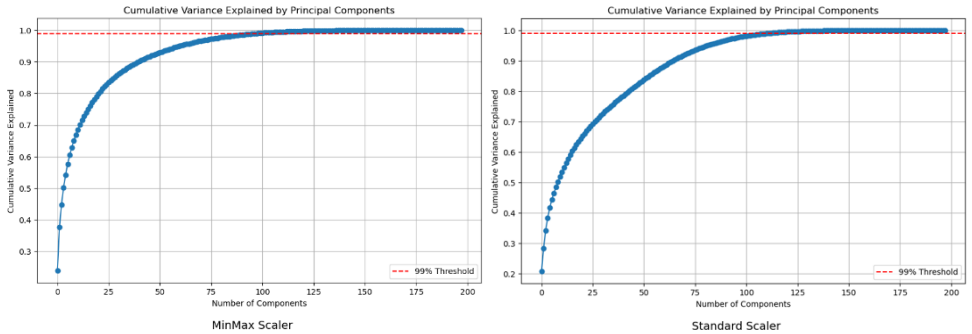


Fig. 4 Cumulative variance plots from principal component analyses with MinMax and Standard Scalers.

The libraries used contained overlapping compounds, and several compounds' SMILES strings could not be successfully parsed by RDKit. So, after removing duplicates and

compounds with invalid SMILES, there were 5363 observations in the training dataset (48 positives and 5315 negatives).

We also explored an alternative method of featurization using Chemprop, a machine-learning package that implements directed message-passing neural networks (D-MPNNs), as employed in the aforementioned study by Wong et al. (2023). In this approach, molecules are represented as graphs, with atoms as nodes and bonds as edges. Unlike RDKit, Chemprop does not use pre-defined features; instead, it extracts its own features to represent the molecules during the training process, although RDKit descriptors can still be added as additional features (Heid et al., 2024). However, our initial attempts with Chemprop yielded poor results—it failed to identify any hits even when trained and tested on the same dataset. Thus, we concluded that the former approach was better suited to our needs.

2.2.3 Exploratory Data Analysis

Chemical Diversity Analysis

To evaluate the chemical diversity of our training data, we performed a K-means clustering analysis on the compounds with the full set of features (198 features at this stage), scaled with Scikit-Learn's StandardScaler, which standardizes features by removing the mean and scaling to unit variance(Scikit-learn, 2024). In K-means clustering, the user specifies the number of clusters (K) for the algorithm to attempt to identify. We specified the range 2–20 K. At each K, we can quantify the quality of the clustering with the silhouette coefficient. In brief, a silhouette coefficient (S) measures the average distance of a point to the other points in its own cluster compared to points in other clusters, thereby proving a measure of how cleanly the different clusters are separated. For any given point, $S \approx 1$ means that the point is well-matched to its own cluster and poorly matched to neighboring clusters, $S \approx 0$ means that the point lies in-between clusters, and $S \approx -1$ means that the point may have been assigned to the wrong cluster. Thus, the average S of all points (also known as the silhouette score) quantifies the separation of K clusters (Shutaywi & Kachouie, 2021). Another measure used was the within-cluster sum of squares (WCSS), which is the sum of the distances from each datapoint to the centroid of its corresponding cluster. See **Fig. 6** for the results.

We have also generated a t-SNE plot (perplexity = 30, n_iter=3000) to visually inspect the spread of our compounds in the chemical space (**Fig. 7**).

2.2.4 Model Tuning and Selection

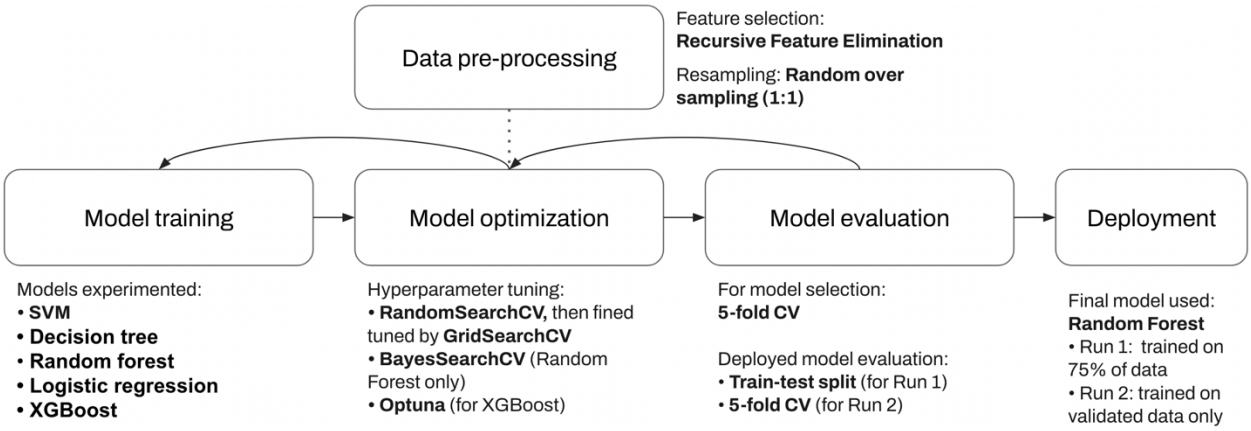


Fig. 5 Model selection overview. The flowchart shows the general ML process after the data preparation and exploratory analysis steps. The texts indicate the techniques used in the end. Not shown here are the techniques explored but not ultimately used.

Fig. 5 illustrates the general workflow for the model selection and optimization process. Initially, a range of machine learning algorithms with default hyperparameters was evaluated to identify potential candidates. These models were subsequently optimized to improve performance, ultimately leading to the selection of the model that demonstrated the most favorable metrics for the task. The looping arrows illustrate the iterative nature of the process, which included trial and error. Note, the data pre-processing step, especially the feature selection aspect, is closely tied to model optimization because the main method we used was Recursive Feature Elimination (RFE), which is a model-based feature selection method whose optimal performance requires model tuning beforehand.

Baseline Models

All the machine learning (ML) algorithms were provided by Scikit-Learn 1.4.2 (Pedregosa et al., 2011) and XGBoost 2.1.0 (Chen & Guestrin, 2016). As training ML models to predict drug properties is still an area of active research, there are no established pipelines, and a range of options needed to be experimented. Thus, we started with several algorithms to use as our baseline models: support vector machine (SVM), decision tree, random forest, and logistic regression. See *Evaluation* for how the models were compared.

Data Preprocessing

Class weighting: Firstly, the extreme imbalance in our dataset (48 positives vs. 5315 negatives) was problematic. Without adjusting the class weights or ratio, the recall and precision scores were zero for all models.

Setting the class_weight parameter to “balanced” (assigning each class a weight that is the inverse of its proportion in the dataset) made a significant difference for the SVM classifier, raising the average recall from 0 to 0.14. However, the precision was only 0.046, on average.

Since our goal was to use the ML model to reduce the number of compounds to screen, we aimed to maximize precision.

Resampling: To address the data imbalance issue, we tried undersampling the minority (positive) class, oversampling the majority (negative class), and a combination of both, across a range of ratios. We found that a 1:1 random oversampling via the Imbalanced-learn module (version 0.12.3) worked best, although the improvements were small before model tuning. Note, the random oversampling was performed anew with each model training cycle as part of the pipeline.

Another problem we discovered was that glucocorticoids represented a disproportionally large number of positive compounds. While the rest of the hits were generally chemically dissimilar, 8 of the 48 hits belong to the same structural family, which greatly biased the model predictions. As a result, many adjustments that seemingly improved the metrics simply enhanced the identification of positive glucocorticoids. Thus, we attempted oversampling only the non-glucocorticoid hits, but that did not seem to work better than a simple random oversampling—metric-wise or in identifying hits of other classes, so we continued using random oversampling.

Scaling: Initially, we reasoned that since not all of our features follow a normal distribution, the MinMaxScaler should be more suitable. But, through experimentation, we found that StandardScaler tended to give a higher precision score than MinMaxScaler. So, any results mentioned onward will be from data scaled with StandardScaler.

Normalization: Since not all of the features are normally distributed, we tried applying log transformation and the Yeo-Johnson transform to our data. Neither of which seemed to improve our metrics, so normalization was not used.

Feature Selection

First, we removed several features with zero variance. There were also a number of sparse features with 0 at most observations; those were usually the chemical fragment descriptors. Thus, we tried filtering features with a range of high sparsity score (more than 90-98% of observations being 0) but that did not yield significant improvements. In the end, we selected 64 features (see **1** in **Appendix**) using recursive feature elimination (RFE), with our best tuned random forest model as the estimator. RFE also filtered out most of the sparse features.

Some of the other feature selection techniques we used included 1) univariate analysis, where we ranked the features by their Pearson correlation with the target variable and tested a range of top n features (where $n < 111$, as suggested by our PCA cumulative variance plot **Fig. 4**). 2) Another univariate analysis where we only included the features that had a significantly different mean between the positive and negative class, which yielded less than 13 significant features. 3) Forward sequential feature selection (SFS), which proved too computationally heavy for our dataset. On a Windows desktop with an Intel® Core™ i7 processor, it took approximately 7.8 days for SFS to finish running with an SVM estimator,

making it impractical to repeat this process with the other models. RFE outperformed all techniques.

Hyperparameter Tuning

The best hyperparameters for the random forest model (see **2** in **Appendix**) were found by Bayesian optimization, as implemented by Scikit-Optimize 0.10.1 (Louppe et al., 2022), while the other models were tuned with a combination of random search and grid search. Random forest was the only model for which Bayesian optimization was successfully completed. For SVM, the optimization process was aborted when it was still running after 14 days on the Windows desktop. Likewise, Bayesian optimization could not be completed in a sensible timeframe for the other models. After tuning, random forest emerged as the model with the best precision and an acceptable recall, on average. See **3** in **Appendix** for the hyperparameter space explored.

Semi-supervised Learning

As mentioned earlier, the imprecise nature of high throughput drug screening means that there could be many false negatives among the training data, which could potentially introduce noise and confuse the ML algorithms. Thus, in an attempt to reduce the effects of potential false negatives, we experimented with two forms of semi-supervised learning: 1) Positive-unlabeled (PU) learning, implemented by the pulearn module (Drouin et al., 2024). It assumes that only some of the positive observations are labeled, while the rest of the data consists of unlabeled positives and negatives. This implementation of PU learning uses a hold-out set of known positives to estimate the proportion of positives within the unlabeld data. So, with this method, we treat all of our negative data as unlabeled (Elkan & Noto, 2008). 2) Self-training, implemented by scikit-learn, starts with a base classifier trained on a smaller set of labeled data and makes predictions on a larger set of unlabeled data. It treats confident predictions as new labeled data, retrains the classifier with the additional labels, thereby iteratively expanding the labeled set of data until improvement stops (van Engelen & Hoos, 2020). So, instead of classifying all of the negatives as unlabeled, we include the 72 “validated” negatives as part of the initial training data for self-training.

Both algorithms seem to promiscuously label data as positive. For example, in one 5-fold CV, a self-training model with random forest as its base classifier identified 25 out of the 48 true positives across the 5 folds—much better than any other model—but also predicted, on average, 525.4 ± 109.7 hits in each fold, when there were only 9-10 hits per fold. Since our goal is to reduce the number of candidates for future screening, and there was no way to confirm whether these imputed positives were real except by experimental validation, we decided to abandon the semi-supervised approach for now.

Evaluation

When comparing the different models, hyperparameters, and pre-processing techniques, we examined the metrics accuracy, precision, recall, and F1. In the beginning, when we could hardly obtain a recall score above a 0, our emphasis was on raising the recall. Later,

when we could consistently obtain a recall of more than 10% in the models being tested, we shifted our focus to maximizing precision, as our goal was to narrow down the number of compounds for screening. The metrics were calculated using stratified 5-fold cross validations (CVs), with sklearn's StratifiedKFold (n_splits=5) and a fixed random seed for equal comparisons. Occasionally, when the differences between two manipulations were small, we would repeat the 5-fold CV five or more times (without a fixed random seed) and perform a paired t-test to discern if there are significant differences between their metrics' means.

Beyond the metrics, we manually inspected the true positives predicted in each CV to check for biases towards any specific chemical groups. As mentioned, our models could often correctly identify glucocorticoid hits but rarely other hit compounds. Since glucocorticoids were of lesser scientific interest, we wanted to ensure that our model could also predict a broader range of compounds. Therefore, between a model with a high precision score—due to correctly predicting many glucocorticoids—and a model with a lower precision score but more diverse predictions, we would prefer the latter.

When deploying our final model, we adhered to the conventional practice of using a train-test split, training the model on a stratified 75% of the data and evaluating it on the remaining 25% (results shown in **Fig. 10**)

2.2.5 Deployment: Virtual Screening

Virtual Screening Data

For the virtual screening, we compiled 5041 unique compounds (which we will refer to as the discovery library) from Selleckchem's L3800 Drug Repurposing Library and Targetmol's L2100 Anti-cancer Compound Library.

Training the Final Model

Initially, we trained our tuned random forest model on 75% of our training data. However, as mentioned in **Data Source**, we only had 72 confidently labeled negatives. Hence, to address the potential false negatives problem, we explored training the model again on only the 120 validated observations (72 negatives + 48 positives), producing a second set of predictions. Given the small size of the validated data, we opted to forgo creating a train-test split.

Hit Selection

After making predictions on the discovery compounds, we ranked them by their predicted probability of being a hit. Since our models' scores were relatively poor, combined with their tendency to predict glucocorticoids, we reasoned that the higher hanks may be polluted with compounds not of our interest. Therefore, instead of simply selecting the top *n* compounds, as done in Smer-Barreto et al. (2023), we decided to manually verify the potential of the highly ranked compounds through searching the literature and select accordingly. We

reviewed the literature for the top ~50 compounds for both sets and manually selected a total of 36 compounds for validation.

We also explored training our model on 100% of the training data to maximize learning examples, and a visual inspection of the predictions appear similar to the first run's. However, we decided not to manually inspect its top compounds due to time constraints.

2.3 Compound Testing on Microglia

2.3.1 Initial Drug Screening

Prior to my arrival at the lab, a phenotypic drug screening was performed on a total of 6651 compounds obtained from three chemical libraries: 1520 from Prestwick Chemical Library, 2131 from Target Mol's Natural Compounds Library, 3000 from Target Mol's FDA-Approved Drug Repurposing Library. However, a few plates were excluded due to poor data quality collected from the control wells.

The phenotypic screen used primary rodent microglia culture (see Microglia Culture below for the culture protocol) incubated with compounds in 384-well microplates for 48 hrs. TGFβ-2 (R&D Systems) served as the positive control, while 0.1% DMSO was used as the negative/vehicle control. After the incubation period, the cells were fixed with 4% paraformaldehyde and subsequently stained with CX3CR1 (fractalkine receptor) primary antibodies, biotin secondary antibodies, and streptavidin-FITC conjugate. Cells were then permeabilized with 0.1% triton and stained with CellMask™ (Invitrogen) and Hoescht stain. Finally, the cells were imaged at 10X and 20X. These steps were all performed with an automated liquid handling system from the Edinburgh Drug Discovery group, who also has more details for the specific reagents used.

The outcome we were studying was the microglia morphology, which correlates with the cells' activation status. A small cell body with many branched, ramified processes indicates homeostatic microglia, while a large, "ameboid"-shaped cell body with retracted processes indicates inflamed or activated microglia. TGFβ-2 is known to induce the ramified morphology, so the shape of the microglia incubated with it was used to establish the baseline for positive outcomes. The cell morphology was quantified with StratomineR, which automatically extracts relevant morphological features, such as circularity, process length, and cell body area.

Based on the morphological data, a selection of promising candidates was further tested with Digital RNA with pertUrbation of Genes (DRUG-seq, provided by Alithea Genomics) to assess their transcriptomic profiles. The cells were incubated and treated under the same conditions. Based on the DRUG-seq results, we identified 48 hits (positives) with reasonable confidence.

We plan to apply the same screening procedures to the compounds selected based on the ML model.

2.3.2 Microglia Culture

The protocol for creating the primary rat microglia culture is based on a protocol published by this lab, which contains step-by-step instructions (Qiu et al., 2018). To begin, brains from rat pups of post-natal day 0–3 were extracted and transferred to a 9:1 mixture of dissection medium (DM) and kynurenic acid during the dissection. The cortices were cut off and kept, while the rest of the brain was discarded. The tissues were incubated with papain at 37°C for 20 minutes, then mechanically homogenized using a serological pipette. For plating, the cell suspension was diluted in DMEM + 10% FBS to a total volume of 100 mL per ~10 brains and distributed into poly-D-lysine (PDL)-coated flasks (10 mL per flask) and put into a humidified tissue culture incubator (37 °C, 5% CO_2, 20% O_2). The medium (DMEM + 10% FBS) was replaced every 3-4 days, and the cells would become ready for experiments after 10-12 days. During this period, neurons would die off, leaving behind only glial cells, as the medium is not designed to support neurons.

Microglia isolation

After 10 days, the microglia were isolated using an orbital shaker. The air vents of the flasks were tightened before being placed in the shaker. Then, the flasks were shaken at 140-180 rpm for 1 h at 37°C. The resulting suspension, containing predominantly microglia, was then centrifuged, the medium was replaced with serum-free Neurobasal A (NBA) medium, and the cells counted using a hemocytometer. The cell concentration was adjusted to approximately 2×10^5 cells/mL before plating 1 mL per well into 24- or 48-well plates.

2.3.3 Drugging

The 48 candidate drugs obtained from the initial screening were generally added to the microglia at concentrations of 3.2 µM, 1, µM, 0.32 µM, and 0.1 µM. Compounds selected through the virtual screening were tested at 1 µM, 0.1 µM, and 0.01 µM. For positive control wells (containing 1 mL of medium), 1 µL of TGFβ-2, stored as 0.25 ng/mL aliquots, was added, while 1 µL of DMSO was added to the negative control wells. After 24 hours of incubation, cells designated for a phagocytosis assay with flow cytometry were extracted for the procedure. For RT-qPCR studies, 1 µL of 200 ng/mL lipopolysaccharide (LPS) was added to each well, followed by an additional 24-hr incubation. LPS, a proinflammatory component of bacterial membranes, was used to study the protective effects of the compounds.

2.3.4 RT-qPCR

RNA was extracted using the High Pure RNA Isolation Kit by Roche, and cDNA was synthesized with Qiagen's QuantiTect Reverse Transcription Kit, as per the manufacturers' instructions. Each well of the 96-well PCR reaction plate contained the following: 1 µL cDNA

sample, 1.2 µL primer, 7.5 µL SYBR Green Master Mix (Thermo Fisher), and 5.3 µL nuclease-free water. The qPCR was performed with a Stratagene Mx3000P QPCR System (Agilent Technologies) using the following program: 10 minutes at 95°C, followed by 40 cycles of 30 seconds at 95°C, 40 seconds at 60°C, and 30 seconds at 72°C. A dissociation curve was measured with an additional cycle of 1 minute at 95°C and 30 seconds at 55°C. Rox was used as the reference dye.

2.3.5 Phagocytosis Assay via Flow Cytometry

After removing the medium, the cells were incubated with pHrodo™ BioParticles™ Conjugates (Invitrogen), diluted 1:10 in Dulbecco's Phosphate Buffered Saline (DPBS), for 1–2 hours. The cells were then mechanically detached from the plates using a pipette tip, and the dye was removed by centrifugation. The resulting cell pellets were resuspended in DPBS to a final volume of 100 µL per sample. To this suspension, 1 µL of the cell death marker 7-AAD (Invitrogen) was added to achieve a final concentration of 1 µg/mL. The flow cytometry analysis was performed with a BD Accuri™ C6 flow cytometer (BD biosciences).

2.3.6 Pathway Analysis of the Hits

We obtained transcriptomic data for 42 hits with DRUG-Seq. 6 of the 48 hits were selected midway through the project and were not included in this analysis. We calculated the log2 fold change in gene expression between the drug-treated condition and the DMSO control, both incubated with LPS, as described in **Drugging**. The gene fold change data were then analyzed using Qiagen's Ingenuity Pathway Analysis (IPA) tool, which identified molecular pathways associated with the differentially expressed genes. A p-value cutoff of 0.1 was applied, resulting in 33 compounds with sufficient gene expression data for IPA's analysis. After IPA returned the differentially regulated molecular pathways for each compound, they were counted and ranked using Python.

2.3.7 Immunofluorescence microscopy

Preliminary testing of the ML-derived hits was conducted using immunofluorescence microscopy, though it did not produce usable results. The cells were fixed with 4% paraformaldehyde (PFA) for 10 minutes and then permeabilized with 0.5% Triton X-100 for another 10 minutes. Following permeabilization, the cells were incubated with CellMask (Invitrogen) prepared by diluting a 1:10 stock in water, followed by a 1:10,000 dilution in PBS. The cells were incubated with CellMask for 20 minutes in the dark. A mounting medium containing DAPI (Invitrogen) and coverslips were added to each well, except when using a 48-well or higher-density plate. The imaging was performed with a Leica AF6000 LX imaging system.

3 Results

3.1 Chemical Diversity Analysis Results

a)

b)

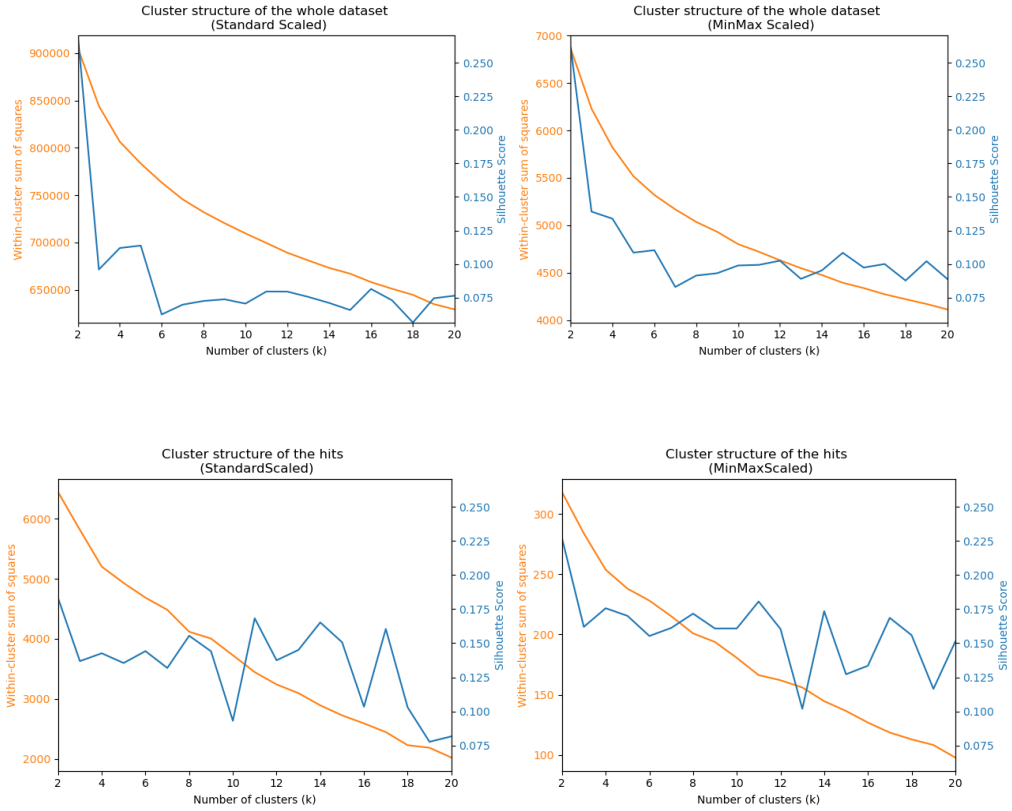


Fig. 6 The silhouette scores of K-means clusters with K = [2,20], with different scalers applied. **a)** Analysis of the whole dataset **b)** Analysis of the hits only.

The silhouette scores were consistently low (**Fig. 6a**), with an average of 0.088 for data normalized with StandardScaler (0.11 for MinMaxScaler), which is desirable in this instance as it indicates a diverse range of compounds that spreads across the continuum of chemical space. We also examined the feature diversity of the hit compounds (**Fig. 6b**). Likewise, the silhouette scores were generally low (an average of 0.14 with StandardScaler and 0.16 with MinMaxScaler). Similarly, the lack of a clear "elbow" in the WCCS scores also indicate poor clustering. Therefore, according to the K-means clustering analysis, our selection of hits and dataset in general have a decent amount of chemical diversity, which increases the likelihood of generating novel findings.

From the t-SNE plot (**Fig. 7**), upon visual inspection, our hits appeared to be quite evenly spread out across the chemical space relative to the negatives.

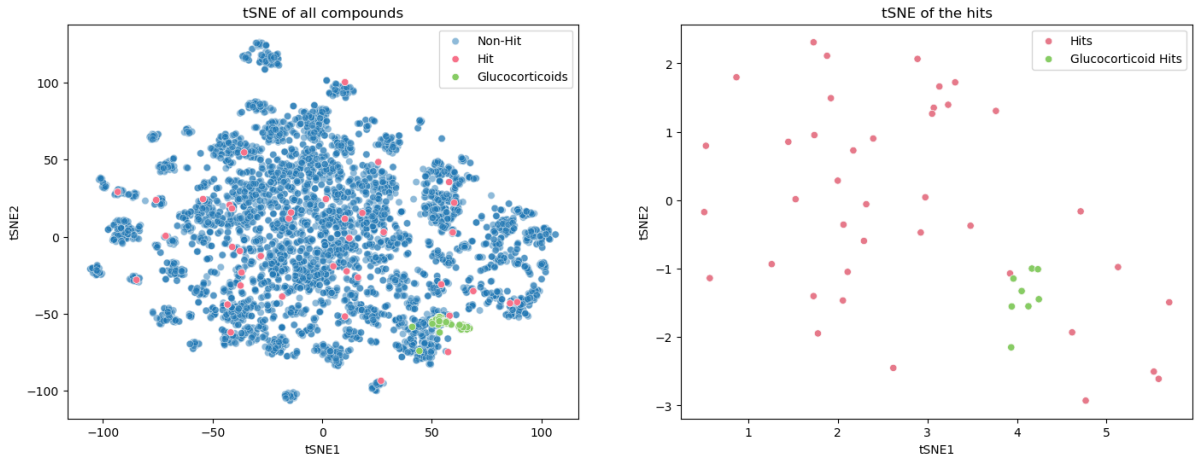
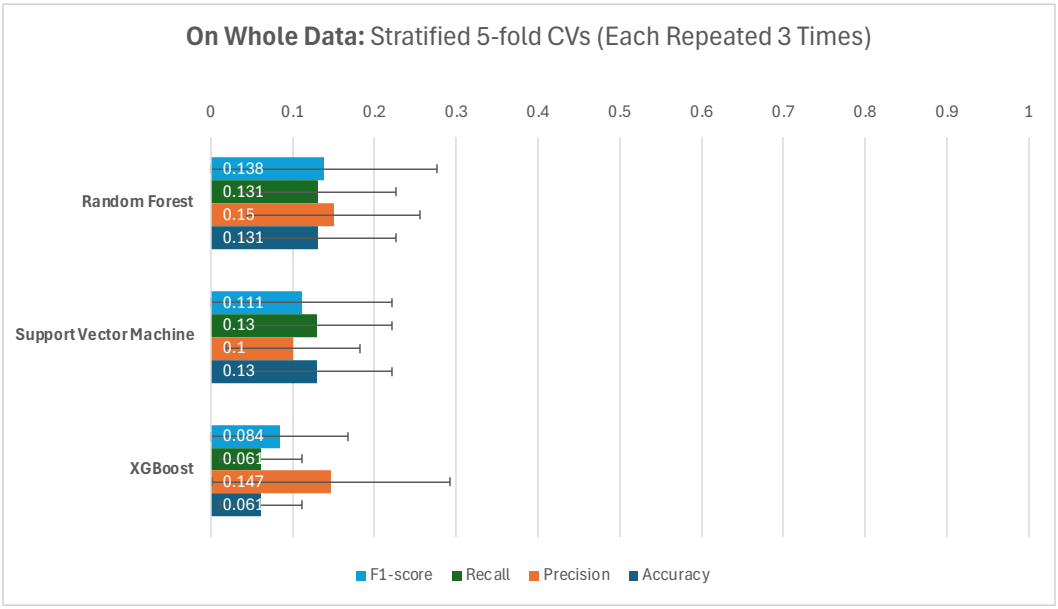


Fig. 7 t-SNE projections of our 198-dimensional data. Left: All training compounds (perplexity = 30, n_iter=3000). Right: Hits only (perplexity = 25, n_iter=1000).

3.2 Model Metrics

The figures below show the CV results of the three finalist models, when evaluated using the whole dataset (**aFig. 8**) or only the validated dataset (**Fig. 9**), as well as the final train-test split evaluation of the deployed model (**Fig. 10**).



aFig. 8 Metrics for Random Forest, Support Vector Machine (SVM), and XGBoost, each evaluated using three repeated stratified 5-fold cross-validations on the entire dataset.

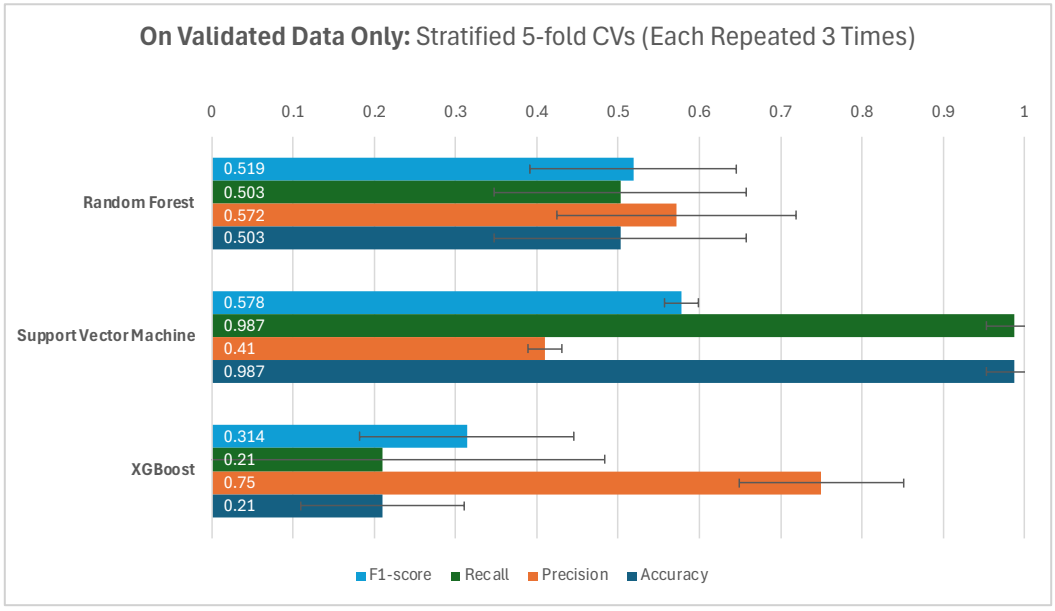


Fig. 9 Metrics for Random Forest, Support Vector Machine (SVM), and XGBoost, each evaluated using three repeated stratified 5-fold cross-validations on the validated dataset.

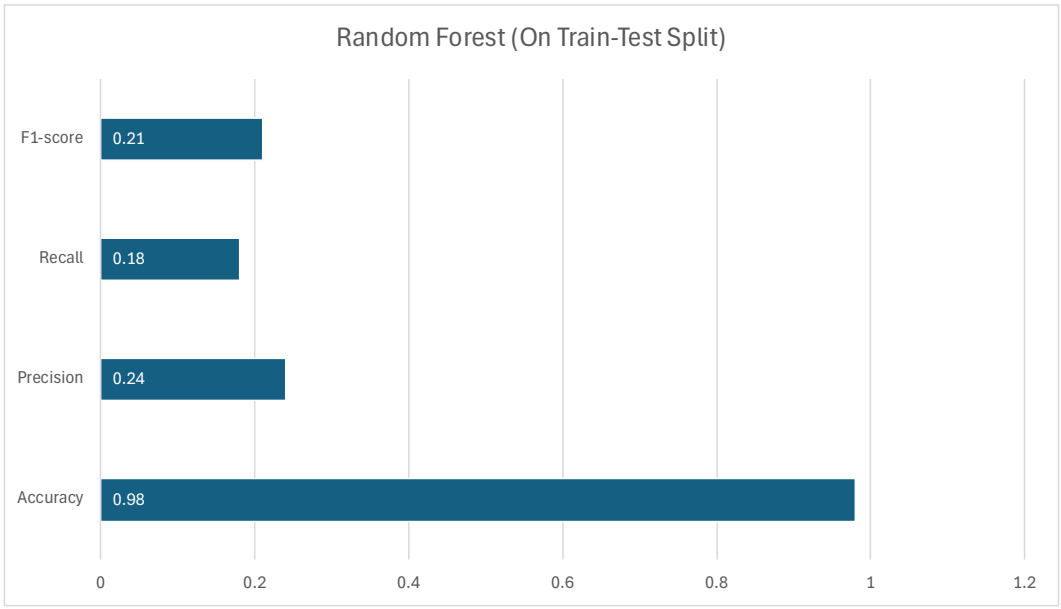


Fig. 10 The train-test split metrics for the deployed random forest model, which obtained the first set of data.

3.3 Machine Learning Model Output

For the two sets of predictions we made, we manually annotated each of the top 50 compounds by examining the literature regarding each compound. We also searched for promising natural compounds until rank 100, as they tend to be more interesting scientifically and better accepted by the public. **Table 1** shows the manually selected compounds from the first set of predictions, along with the rationales used for their selection. See **Supplementary Materials** for the full set of predictions.

Rank	Name	Probability	Rationale	References
1	**Betamethasone Valerate**	0.838	Atlhough it is a glucocorticoid cream, it is supposed to have a lower potency and so generally safer. It is also the top-ranked compound in the predictions. It is also a member of a distinct cluster among the top 60 drugs	MedChemExpress
7	**Decursinol angelate**	0.699	A natural compound from the roots of *Angelica gigas*. It seems to be in another cluster that's distinct from the rest. It also seems to inhibit LPS-Induced macrophage responses through the NFκB and MAPK signaling pathways. Also induces PKC activation and promotes apoptosis	MedChemExpress
12	**YK-3-237**	0.555	A synthetic activator of SIRT1. There is increaing interests in sirtuins because of their anti-aging effects. Popular natural compounds like resvertrol (found in grape skins) are thought to activate sirtuins. YK-3-237 itself seems to be a little studied compound, however.	MedChemExpress
13	**Ouabain**	0.547	Compound discovered to be senolytic in Smer-Bareto's paper so might have undiscovered anti-inflammatory effects as well	PubChem
20	**Chrysotoxine**	0.483	A natural bibenzyl compound that protects dopaminergic neurons from MPP(+)-induced toxicity, protects mitochondrial health, and inhibits NF-kB like the other natural compounds	MedChemExpress
21	**6-Demethoxytangeretin**	0.448	A naturally occurring flavonoid found in citrus plants (like the mandarin orange) that targets similar anti-cancer/anti-oxidative pathways, but it also seems to induce the CRE (cAMP Response Element)-dependent transcription pathway, which has been implicated in learning and memory (can stimulate neurite growth), glucose and lipid metabolism, and other long-term changes. But there hasn't been a lot of studies on 6-demethoxytangeretin itself	MedChemExpress
23	**Myrislignan**	0.418	A natural compound from nutmeg, which is a widely available and popular spice. Unlike the other natural compounds, the top search results for myrislignan reports on it can be anti-parasitic by disrupting the redox balance and mitochondrial functions of the parasitic protozoan T. gondii. However, it also seems to be able to reduce inflammation like the other natural compounds, by inhibiting NF-κB signaling and reducing IL-6, TNF-α, and NO production	MedChemExpress
25	**Pectolinarigenin**	0.405	Another compound from this cluster. Pectolinarigenin is a natural flavonoid primarily isolated from *Cirsium* (thistles), which is the national flower of Scotland. So I think it's appropriate that we study this in Edinburgh. It has been reported to reduce LPS-induced inflammation in astrocytes by blocking NFκB and MAPK activation and inhibiting IL-6 and IL-1β production and stimulate IL-10 release.	Ge et al. (2023)

			Pectolinarigenin is also a component of the herb Daji (Japanese thistle) from traditional Chinese medicine	
26	**Okanin**	0.403	This is a natural flavone that seems to be from another cluster as the natural compounds mentioned above (see the graph). It has been reported to attenuate LPS-induced microglial activation through inhibition of the TLR4/NF-κB signaling pathways	MedChemExpress
27	**Octahydrocurcumin**	0.401	See below	MedChemExpress
32	**Hexahydrocucurmin**	0.386	Both are metabolites of curcumin. Might be more biologically relevant than curcumin itself, which is mainly metabolized in the kidney and liver. But, individually, they may also have less diverse effects as curcumin. Might be worth trying both at the same time.	Izadi et al. (2024)
33	**WAY-272077**	0.382	A sphingosine kinase inhibitor with anti-inflammatory, antitumor and hemostatic effects. A study reported that sphingosine 1-phosphate (S1P) accumulation in neural cells causes neuroinflammation	PubChem
37	**Salvigenin**	0.374	An increasingly popular natural compound isolated from sage and related plants. Mainly studied for its anti-tumor effects. A 2024 studied showed that it protected liver cells against herbicide-induced toxicity by modulating the Nrf-2/Keap-1 and NF-κB pathways.	MedChemExpress
39	**L-165041**	0.372	A synthetic molecule that's located close to the curcumin metabolites on our map. It activates the PPARβ/δ receptor, which is being key target in treating metabolic disorders and can also reduce inflammation. It might be helpful to include this compound because it is very different from the rest.	MedChemExpress
42	**Hispidulin**	0.369	A natural compound that is most well-known for being able to allosterically bind to the benzodiazepine receptor and enhance GABAa signaling, thereby ameliorating anxiety and epilepsy. However, it also has anti-inflammatory effects and can cross the BBB readily	MedChemExpress
46	**Dimethylcurcumin**	0.362	This is a synthetic analog of curcumin that is supposed to be more stable in cellular environments than curcumin and may have enhanced anti-cancer properties in some cell lines.	Al Joseph et al. (2018)
48	**7,3',4'-Tri-O-methylluteolin**	0.361	Reduces inflammation in macrophages by inhibiting the mRNA expressions of inducible nitric oxide synthase (iNOS) and cyclooxygenase-2 (COX-2); Only *in vitro* studies yet.	MedChemExpress
50	**Phyllanthin**	0.351	A natural compound that's close to curcumin in this plot. It's reported to reduce inflammation via downregulating the NF-κB and AMPK/Nrf2 pathways. It also seems to protect mice against obesity (via upregulating insulin receptors), which seems rare among papers on plant-derived compounds. That might confer extra benefits because there's increasing discussion on how neurological disorders may in part be metabolic problems	MedChemExpress
58	**Prosapogenin A**	0.332	A natural compound that seems to be in a cluster distinct from the rest of the top predicted compounds. It can supposedly reduce inflammation in LPS-activated RAW264.7	Han et al. (2013)

			macrophages by inhibiting NO production and the MAPK/NF-κB pathway.	
71	**Caudatin**	0.312	A plant-derived steroid, which is unusual among our list of compounds. It seems to have anti-cancer and anti-inflammatory effects.	MedChemExpress
88	**Hydroxygenkwanin**	0.296	A natural compound found in *Daphne genkwa* "a deciduous shrub and one of the 50 fundamental herbs used in traditional Chinese medicine"	Li et al. (2022)

Table 1. Annotated first set of candidates selected among the top 100 predictions.

3.4 Common Hit Pathways

Using the transcriptomics data from the DRUG-seq of 42 hits, we analyzed the pathways implicated by the gene expression changes. 33 compounds had sufficient significantly ($p \leq 0.1$) altered genes to be analyzed by IPA. See **Table 2** for the 8 most commonly modulated pathways. See **Supplementary Materials** for the full table.

Drug Count	Most Commonly Upregulated Pathways	Drug Count	Most Commonly Downregulated Pathways	Drug Count	Pathways with Both Up- & Downregulated Genes
16	CLEAR Signaling Pathway	18	Mitochondrial Dysfunction	5	Class A/1 (Rhodopsin-like receptors)
15	Oxidative Phosphorylation	14	Parkinson's Signaling Pathway	4	Hematoma Resolution Signaling Pathway
15	Chaperone Mediated Autophagy Signaling Pathway	13	HER-2 Signaling in Breast Cancer	3	Docosahexaenoic Acid (DHA) Signaling
15	Neutrophil degranulation	11	Integrin Signaling	3	Neuroinflammation Signaling Pathway
14	Electron transport, ATP synthesis, and heat production by uncoupling proteins	11	Phagosome Formation	3	Xenobiotic Metabolism AHR Signaling Pathway
11	Post-translational protein phosphorylation	11	Actin Cytoskeleton Signaling	3	Breast Cancer Regulation by Stathmin1
11	Regulation of Insulin-like Growth Factor (IGF) transport and uptake by IGFBPs	11	Serotonin Receptor Signaling	3	Ferroptosis Signaling Pathway
10	Response to elevated platelet cytosolic Ca2+	11	Systemic Lupus Erythematosus in B Cell Signaling Pathway	2	Gap Junction Signaling

Table 2. The 8 most commonly modulated pathways containing 1) significantly upregulated genes 2) significantly downregulated genes 3) both up- and downregulated genes, as determined by Qiagen's IPA tool.

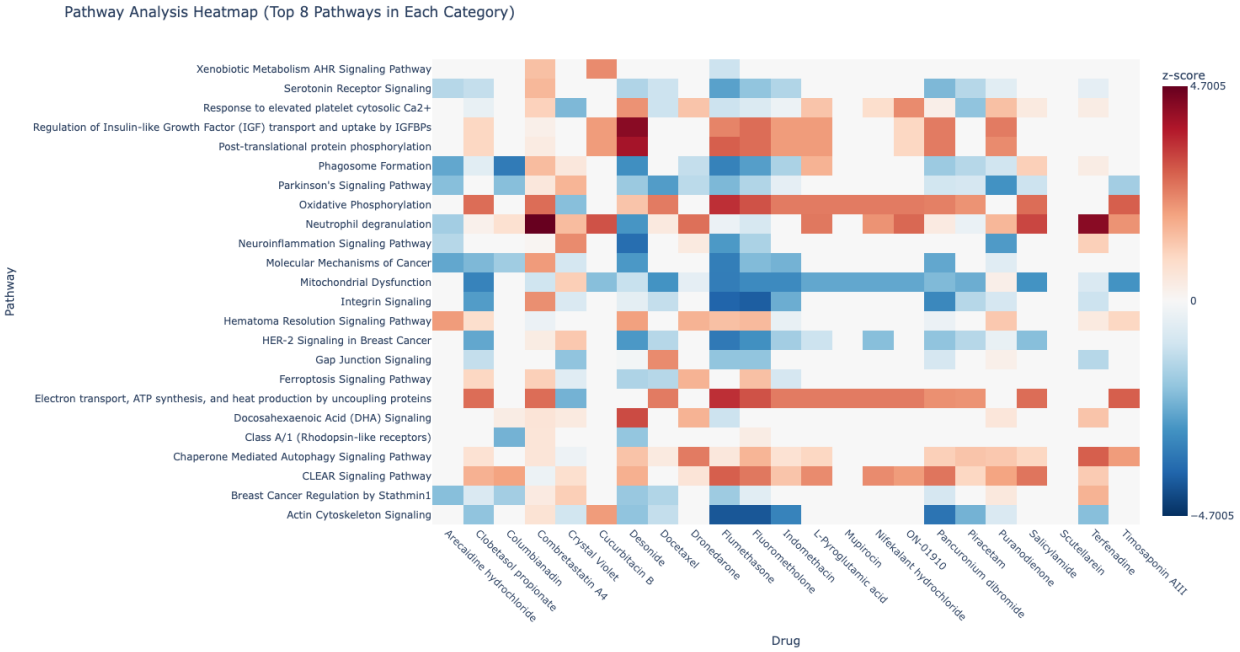


Fig. 11 A heatmap of the 8 most commonly modulated pathways in each category. Note, instead of tallying the modulation categorically (up, down, and both), the heatmap shows the z-scores outputted by IPA. A z-score of 0

means there are an equal number of significantly up- and downregulated genes in pathway. Note, only the compounds with changes in these pathways are shown.

4 Discussion

4.1 Model Evaluation & Predictions

First set of predictions: model trained on 75% of the entire training data

What constitutes "good" metrics varies across fields, but the metrics for our random forest model would likely be considered suboptimal for most applications. Specifically, our primary metric of interest, precision, averaged only 0.15 in cross-validation (**aFig. 8**) and 0.24 in the 75:25 train-test split evaluation conducted prior to deployment (**Fig. 10**). It should also be noted that the standard deviations were high, and the precision score (0.15 ± 0.106) can be as low as 0.044 in the worst-case scenario. Additionally, even though we selected the random forest model due to its marginally better average performance compared to the rest, the error bars have significant overlap.

Thus, the resulting ranking of predictions should be interpreted as approximate; for instance, a compound ranked 30th is not necessarily more likely to be a hit than one ranked 50th but should have more potential than one ranked 150th. Moreover, we recognized that our models were biased toward glucocorticoids. Therefore, as described in **Hit Selection**, we manually reviewed the literature to assess the potential of the highly ranked compounds and selected them accordingly. The resulting selection of 36 candidates was primarily influenced by our constraints on time and resources, rather than an exhaustion of viable options.

As expected, several glucocorticoids ranked highly in our predictions. While we generally excluded glucocorticoids from further consideration due to their well-documented properties and limited chemical novelty, we included one (betamethasone valerate) in our drug validation as a positive control. A literature review of the top 50 or so compounds identified numerous natural compounds with known anti-inflammatory properties, suggesting that the model is performing acceptably. For example, ranking 25th in the first set of predictions is pectolinarigenin, a natural flavonoid primarily isolated from thistles[1], especially of the genus *Cirsium*. It has been shown to exert anti-cancer and anti-inflammatory effects by inhibiting pathways such as NF-κB and MAPK. While there are not yet human studies on pectolinarigenin, the compound is found abundantly in the Chinese herb *Chromolaena odorata*, used in traditional medicine (N. Patel et al., 2024). In fact, most of the relevant compounds only have preclinical data, but that aligns with our goal of characterizing novel compounds.

Also ranked highly are octahydrocurcumin (27th) and hexahydrocurcumin (32th), both metabolites of curcumin, which is well-known for its anti-inflammatory properties and

[1] Which are, coincidentally, the national flower of Scotland, though no specific species is designated.

widely sold as a supplement. This observation lends additional support to the utility of our model.

An examination of the bottom ~30 predictions revealed predominantly unrelated compounds, though a few molecules with reported anti-inflammatory effects were still identified. This observation underscores the relatively low recall of our model, as well as the limited number of hits in our training data, which may not fully capture the diversity of anti-inflammatory molecules. Additionally, the majority of the compounds screened came from an anti-cancer library. Given the close relationship between inflammation and cancer (Korniluk et al., 2017) this may have influenced the results. For instance, inhibiting NF-κB can suppress both inflammation and tumor progression (Verzella et al., 2020). Consequently, it is possible that the use of an anti-cancer library inflated the number of anti-inflammatory compounds among both the top and bottom ranks.

Overall, our predictions seem sensible, and the pending experimental validation will further determine the success of our approach.

Second set of predictions: model trained on 100% of the validated data

As noted in **Data Source**, we are reasonably confident in all of our hits but were only able to confidently label 72 negatives in our training data. Therefore, to explore the outcome of using a "purer" dataset, we created a second set of predictions using the random forest model trained only on the validated data. Given the small size of the dataset, we opted not to follow the convention of withholding a testing set.

The CV evaluation using only the validated data produced markedly better metrics (**Fig. 9**). However, it is unclear whether these enhanced metrics reflect genuine model performance or are the result of overfitting and reduced variability in the training data. For that reason, even though the support vector machine model had an average precision of 0.98 in this context, we decided to continue using the random forest model for its more realistic scores.

While there are differences, the top-ranking compounds in the second set of predictions (see **Supplemental Materials**) include many that were also ranked highly in the first set, which reassures us of the model's stability.

It should also be noted that the predicted probabilities for the model trained on the validated data are generally much higher than those of the first model. In fact, the first set of predictions includes only 16 compounds with >0.5 probabilities, while the second set includes 1232 compounds (see **4** in **Appendix** for visualizations of the predictions). This difference was another reason we chose not to rely solely on the binary prediction outcome of the model but instead examined the rankings.

Considering both sets of predictions, we selected 36 compounds (see **5** in **Appendix** for the list) for experimental validation.

We have performed some preliminary validation on several of the candidates with immunofluorescence microscopy. However, the resulting images were uninterpretable due to unknown issues with the cell culture. For the purpose of the thesis, some representative images are featured in the **Appendix** (item **6**)**.**

4.2 Common Pathways

IPA analysis of the differentially expressed genes from our training hits revealed several intriguing pathways shared across multiple compounds (**Table 1**). The most commonly upregulated pathway, observed in 16 hits, is the Coordinated Lysosomal Expression and Regulation (CLEAR) pathway. This gene network plays a critical role in cellular homeostasis by regulating a variety of lysosome-associated processes, including autophagy, endocytosis, exocytosis, membrane repair, immune response, and phagocytosis (Palmieri et al., 2011).

The **Introduction** highlighted the importance of phagocytosis in normal microglial functions, so this observation is consistent with current understanding and provides additional validation for the 16 hits. While the CLEAR pathway predominantly regulates autophagy, the phagocytosis and autophagy pathways converge on the lysosomal degradation and recycling process. Hence, changes in one pathway influence the other.

On the other hand, the most commonly downregulated pathway, observed in 16 hits, is related to mitochondrial dysfunction. There is growing recognition of the role mitochondrial dynamics and bioenergetics play in neurodegenerative diseases. This connection is exemplified by the use of rotenone, a complex I inhibitor, to model Parkinson's disease in animals (Johri & Beal, 2012). The late age of onset for most neurodegenerative conditions provides another hint at the role of energy metabolism.

Autophagy, upregulated by many of our hits, is a major mechanism for maintaining mitochondrial health. The autophagic removal of damaged mitochondrial, known as mitophagy, maintains the integrity of the mitochondrial population. Failure to do so leads to an accumulation of dysfunctional mitochondria, which triggers apoptosis (Ashrafi & Schwarz, 2013).

Emerging evidence is also illuminating microglia-specific metabolic changes in neurodegeneration. Homeostatic microglia primarily rely on oxidative phosphorylation (OXPHOS)—a mitochondria-dependent process—for energy production, whereas activated microglia rely more on glycolysis. This increased glucose metabolism supports the elevated production of reactive oxygen and nitrogen species (RONS) in inflammation and facilitates the assembly of inflammasomes, which activate caspase-1 and promote the secretion of pro-inflammatory cytokines, such as IL-1β and IL-18 (Ghosh et al., 2018). The shifted metabolism can also induce inflammation by depleting OXPHOS substrates, which seems to upregulate the pro-inflammatory NF-κB activity (Shen et al., 2017). Therefore, the fact that 16 of our training hits ameliorate mitochondrial dysfunction is a promising sign.

A seemingly contradictory observation in the pathway analysis is that 11 of the hits downregulated phagosome formation (**Table 2**). This seems to contradict the upregulation in the CLEAR pathway and the understanding that phagocytosis should be increased in homeostatic microglia. If this is not a quirk of the *in vitro* model, then one should consider the possibility that anti-inflammatory actions and normal phagocytic functions can be decoupled. This also suggests a diversity in mechanisms of action among the hits. Thus, future therapeutic research may consider combining two or more compounds with complimentary effects.

Interestingly, many of the commonly altered pathways—mitochondrial function, OXPHOS, autophagy, IGF-1 signaling—are also modulated Caloric restriction (CR) and the ketogenic diet (KD, a diet that avoids carbohydrates) (Maalouf et al., 2009). Aging research has pinpointed energy metabolism as a central determinant of longevity, beginning with the discovery that *Caenorhabditis elegans* carrying a partial loss-of-function mutation in the *daf-2* gene live twice as long as wild-type counterparts (Kenyon et al., 1993). The mammalian orthologs of *daf-2* are the insulin receptor (IR) and insulin-like growth factor 1 receptor (IGF-1R). Subsequent studies have shown that suppressing nutrient-sensing pathways, particularly the insulin/insulin-like growth factor signaling (IIS) and mTOR pathways, can extend lifespan across a range of species, including yeast, worms, flies, and mice (Dobrenel et al., 2016).

CR and KD inhibit IIS and mTOR and have been shown to exert beneficial effects in humans. These include reversing type II diabetes, improving biomarkers associated with aging and chronic diseases, and exhibiting well-documented anti-inflammatory properties; Notably, KD has been prescribed to treat drug-resistant epilepsy (Napoleão et al., 2021). The primary effect of inhibiting mTOR is shifting cells from an overall anabolic state to an overall catabolic state, a process closely tied to autophagy (Dobrenel et al., 2016); under nutrient-rich conditions, mTOR inhibits the CLEAR pathway (Pan & Valapala, 2023). Therefore, it is noteworthy that our hits modulate many of the pathways also modulated by CR and KD, exerting positive effects through improving lysosomal degradation and bioenergetics. However, KD and CR studies typically do not report increased microglial phagocytosis as an effect—although it may still be indirectly supported by reduced inflammation and improved bioenergetics. Therefore, improved health outcomes are not necessarily incompatible with the downregulated phagosome formation observed in the 6 hits.

It will be interesting to see whether the virtual screening's hits share the same pathways modulated by the current hits. The results will advance our understanding of the genetic regulation of microglial health.

4.3 Speculations

As the results of our new candidates' validation are still pending, the potential outcomes remain speculative. Firstly, even one successfully validated candidate from the ML approach

would be an improvement over the traditional screening we performed, raising the working hit rate from 0.008 to 0.027. If two candidates are confirmed, the working hit rate would achieve a sixfold increase, matching the improvement reported by Wong et al. (2023) in senolytic identification.

Should more candidates prove viable, the results will also inform us of this approach's ability to discover chemically diverse compounds. By analyzing the structural similarity of newly confirmed hits, we can assess the model's generalizability: if they all closely resemble the current hits, then it would indicate that, in the context of an imbalanced dataset and drugs of this nature, the combination of physiochemical descriptors and traditional ML algorithms cannot generalize far beyond the training structures; conversely, if the new hits include structurally diverse compounds, it would suggest that this approach could uncover some hidden underlying patterns within the physicochemical descriptors that are associated with functional properties. These patterns might link seemingly dissimilar molecules through shared chemical features, such as specific combinations of functional groups and stereochemistry, indicating the model's ability to generalize beyond the apparent structures.

Should we discover sufficient new hits, the results will also provide insights into the model's ability to generalize across underlying mechanisms. If the pathways modulated by the new hits are similar to those of the current hits, then this approach may not be suitable for discovering mechanistically novel compounds, although it could also mean that there is a limited set of pathways capable of improving microglial health. Conversely, discovering novel pathways would indicate the model's capacity to predict the outcome independent of specific mechanisms. However, as ML models can only identify compounds that share some commonalities in the data with the training compounds, any novel pathways discovered must be in some ways linked to shared chemical features. Thus, if the new hits modulate additional pathways, it might indicate that these pathways are correlated or co-regulated with those already identified, or that shared chemical features among the compounds can affect distinct but related mechanisms.

4.4 Limitations

Data quality

While hit validation is still pending, it remains difficult to fully assess the success of our method and identify the potential causes of its shortcomings. However, certain limitations are already apparent, the most notable being the comparatively poor performance metrics. The most likely explanation was the noisy screening data combined with the limited positive examples. Initially, we selected only 42 hits, and our models struggled to achieve a non-zero recall score. Expanding the hits to 48 made it significantly easier to obtain recall scores of at least 5-10%, underscoring the importance of even small increases in datapoints when working with a highly imbalanced dataset, particularly when chemical diversity is also a priority. Even though, looking at **Fig. 7**, our hits seem chemically diverse, each cluster only has one or two examples for the model to learn from, which likely meant that a larger sample size was needed to account for the variability. If our goal was to identify chemically similar

molecules, a dataset of this size might be sufficient, as demonstrated by the ease of recognizing glucocorticoids in our attempts.

Additionally, the significantly better metrics observed in the CVs performed solely on validated data suggest that false negatives in the dataset had a substantial impact. However, the improvement may also be primarily attributed to the smaller dataset size and reduced chemical diversity, making it difficult to determine with confidence which model would generalize better to unseen data. Moving forward, this issue could be mitigated by prioritizing data quality over quantity—a mid-sized dataset (e.g., ~2,500 datapoints for this application) with accurate labels would likely produce much better models than a larger dataset contaminated with false labels.

Molecular representations

Exploring alternative methods for representing molecules could also have produced different results; however, the limited timeframe of this project constrained our ability to do so. As discussed in **Featurization and Data Cleaning**, in addition to using physiochemical descriptors—which is considered a more traditional approach—one could also employ graph-based deep learning techniques, or Graph Neural Networks (GNNs), as seen in Wong et al. (2023). Although we initially attempted to implement GNNs, the approach was abandoned following poor early results. A more comprehensive exploration of this method might have led to different outcomes. However, while deep learning techniques are considered newer, there is an on-going debate about their effectiveness relative to traditional descriptor-based, with recent studies suggesting that their performance may be context-dependent (Baptista et al., 2022; Jiang et al., 2021). Now that we have performed feature selection, we are better equipped to explore a hybrid approach—integrating chemical descriptors with GNNs—as a way to leverage the strengths of both methods. Additionally, a recent paper by Snyder et al. (2024) proposed a rule-of-thumb for model selection based on data size and diversity. Their analysis suggests that, in cases such as ours, where there are few data points in the target class, few-shot learning (FSLC) models may offer the best generalization.

Heterogeneity in mechanism of action

Our project was inspired by the successful application of ML in identifying senolytic drugs (Smer-Barreto et al., 2023; Wong et al., 2023). However, the known senolytics generally act by suppressing pro-apoptotic pathways and the referenced studies labeled compounds based solely on their ability to induce senescent cell death—a relatively straightforward measure. Consequently, their target data may have a more uniform mechanism of action. In contrast, there may be more heterogeneous paths to improving apparent microglial health, as suggested by our gene expression analysis (**Table 2**), where the most commonly modulated pathway is only shared by 18 out of the 33 analyzed hits, with some seeming to downregulate phagocytosis, a process typically associated with normal microglial function. Therefore, we may be attempting to extrapolate based on a small number of compounds with diverse underlying actions. Moreover, we conducted our screening by first examining

morphological changes, then inflammatory markers. These phenotypes exhibit more gradation compared to the binary outcome of cell death, which may introduce additional variability in our data. Thus, this probable heterogeneity further highlights the need for a large sample size in our application.

4.5 Future Directions

In addition to the potential improvements to our model described above, we are also interested in the mechanisms that promote microglial health. Should we obtain validated hits, their transcriptomic profiles can be compared with our current pathways. Do successful hits generally act on similar pathways, or can they achieve the same effects through different mechanisms? Can compounds acting on different pathways improve microglial health in the same manner, or do they each improve different aspects of health? These questions could provide critical insights into the diversity of mechanisms underlying microglial function.

Future research will also benefit from a more precise and comprehensive definition of microglial health. As described in the **Introduction**, microglia exist along a broad continuum of states, exhibiting heterogenous phenotypes even in diseased conditions. Thus, metrics such as morphology, inflammation levels, and even phagocytic ability may be insufficient to fully characterize microglial health, which may have more than a singular manifestation. For instance, as discussed in **Common Pathways**, CR and KD exert well-documented anti-inflammatory and neuroprotective effects, while not reported to directly upregulate phagocytosis. Therefore, it is possible that interventions may restore selective aspects of microglial health without necessarily reinstating other homeostatic features, such as phagocytic capacity.

One avenue for measuring homeostasis that warrants further exploration is to examine the known microglial (homeostatic) signature genes, which are known to be lost under *in vitro* and neurodegenerative conditions (Butovsky & Weiner, 2018). By combining these markers with our current measures, we may develop more comprehensive indicators of microglial health and identify pathways that are critical for maintaining or restoring homeostasis.

Aside from their diversity in functional states, microglia also exhibit distinct subpopulations. Currently, at least two well-studied lineages of microglia have been identified. Most microglia originate from progenitor cells in the embryonic yolk sac that migrate directly into the developing brain. However, a small population, distinguishable by the transcription factor *Hoxb8*, first migrates to the fetal liver for further expansion before entering the brain. These *Hoxb8* microglia have unique localization in the brain and seem to be involved in anxiety- and OCD-related neural circuits. Notably, selectively ablating these *Hoxb8* microglia induces OCD-like symptoms in mice, highlighting their specialized role (Tränkner et al., 2019). Given this observation, it would be interesting to see if different subpopulations respond differently to a given compound. Being more specific with our microglia lineage may also help reduce variability in our screening.

The proposed explorations will help advance our goal of delineating homeostatic pathways and developing effective therapeutics, as well as improving our understanding of effective ML approaches for drug discovery.

5 References

Ashrafi, G., & Schwarz, T. L. (2013). The pathways of mitophagy for quality control and clearance of mitochondria. In *Cell Death and Differentiation* (Vol. 20, Issue 1). https://doi.org/10.1038/cdd.2012.81

Augusto-Oliveira, M., Arrifano, G. P., Lopes-Araújo, A., Santos-Sacramento, L., Takeda, P. Y., Anthony, D. C., Malva, J. O., & Crespo-Lopez, M. E. (2019). What do microglia really do in healthy adult brain? In *Cells* (Vol. 8, Issue 10). https://doi.org/10.3390/cells8101293

Badimon, A., Strasburger, H. J., Ayata, P., Chen, X., Nair, A., Ikegami, A., Hwang, P., Chan, A. T., Graves, S. M., Uweru, J. O., Ledderose, C., Kutlu, M. G., Wheeler, M. A., Kahan, A., Ishikawa, M., Wang, Y. C., Loh, Y. H. E., Jiang, J. X., Surmeier, D. J., … Schaefer, A. (2020). Negative feedback control of neuronal activity by microglia. *Nature*, *586*(7829). https://doi.org/10.1038/s41586-020-2777-8

Baptista, D., Correia, J., Pereira, B., & Rocha, M. (2022). Evaluating molecular representations in machine learning models for drug response prediction and interpretability. *Journal of Integrative Bioinformatics*, *19*(3). https://doi.org/10.1515/jib-2022-0006

Baxter, P. S., Dando, O., Emelianova, K., He, X., McKay, S., Hardingham, G. E., & Qiu, J. (2021). Microglial identity and inflammatory responses are controlled by the combined effects of neurons and astrocytes. *Cell Reports*, *34*(12). https://doi.org/10.1016/j.celrep.2021.108882

Bohlen, C. J., Friedman, B. A., Dejanovic, B., & Sheng, M. (2019). Microglia in Brain Development, Homeostasis, and Neurodegeneration. In *Annual Review of Genetics* (Vol. 53). https://doi.org/10.1146/annurev-genet-112618-043515

Brown, A. S., Begg, M. D., Gravenstein, S., Schaefer, C. A., Wyatt, R. J., Bresnahan, M., Babulas, V. P., & Susser, E. S. (2004). Serologic evidence of prenatal influenza in the etiology of schizophrenia. *Archives of General Psychiatry*, *61*(8). https://doi.org/10.1001/archpsyc.61.8.774

Butovsky, O., & Weiner, H. L. (2018). Microglial signatures and their role in health and disease. In *Nature Reviews Neuroscience* (Vol. 19, Issue 10). https://doi.org/10.1038/s41583-018-0057-5

Cadiz, M. P., Jensen, T. D., Sens, J. P., Zhu, K., Song, W. M., Zhang, B., Ebbert, M., Chang, R., & Fryer, J. D. (2022). Culture shock: microglial heterogeneity, activation, and disrupted single-cell microglial networks in vitro. *Molecular Neurodegeneration*, *17*(1). https://doi.org/10.1186/s13024-022-00531-1

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939785

Cunningham, C. L., Martínez-Cerdeño, V., & Noctor, S. C. (2013). Microglia regulate the number of neural precursor cells in the developing cerebral cortex. *Journal of Neuroscience*, *33*(10). https://doi.org/10.1523/JNEUROSCI.3441-12.2013

Davalos, D., Grutzendler, J., Yang, G., Kim, J. V., Zuo, Y., Jung, S., Littman, D. R., Dustin, M. L., & Gan, W. B. (2005). ATP mediates rapid microglial response to local brain injury in vivo. *Nature Neuroscience*, *8*(6). https://doi.org/10.1038/nn1472

Deng, J., Yang, Z., Ojima, I., Samaras, D., & Wang, F. (2022). Artificial intelligence in drug discovery: Applications and techniques. In *Briefings in Bioinformatics* (Vol. 23, Issue 1). https://doi.org/10.1093/bib/bbab430

Dobrenel, T., Caldana, C., Hanson, J., Robaglia, C., Vincentz, M., Veit, B., & Meyer, C. (2016). TOR Signaling and Nutrient Sensing. In *Annual Review of Plant Biology* (Vol. 67). https://doi.org/10.1146/annurev-arplant-043014-114648

Drouin, A., Annavajjala, A., & Wright, R. (2024). *pulearn*. Https://Pulearn.Github.Io/Pulearn/.

Elkan, C., & Noto, K. (2008). Learning classifiers from only positive and unlabeled data. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/1401890.1401920

Faust, T. E., Gunner, G., & Schafer, D. P. (2021). Mechanisms governing activity-dependent synaptic pruning in the developing mammalian CNS. In *Nature Reviews Neuroscience* (Vol. 22, Issue 11). https://doi.org/10.1038/s41583-021-00507-y

Ghosh, S., Castillo, E., Frias, E. S., & Swanson, R. A. (2018). Bioenergetic regulation of microglia. In *GLIA* (Vol. 66, Issue 6). https://doi.org/10.1002/glia.23271

Ginhoux, F., Greter, M., Leboeuf, M., Nandi, S., See, P., Gokhan, S., Mehler, M. F., Conway, S. J., Ng, L. G., Stanley, E. R., Samokhvalov, I. M., & Merad, M. (2010). Fate mapping analysis reveals that adult microglia derive from primitive macrophages. *Science*, *330*(6005). https://doi.org/10.1126/science.1194637

Guo, K., Luo, J., Feng, D., Wu, L., Wang, X., Xia, L., Tao, K., Wu, X., Cui, W., He, Y., Wang, B., Zhao, Z., & Zhang, Z. (2021). Single-Cell RNA Sequencing With Combined Use of Bulk RNA Sequencing to Reveal Cell Heterogeneity and Molecular Changes at Acute Stage of Ischemic Stroke in Mouse Cortex Penumbra Area. *Frontiers in Cell and Developmental Biology*, *9*. https://doi.org/10.3389/fcell.2021.624711

Hagberg, H., Gressens, P., & Mallard, C. (2012). Inflammation during fetal and neonatal life: Implications for neurologic and neuropsychiatric disease in children and adults. In *Annals of Neurology* (Vol. 71, Issue 4). https://doi.org/10.1002/ana.22620

Hanisch, U. K. (2002). Microglia as a source and target of cytokines. In *GLIA* (Vol. 40, Issue 2). https://doi.org/10.1002/glia.10161

Haupt, M., Gerner, S. T., & Doeppner, T. R. (2024). The dual role of microglia in ischemic stroke and its modulation via extracellular vesicles and stem cells. *Neuroprotection*, *2*(1), 4–15. https://doi.org/10.1002/nep3.39

Heid, E., Greenman, K. P., Chung, Y., Li, S. C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., & McGill, C. J. (2024). Chemprop: A Machine Learning Package for Chemical

Property Prediction. *Journal of Chemical Information and Modeling*, *64*(1). https://doi.org/10.1021/acs.jcim.3c01250

Hopp, S. C., Lin, Y., Oakley, D., Roe, A. D., Devos, S. L., Hanlon, D., & Hyman, B. T. (2018). The role of microglia in processing and spreading of bioactive tau seeds in Alzheimer's disease. *Journal of Neuroinflammation*, *15*(1). https://doi.org/10.1186/s12974-018-1309-z

Iadecola, C., & Anrather, J. (2011). The immunology of stroke: From mechanisms to translation. In *Nature Medicine* (Vol. 17, Issue 7). https://doi.org/10.1038/nm.2399

Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-020-00479-8

Johri, A., & Beal, M. F. (2012). Mitochondrial Dysfunction in Neurodegenerative Diseases. *Journal of Pharmacology and Experimental Therapeutics*, *342*(3), 619–630. https://doi.org/10.1124/jpet.112.192138

Kato, G., Inada, H., Wake, H., Akiyoshi, R., Miyamoto, A., Eto, K., Ishikawa, T., Moorhouse, A. J., Strassman, A. M., & Nabekura, J. (2016). Microglial contact prevents excess depolarization and rescues neurons from excitotoxicity. *ENeuro*, *3*(3). https://doi.org/10.1523/ENEURO.0004-16.2016

Kenyon, C., Chang, J., Gensch, E., Rudner, A., & Tabtiang, R. (1993). A C. elegans mutant that lives twice as long as wild type. *Nature*, *366*(6454). https://doi.org/10.1038/366461a0

Korniluk, A., Koper, O., Kemona, H., & Dymicka-Piekarska, V. (2017). From inflammation to cancer. In *Irish Journal of Medical Science* (Vol. 186, Issue 1). https://doi.org/10.1007/s11845-016-1464-0

Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., Friederich, P., Gaudin, T., Gayle, A. A., Jablonka, K. M., Lameiro, R. F., Lemm, D., Lo, A., Moosavi, S. M., Nápoles-Duarte, J. M., Nigam, A. K., Pollice, R., Rajan, K., Schatzschneider, U., … Aspuru-Guzik, A. (2022). SELFIES and the future of molecular string representations. In *Patterns* (Vol. 3, Issue 10). https://doi.org/10.1016/j.patter.2022.100588

Kurematsu, C., Sawada, M., Ohmuraya, M., Tanaka, M., Kuboyama, K., Ogino, T., Matsumoto, M., Oishi, H., Inada, H., Ishido, Y., Sakakibara, Y., Nguyen, H. B., Thai, T. Q., Kohsaka, S., Ohno, N., Yamada, M. K., Asai, M., Sokabe, M., Nabekura, J., … Sawamoto, K. (2022). Synaptic pruning of murine adult-born neurons by microglia depends on phosphatidylserine. *Journal of Experimental Medicine*, *219*(4). https://doi.org/10.1084/jem.20202304

Lane, C. A., Hardy, J., & Schott, J. M. (2018). Alzheimer's disease. In *European Journal of Neurology* (Vol. 25, Issue 1). https://doi.org/10.1111/ene.13439

Li, T., Zhao, J., Xie, W., Yuan, W., Guo, J., Pang, S., Gan, W. B., Gómez-Nicola, D., & Zhang, S. (2021). Specific depletion of resident microglia in the early stage of stroke reduces cerebral ischemic damage. *Journal of Neuroinflammation*, *18*(1). https://doi.org/10.1186/s12974-021-02127-w

Liu, P. P., Xie, Y., Meng, X. Y., & Kang, J. S. (2019). History and progress of hypotheses and clinical trials for alzheimer's disease. In *Signal Transduction and Targeted Therapy* (Vol. 4, Issue 1). https://doi.org/10.1038/s41392-019-0063-8

Louppe, G., Head, T., Kumar, M., Nahrstaedt, H., & Shcherbatyi, L. (2022). *scikit-optimize* . Https://Github.Com/Scikit-Optimize.

Lv, Q. K., Tao, K. X., Wang, X. B., Yao, X. Y., Pang, M. Z., Liu, J. Y., Wang, F., & Liu, C. F. (2023). Role of α-synuclein in microglia: autophagy and phagocytosis balance neuroinflammation in Parkinson's disease. In *Inflammation Research* (Vol. 72, Issue 3). https://doi.org/10.1007/s00011-022-01676-x

Maalouf, M., Rho, J. M., & Mattson, M. P. (2009). The neuroprotective properties of calorie restriction, the ketogenic diet, and ketone bodies. In *Brain Research Reviews* (Vol. 59, Issue 2). https://doi.org/10.1016/j.brainresrev.2008.09.002

Marino Lee, S., Hudobenko, J., McCullough, L. D., & Chauhan, A. (2021). Microglia depletion increase brain injury after acute ischemic stroke in aged mice. *Experimental Neurology*, *336*. https://doi.org/10.1016/j.expneurol.2020.113530

Morgan, J. T., Chana, G., Pardo, C. A., Achim, C., Semendeferi, K., Buckwalter, J., Courchesne, E., & Everall, I. P. (2010). Microglial activation and increased microglial density observed in the dorsolateral prefrontal cortex in autism. *Biological Psychiatry*, *68*(4). https://doi.org/10.1016/j.biopsych.2010.05.024

Napoleão, A., Fernandes, L., Miranda, C., & Marum, A. P. (2021). Effects of calorie restriction on health span and insulin resistance: classic calorie restriction diet vs. Ketosis-inducing diet. In *Nutrients* (Vol. 13, Issue 4). https://doi.org/10.3390/nu13041302

Neumann, H., Kotter, M. R., & Franklin, R. J. M. (2009). Debris clearance by microglia: An essential link between degeneration and regeneration. In *Brain* (Vol. 132, Issue 2). https://doi.org/10.1093/brain/awn109

Nguyen, P. T., Dorman, L. C., Pan, S., Vainchtein, I. D., Han, R. T., Nakao-Inoue, H., Taloma, S. E., Barron, J. J., Molofsky, A. B., Kheirbek, M. A., & Molofsky, A. V. (2020). Microglial Remodeling of the Extracellular Matrix Promotes Synapse Plasticity. *Cell*, *182*(2). https://doi.org/10.1016/j.cell.2020.05.050

Nimmerjahn, A., Kirchhoff, F., & Helmchen, F. (2005). Neuroscience: Resting microglial cells are highly dynamic surveillants of brain parenchyma in vivo. *Science*, *308*(5726). https://doi.org/10.1126/science.1110647

Palmieri, M., Impey, S., Kang, H., di Ronza, A., Pelz, C., Sardiello, M., & Ballabio, A. (2011). Characterization of the CLEAR network reveals an integrated control of cellular clearance pathways. *Human Molecular Genetics*, *20*(19). https://doi.org/10.1093/hmg/ddr306

Pan, H. Y., & Valapala, M. (2023). Role of TFEB in Diseases Associated with Lysosomal Dysfunction. In *Advances in Experimental Medicine and Biology* (Vol. 1415). https://doi.org/10.1007/978-3-031-27681-1_46

Paolicelli, R. C., Bolasco, G., Pagani, F., Maggi, L., Scianni, M., Panzanelli, P., Giustetto, M., Ferreira, T. A., Guiducci, E., Dumas, L., Ragozzino, D., & Gross, C. T. (2011). Synaptic pruning by microglia is necessary for normal brain development. *Science*, *333*(6048). https://doi.org/10.1126/science.1202529

Paolicelli, R. C., Sierra, A., Stevens, B., Tremblay, M. E., Aguzzi, A., Ajami, B., Amit, I., Audinat, E., Bechmann, I., Bennett, M., Bennett, F., Bessis, A., Biber, K., Bilbo, S., Blurton-Jones, M., Boddeke, E., Brites, D., Brône, B., Brown, G. C., … Wyss-Coray, T. (2022). Microglia

states and nomenclature: A field at its crossroads. In *Neuron* (Vol. 110, Issue 21). https://doi.org/10.1016/j.neuron.2022.10.020

Parhizkar, S., Arzberger, T., Brendel, M., Kleinberger, G., Deussing, M., Focke, C., Nuscher, B., Xiong, M., Ghasemigharagoz, A., Katzmarski, N., Krasemann, S., Lichtenthaler, S. F., Müller, S. A., Colombo, A., Monasor, L. S., Tahirovic, S., Herms, J., Willem, M., Pettkus, N., … Haass, C. (2019). Loss of TREM2 function increases amyloid seeding but reduces plaque-associated ApoE. *Nature Neuroscience*, *22*(2). https://doi.org/10.1038/s41593-018-0296-9

Parkhurst, C. N., Yang, G., Ninan, I., Savas, J. N., Yates, J. R., Lafaille, J. J., Hempstead, B. L., Littman, D. R., & Gan, W. B. (2013). Microglia promote learning-dependent synapse formation through brain-derived neurotrophic factor. *Cell*, *155*(7). https://doi.org/10.1016/j.cell.2013.11.030

Patel, A. R., Ritzel, R., McCullough, L. D., & Liu, F. (2013). Microglia and ischemic stroke: A double-edged sword. In *International Journal of Physiology, Pathophysiology and Pharmacology* (Vol. 5, Issue 2).

Patel, N., Kulshrestha, R., Bhat, A. A., Mishra, R., Singla, N., Gilhotra, R., & Gupta, G. (2024). Pectolinarigenin and its derivatives: Bridging the gap between chemical properties and pharmacological applications. In *Pharmacological Research - Modern Chinese Medicine* (Vol. 10). https://doi.org/10.1016/j.prmcm.2024.100378

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Cournapeau, D., Brucher, M., & Perrot, M. (2011). Scikit-learn: Machine Learning in Python Pedregosa, Varoquaux, Gramfort et al. *Journal of Machine Learning Research*, *12*.

Pizzorusso, T., Medini, P., Berardi, N., Chierzi, S., Fawcett, J. W., & Maffei, L. (2002). Reactivation of ocular dominance plasticity in the adult visual cortex. *Science*, *298*(5596). https://doi.org/10.1126/science.1072699

Qiu, J., Dando, O., Baxter, P. S., Hasel, P., Heron, S., Simpson, T. I., & Hardingham, G. E. (2018). Mixed-species RNA-seq for elucidation of non-cell-autonomous control of gene transcription. *Nature Protocols*, *13*(10). https://doi.org/10.1038/s41596-018-0029-2

Sakai, J. (2020). How synaptic pruning shapes neural wiring during development and, possibly, in disease. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(28). https://doi.org/10.1073/pnas.2010281117

Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease. *The Lancet*, *397*(10284), 1577–1590. https://doi.org/10.1016/S0140-6736(20)32205-4

Scikit-learn. (2024). *StandardScaler*. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Shen, Y., Kapfhamer, D., Minnella, A. M., Kim, J. E., Won, S. J., Chen, Y., Huang, Y., Low, L. H., Massa, S. M., & Swanson, R. A. (2017). Bioenergetic state regulates innate inflammatory responses through the transcriptional co-repressor CtBP. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00707-0

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, *23*(6). https://doi.org/10.3390/e23060759

Smer-Barreto, V., Quintanilla, A., Elliott, R. J. R., Dawson, J. C., Sun, J., Campa, V. M., Lorente-Macías, Á., Unciti-Broceta, A., Carragher, N. O., Acosta, J. C., & Oyarzún, D. A. (2023). Discovery of senolytics using machine learning. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-39120-1

Snyder, S. H., Vignaux, P. A., Ozalp, M. K., Gerlach, J., Puhl, A. C., Lane, T. R., Corbett, J., Urbina, F., & Ekins, S. (2024). The Goldilocks paradigm: comparing classical machine learning, large language models, and few-shot learning for drug discovery applications. *Communications Chemistry*, *7*(1), 134. https://doi.org/10.1038/s42004-024-01220-4

Stefanova, N. (2022). Microglia in Parkinson's Disease. In *Journal of Parkinson's Disease* (Vol. 12). https://doi.org/10.3233/JPD-223237

Strackeljan, L., Baczynska, E., Cangalaya, C., Baidoe-Ansah, D., Wlodarczyk, J., Kaushik, R., & Dityatev, A. (2021). Microglia depletion-induced remodeling of extracellular matrix and excitatory synapses in the hippocampus of adult mice. *Cells*, *10*(8). https://doi.org/10.3390/cells10081862

Tränkner, D., Boulet, A., Peden, E., Focht, R., Van Deren, D., & Capecchi, M. (2019). A Microglia Sublineage Protects from Sex-Linked Anxiety Symptoms and Obsessive Compulsion. *Cell Reports*, *29*(4). https://doi.org/10.1016/j.celrep.2019.09.045

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2). https://doi.org/10.1007/s10994-019-05855-6

Verzella, D., Pescatore, A., Capece, D., Vecchiotti, D., Ursini, M. V., Franzoso, G., Alesse, E., & Zazzeroni, F. (2020). Life, death, and autophagy in cancer: NF-κB turns up everywhere. In *Cell Death and Disease* (Vol. 11, Issue 3). https://doi.org/10.1038/s41419-020-2399-y

Wang, M. Y., Liu, W. J., Wu, L. Y., Wang, G., Zhang, C. L., & Liu, J. (2023). The Research Progress in Transforming Growth Factor-β2. In *Cells* (Vol. 12, Issue 23). https://doi.org/10.3390/cells12232739

Weininger, D., Weininger, A., & Weininger, J. L. (1988). SMILES (Simplified Molecular Input Line Entry System). *J. Chem. Inf. Comput. Sci*, *28*.

Wolf, S. A., Boddeke, H. W. G. M., & Kettenmann, H. (2017). Microglia in Physiology and Disease. In *Annual Review of Physiology* (Vol. 79). https://doi.org/10.1146/annurev-physiol-022516-034406

Wong, F., Omori, S., Donghia, N. M., Zheng, E. J., & Collins, J. J. (2023). Discovering small-molecule senolytics with deep neural networks. *Nature Aging*, *3*(6). https://doi.org/10.1038/s43587-023-00415-z

Wu, Y., & Hirschi, K. K. (2021). Tissue-Resident Macrophage Development and Function. In *Frontiers in Cell and Developmental Biology* (Vol. 8). https://doi.org/10.3389/fcell.2020.617879

Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., Hsieh, C. Y., & Hou, T. (2023). Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-38192-3

6 Appendix

1. 64 Features Selected by Recursive Feature Elimination

MaxAbsEStateIndex, MaxEStateIndex, MinAbsEStateIndex, MinEStateIndex, qed, SPS, MolWt, HeavyAtomMolWt, ExactMolWt, FpDensityMorgan1, FpDensityMorgan2, FpDensityMorgan3, AvgIpc, BalabanJ, BertzCT, Chi0, Chi0n, Chi0v, Chi1, Chi1n, Chi1v, Chi2n, Chi2v, Chi3n, Chi3v, Chi4n, Chi4v, HallKierAlpha, Kappa1, Kappa2, Kappa3, PEOE_VSA1, PEOE_VSA11, PEOE_VSA3, PEOE_VSA6, PEOE_VSA7, PEOE_VSA8, SMR_VSA1, SMR_VSA10, SMR_VSA3, SMR_VSA5, SlogP_VSA12, SlogP_VSA2, SlogP_VSA3, SlogP_VSA5, SlogP_VSA6, TPSA, EState_VSA1, EState_VSA2, EState_VSA4, EState_VSA5, EState_VSA8, EState_VSA9, VSA_EState1, VSA_EState4, VSA_EState7, VSA_EState8, FractionCSP3, NumAliphaticCarbocycles, NumAromaticRings, NumRotatableBonds, MolLogP, MolMR, fr_bicyclic

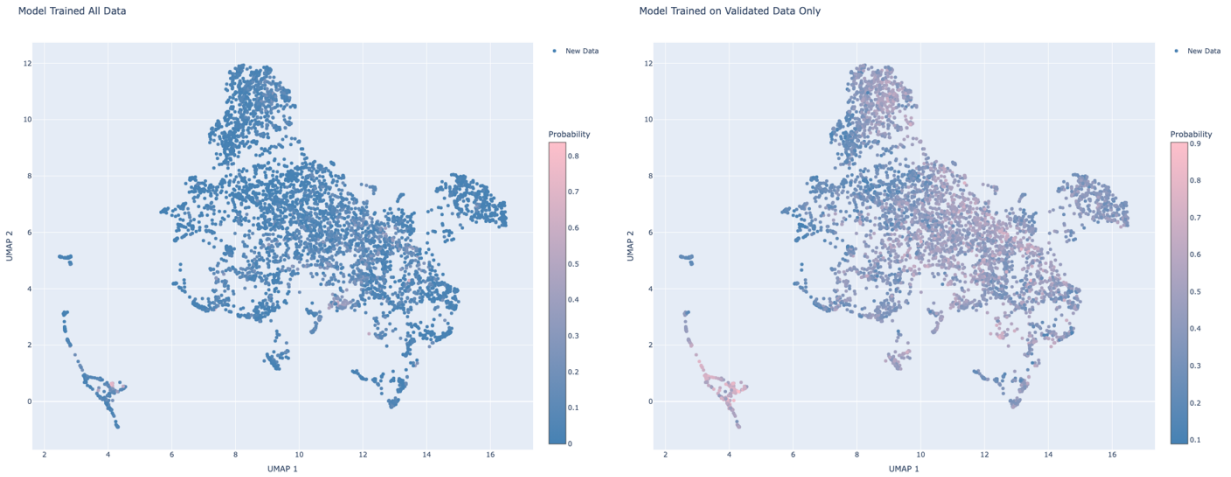
2. Best Random Forest hyperparameters found by BayesianSearchCV

'ccp_alpha': 0.006370568744783853,
'class_weight': 'balanced_subsample',
'criterion': 'entropy',
'max_depth': 15,
'max_features': 0.1,
'max_leaf_nodes': 597,
'min_impurity_decrease': 0.0,
'min_samples_split': 8,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 288

3. The hyperparameter space explored by BayesianSearchCV for Random Forest (along with some parameters for the search algorithm)

```
BayesSearchCV(
    rf_pipeline,
    search_spaces = param_space,
    n_iter=1000,
    n_points=10, # I don't really understand what this does, tbh; supposedly the # of parameters being tested in parallel
    random_state=0,
    scoring = 'precision',
    n_jobs = -1,
    cv = 5,
    return_train_score=True
)
```

4. **Visualizations of the model outputs when trained on 75% of the full data vs.100% of the validated data**

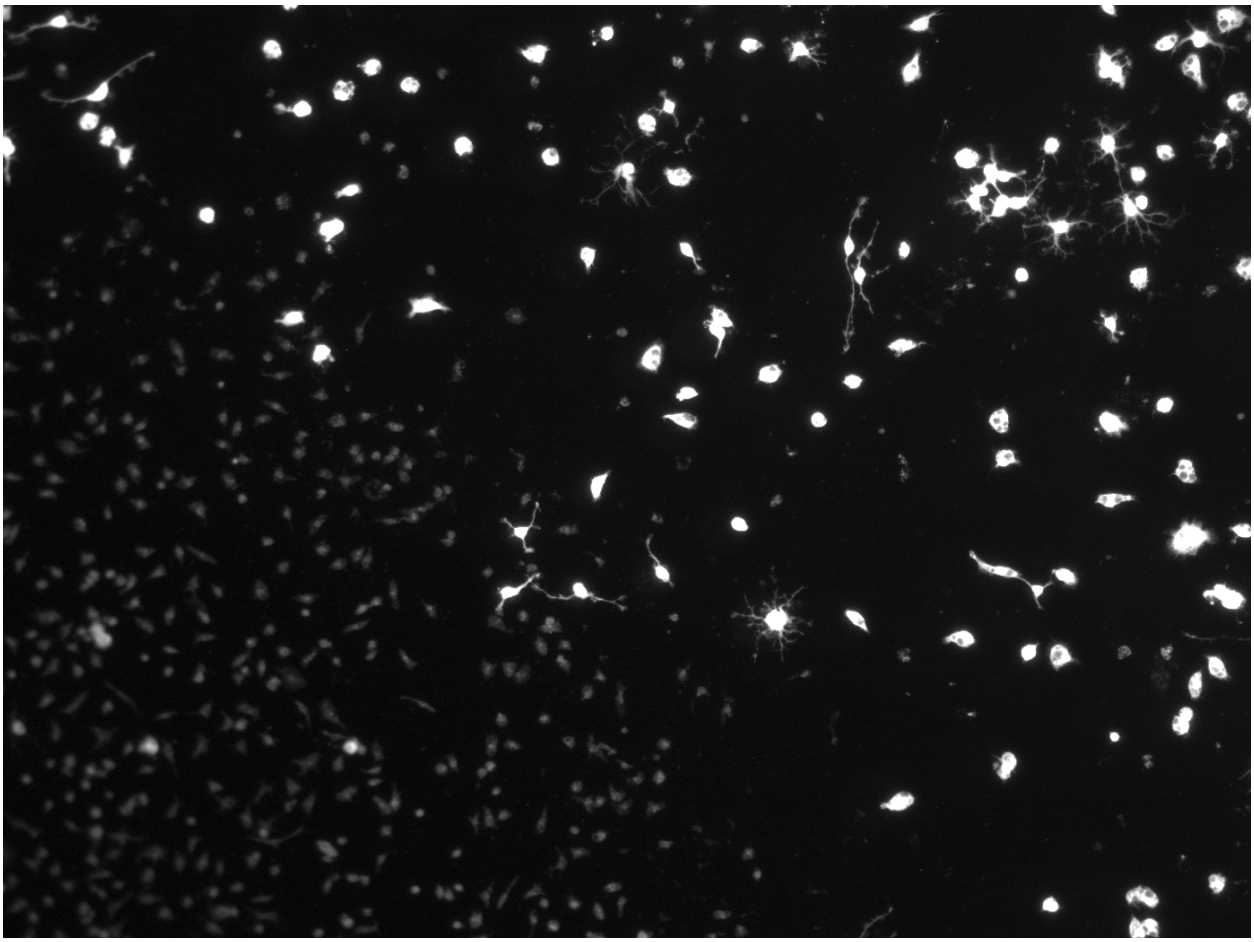


Left: Model trained a 75% split of the full data. *Right*: Model trained on the validated data only. Shown here is a UMAP projection of the discovery (new) data, with the RF model's predicted probabilities shown as colors. (UMAP parameters: n_neighbors=n_13, n_components=2, n_jobs=-1, random_state=0). As can be seen here, the validated-data-only model had much less selective predictions.

5. **The 36 compounds selected for experimental validation**

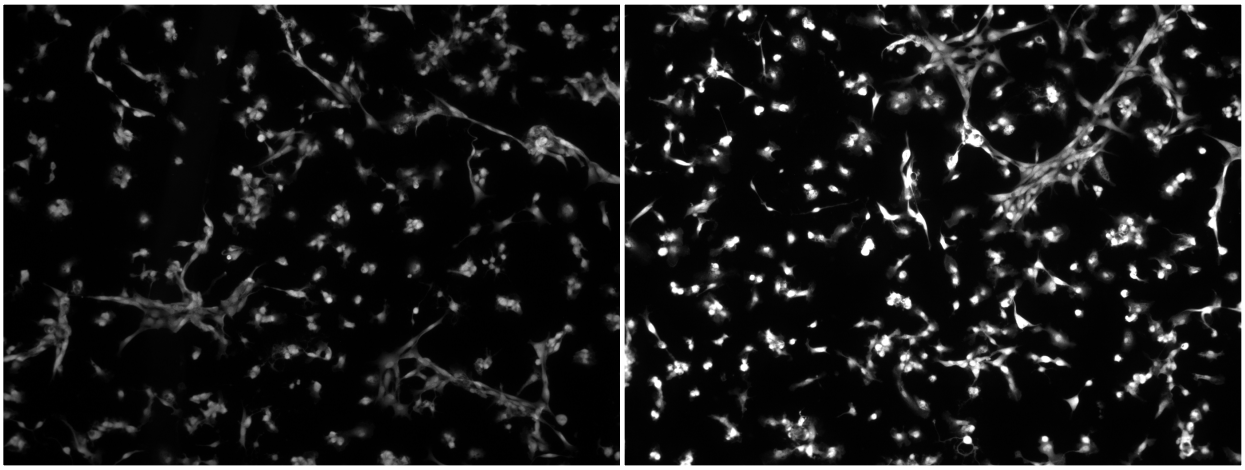
Betamethasone valerate, Ouabain octahydrate, Caudatin, Hydroxygenkwanin, Pectolinarigenin, Salvigenin, Myrislignan, L-165041, 6-Demethoxytangeretin, Prosapogenin A, Decursinol angelate, Hispidulin, Dimethylcurcumin, Hexahydrocurcumin, 7,3',4'-Tri-O-methylluteolin, Octahydrocurcumin, Phyllanthin, Fexofenadine, YK-3-237, Chrysotoxine, SphK1&2-IN-1(WAY-272077), Okanin, Cucurbitacin E, Bigelovin, Brevilin A, Pachymic acid, Mogroside I E1, Microhelenin C, Eriocalyxin B, 11-Keto-beta-boswellic acid, Cimiracemoside C, Alisol F 24-acetate, Ergolide, Caudatin, Ganoderic acid D, Ganoderic acid B (The pink compounds are selected from the second set of predictions)

6. **Representative immunofluorescence images that demonstrate "strangeness" in our cell culture**



Microglia incubated with pectolinarigenin at 1mM concentration. For some reason, almost every well (including the negative control wells with only DMSO) had two distinct-looking populations of cells, as seen here: smaller cells that are generally spindle shaped (left) and larger cell that either amoeboid or oddly shaped or have many branches (right). We generally don't see this in our positive-control (TGF-β) wells, however. Since all of our substances, except TGF-β, were diluted in DMSO, I personally suspect that DMSO might have something to do with this—especially since the different cell populations are often distributed concentrically (e.g., the smaller cells in the center and the larger cells surrounding them in the perimeter). However, I was informed that this never occurred in the past.

It is also possible that the well bottom's shape was incorrect (e.g., concave instead of flat), or that the samples were contaminated with oligodendrocytes, which can be very proliferative even in *in vitro* environments.



DMSO (left) and Caudatin (right) from another imaging session. In another set of experiments, we did not see the different populations of cells, but the cells looked weird—there are many large "clumps" of cells.

7 Supplementary Materials

The full annotated predictions, pathway analysis, data, and code used can be found at https://github.com/Allen2019/Thesis_Supplementary_Materials