UNIVERSITY OF OXFORD

DPHIL THESIS

Barriers to generalisation in acoustic hive monitoring

Author:
Stella Marie FELSINGER

Supervisors:
Prof. Geraldine WRIGHT
Dr. Beth MORTIMER
Dr. Huw EVANS (*Canetis SRL*)
Sandra EVANS (*Canetis SRL*)

A thesis submitted for the degree of Doctor of Philosophy

in the

Oxford Bee Lab
Department of Biology
University of Oxford

Hilary Term 2025

To my grandparents:

Opa Lois, for my first introduction to bioacoustics.

Oma Dorle, for nurturing my love of animals.

Opa Sepp, for inspiring excellence.

Oma Gitti, for reminding me what is really important.

Declaration of authorship

I declare that this thesis was composed by myself and that the work contained herein is my own except where explicitly stated in the text. The work has not been submitted for any other degree or professional qualification.

— Stella Marie Felsinger, Hilary 2025

Acknowledgements

First and foremost, a big thank you to my supervisors, Jeri and Beth, for your support and mentorship over the past 4.5 years. I count myself immensely lucky to have worked with you both.

I'm very grateful to Sandra and Huw Evans, my industrial supervisors at Canetis SRL, for generously sharing their data and expertise with me. A huge thanks also to the team at BeeHero for allowing me to use their monitoring equipment, which was essential to my research, and inviting me to visit their headquarters.

Thank you to my examiners Andrew and Ellie for offering valuable advice during my transfer and confirmation examinations and reassuring me that I was on the right track.

To the members of the Bee Lab, I learnt so much from each of you: how to run an experiment, make diets, dissect bees, extract honey, ... In particular, I want to thank Rui, my beekeeping teacher, JenChen, my glasshouse and lab buddy, and Léa, my research assistant over two seasons. Most importantly, thank you all for making me feel welcome from the start. I truly could not have wished for a better lab with kinder people.

I've been lucky to be surrounded by good friends during my time here. Matt, Emma, and Pepe — Oxford won't be the same without you. To many happy returns. Léa, Marti, and Marci — to many more adventures.

My dear friend Emma, the Mulder to my Scully, I'm so proud of us both.

To my partner Will, thank you for convincing me it's not all terrible. You make my life better every day.

Mama, Papa und Mathis — danke für alles. Ich hab's geschafft!

Contents

0.1 Abstract 1

1 General Introduction **2**

1.1 Introduction 2

1.2 Honeybees 4

1.3 Sounds & Vibrations in honeybee hives 13

1.4 Hive Monitoring 23

1.5 Outline 28

2 Robust design and testing of hive monitoring models **30**

2.1 Abstract 30

2.2 Introduction 31

2.3 Data analysis 35

2.4 Results 44

2.5 Discussion 54

3 Explainable AI for detecting honeybee queen loss **66**

3.1 Abstract 66

3.2 Introduction 67

3.3 Experiment methods 69

3.4 Analytical methods 72

3.5 Results 76

3.6 Discussion 94

4 Acoustic monitoring of colony nutritional state **108**

4.1 Abstract 108

4.2 Introduction 109

4.3 Experiment methods 112

4.4 Analytical methods . . . 124
4.5 Experiment results . . . 135
4.6 Modelling results . . . 140
4.7 Discussion . . . 146

5 General Discussion **158**
5.1 Summary of key findings . . . 158
5.2 What factors influence the performance of colony state prediction models? 160
5.3 What constitutes best practice in hive monitoring? . . . 163
5.4 Conclusion . . . 167

A Supplement to Chapter 2 **168**
A.1 Model size and dataset splitting . . . 168
A.2 Model performance by hive, feature set, and sensor position . . . 171

B Supplement to Chapter 3 **173**
B.1 Inferring hive metrics . . . 173
B.2 Model performance . . . 175
B.3 Mean SHapley Additive exPlanations (SHAP) values in single-hive and multi-hive models across hives . . . 180

C Supplement to Chapter 4 **182**
C.1 Experimental . . . 182
C.2 Soundscape differences . . . 183
C.3 Modelling . . . 189

Bibliography **193**

List of Figures

1.2.1 Deconstructed beehive. 5
1.3.1 Queen piping signals. 18

2.3.1 BeeHero IHS in hive. 36
2.3.2 Illustration of ROC curve. 40
2.3.3 Dataset splitting scheme. 41
2.4.1 Feature correlation. 45
2.4.2 Test AUC by test type. 46
2.4.3 Test AUC by sensor position. 47
2.4.4 Test AUC by model type. 48
2.4.5 Test AUC by feature set. 49
2.4.6 Test AUC by hive. 51
2.4.7 Cross-test accuracy of single-hive models. 52
2.4.8 OOG AUC of MHMs by feature set, hive, and sensor position. 53
2.5.1 Hive configuration affects the frequency spectrum. 61

3.3.1 Experiment schedule and queen states. 70
3.3.2 Interpolation of inspection data 71
3.4.1 Reading 'beeswarm' plots 75
3.5.1 2D views of 3D manifold — queen presence. 77
3.5.2 2D views of 3D audio manifold — queen state by hive 79
3.5.3 2D views of 3D audio manifold — hive population. 80
3.5.4 2D views of 3D audio manifold — food stores. 81
3.5.5 2D views of 3D audio manifold — hive ID, pallet, and sensor position. 82
3.5.6 2D views of 3D audio manifold — day and time of day. 83
3.5.7 Model results — C2 Centre 86
3.5.8 Model results — A3 Inside 88

3.5.9 Mean feature importance for multi-hive models — 20 best features. . 90
3.5.10 Mean feature importance comparison between two models 92
3.5.11 Mean feature importance comparison between two models and mean 93

4.3.1 Experiment setup 113
4.3.2 Hive configuration 114
4.3.3 Experiment timeline. 117
4.3.4 "BroodCam" — custom frame imaging solution. 122
4.3.5 Frame photo with cell content annotations 122
4.3.6 Regression of brood counts to frame estimates. 123
4.4.1 Mean larval weight by day of development. 130
4.4.2 Timeshifted larva and capped brood counts for hive 1. 131
4.4.3 Dataset split for single-hive models. 132
4.4.4 Dataset split for multi-hive models. 133
4.5.1 Brood counts by experiment group at each milestone. 137
4.5.2 Interpolated brood counts by experiment group. 138
4.5.3 Chalkbrood severity by experiment group. 139
4.6.1 Soundscape differences between each treatment group and control. . 141
4.6.2 Model goodness-of-fit by model type and prediction target. 143
4.6.3 Multi-hive model (MHM) goodness-of-fit by experiment group and prediction target. 144
4.6.4 Multi-hive model (MHM) goodness-of-fit by test hive group and prediction target. 145
4.7.1 Correlation of background noise PSD and colony strength. 152

A.1.1 Preliminary testing — Sensor position and learning rate. 169
A.1.2 Train-validation-test splitting schemes. 170
A.2.1 OOG AUC of MHMs by hive, feature set, and sensor position. 172

B.1.1 Interpolated inspection data. 174
B.2.1 Model results on training set — C2. 175
B.2.2 Model results on training set — A3. 176
B.2.3 Mean correctness polar plots — C2. 177

B.2.4 Mean correctness polar plots — A3. 179
B.3.1 Mean feature importance for single-hive and multi-hive classifiers. . . 181

C.1.1 Regression of chalkbrood severity and colony health. 185
C.2.1 Soundscape differences in the morning. 186
C.2.2 Soundscape differences in the evening. 187
C.2.3 Soundscape differences at night. 188
C.3.1 Impact of weighting function on larval brood deviation metric. 190
C.3.2 Impact of model checkpoint on model performance. 190
C.3.3 Model goodness-of-fit by learning rate. 191
C.3.4 Multi-hive model goodness-of-fit by test hive. 192

List of Tables

1.3.1 Fundamental frequencies of in-hive sounds and vibrations. 21

2.2.1 Common hyperparameters. 33

3.4.1 Number of recordings in sensor subsets. 73

4.3.1 Allocation of hives into experimental groups. 116
4.3.2 Mapping between notations of experiment time. 117
4.3.3 Pollen and candy patty recipes. 119
4.4.1 Honeybee brood cycle 127
4.4.2 Mean larval weight by day. 129
4.4.3 Hyperparameters for nutritional state detection models. 134

B.1.1 Queen status by day. 173
B.2.1 Mean model correctness by time of day — C2. 178
B.2.2 Mean model correctness by time of day — A3. 179

C.1.1 Colony health assessment at D0. 184
C.1.2 Regression between brood frame estimate and cell count. 184
C.3.1 Cohort weighting for nutritional state label. 189

Acronyms

ML Machine Learning 1

AI Artificial Intelligence 2

ANN Artificial Neural Network 31

SLP single-layer perceptron 32

MLP multi-layer perceptron 32

CNN Convolutional Neural Network 36

FNO Fourier Neural Operator 133

MIL multi-instance learning 62

PCHIP Piecewise Cubic Hermite Interpolating Polynomial 70

UMAP Uniform Manifold Approximation & Projection 72

SHAP SHapley Additive exPlanations 68

PAM Passive Acoustic Monitoring 3

WG within-group 28

OOG out-of-group 27

SHM single-hive model 41

MHM multi-hive model 41

ZCR zero-crossing rate 38

PSD Power Spectral Density 34

CWT Continuous Wavelet Transform 36

STFT Short-Time Fourier Transform 36

IoT Internet of Things 25

IHS in-hive sensor 23

AWS Amazon Web Services 25

MSE Mean Standard Error 134

AUC Area Under Curve 27

ROC Receiver Operating Characteristic 40

FPR false positive rate 84

FNR false negative rate 84

SNR signal-to-noise ratio 55

JH Juvenile Hormone 7

QMP Queen Mandibular Gland Pheromone complex 7

HPG hypopharyngeal gland 6

0.1 Abstract

In industrial apiculture, honeybee colonies are exposed to a range of interacting stressors which impact their wellbeing and productivity. Machine Learning (ML) offers the promise to detect colony condition from sensor data to optimise pollination performance and honey yields while maintaining colony health. However, robust testing of these methods is lacking, limiting real-world applications. In this work, I develop best practices for colony state prediction and attempt to characterise the major barriers to cross-hive generalisation.

In Chapter 2, I develop a pipeline to extract handcrafted and deep-learned features from honeybee audio samples and train neural networks for binary classification. Using a dataset of 10 queenless and queenright hives, I test the effect of different feature extraction methods, sensor positions, training set diversity, and hive ID on prediction accuracy. Crucially, I demonstrate how test set choice can artificially inflate model performance and discuss best practices for model testing.

In Chapter 3, I investigate how different hive metrics shape the in-hive soundscape using UMAP embeddings. I then delve into the prediction time series of individual models, testing different methods for increasing prediction fidelity. Finally, I use SHAP analysis to identify which audio features contribute to detecting queenlessness across hives and relate these features to our existing knowledge of honeybee vibrational communication.

In Chapter 4, I present a dataset of 25 colonies undergoing varying levels of pollen starvation in combination with chalkbrood infection and propose an intuitive visualisation of soundscape differences over time. I analyse the effect of periods of pollen dearth on brood production and test the ability of neural networks to predict nutritional state from a multi-modal feature set of audio, temperature, and humidity features.

Overall, I identify an overwhelming lack of robust testing in the literature, leading to many overstated claims. I demonstrate a strong confounding effect of hive ID and encounter difficulties with cross-hive generalisation in my own models, dependent on dataset and prediction task. Large multi-hive, multi-context datasets will be required to one day achieve generalisation across apiaries, seasons, and climates.

Chapter 1

General Introduction

1.1 Introduction

Honeybees are important pollinators with high commercial value. Many luxury crops like almonds, cherries, or avocados depend in large part on pollination services provided by honeybees. Even for crops that can be wind- or self-pollinated, pollination by bees can improve quality, shelf life and commercial value (Klatt et al., 2014). Pollination services from commercially managed honeybee colonies are therefore especially needed in monocultures and other cropping systems lacking natural populations of pollinators. To ensure harvests and improve yields of pollinator-dependent crops, commercial agriculture depends on a stable supply of honeybee colonies that can be moved to the orchard or field where their services are needed. Pollination as a service is now a $200m+ industry in the US (NASS, 2025). Maintaining the health of honeybee colonies within this system is therefore not only a major welfare concern for our most intensively managed insect, but an economic concern (Goodrich, 2019).

On pollination deployments, honeybee colonies face a variety of stressors through the year including exposure to pesticides (Pettis et al., 2013), disturbance due to transport (Ahn et al., 2012), enhanced spread of diseases and parasites as colonies from different geographical regions mix (Cavigli et al., 2016), and poor nutrition due to their monofloral diet while deployed. Additionally, these stressors are known to compound – for example, bees are less likely to survive disease if they are fed a monofloral pollen diet (Dolezal et al., 2019).

To help colonies survive and thrive under these conditions, remote hive monitoring using Artificial Intelligence (AI), or "precision beekeeping", could aid beekeepers

in streamlining operations and targeting interventions to those colonies most in need. In-hive audio is a particularly promising data source, given the rich fabric of vibrational communication in a honeybee colony (Hrncir, Barth, and Tautz, 2006). Passive Acoustic Monitoring (PAM) is gaining in popularity for many biological applications, such as species distribution modelling or activity monitoring (Gallus et al., 2012; Sugai et al., 2019; Favaro et al., 2021). Machine Learning is often used to extract insights from the vast datasets that can be gathered by automated soundscape recording (Sugai et al., 2019; Duarte et al., 2024).

Remote monitoring of honeybee colony foraging strength is already in commercial use to optimise colony placement within orchards and monitor pollination. Similar academic work centred around detecting colony condition from in-hive sound, mainly queen loss and reproductive swarming, has been underway for decades (Meikle and Holst, 2015). However, these studies are typically based on limited datasets with few hives, lack experimental protocols to control for confounding factors, and do not test models under real-world conditions.

In this thesis, I aim to characterise the challenges surrounding colony state prediction and identify barriers to cross-hive generalisation. Much of the research into hive monitoring does not go beyond reporting model accuracy. To better understand variations in model performance, I introduce a method to link model predictions to input features and honeybee vibrational communication using Shapley values (Lundberg and Lee, 2017). Furthermore, hive monitoring typically targets rare events like swarming and queen loss. Monitoring nutritional state to detect periods of pollen dearth has never been attempted before. I present the first dataset of in-hive audio, temperature, and humidity in 25 colonies undergoing pollen starvation under controlled conditions and develop neural network models for predicting colony strength and nutritional state. Finally, I give recommendations for best practice in hive monitoring, including experiment design, data collection, and model testing.

1.1.1 Structure of introduction

In section 1.2, I describe typical beekeeping practices for colony maintenance and the functioning of healthy honeybee colonies as well as colonies experiencing queen loss and starvation. In section 1.3, I review the literature on sounds and vibrations in

honeybee hives, including their generation and perception by bees, signals used in vibrational communication, and ways of measuring in-hive sound and vibration. In section 1.4, I present the current state of research on precision beekeeping. I discuss the aims of hive monitoring, typical sensors used, and common targets and challenges of hive monitoring models.

1.2 Honeybees

1.2.1 Beekeeping

Humans have harvested honeybee products and cultivated swarms since prehistoric times (Roffet-Salque et al., 2015; Kritzky, 2015). The shift from harvesting wild nests to maintaining colonies in proximity to settlements may have happened as early as the Neolithic (Roffet-Salque et al., 2015), and had certainly happened by the 2nd millenium BC (Kritzky, 2015). The nest usually needed to be partially destroyed in order to harvest wax or honey.

Modern Western beekeeping practice uses hives with removable frames or bars, which facilitate regular hive inspections and yearly honey harvests without killing the colony and/or destroying the nest. Hives are often modular, allowing the nest cavity to be expanded or reduced depending on the size of the colony and seasonal need, eg honey production or swarming prevention (Fig. 1.2.1).

During an inspection, the hive is opened and the beekeeper estimates colony strength by the size of the cluster viewed from above or the proportion of frames covered with bees. The beekeeper then looks at the frame contents to gauge brood production and stores of honey and bee bread. The beekeeper particularly checks for recently-laid eggs (standing up in the centre of the cell) to make sure that the queen is actively laying. If young eggs are missing or there are concerns about the queen's productivity or condition, the beekeeper attempts to find the queen herself. Inspections also allow the beekeeper to identify signs of disease or parasites, such as the parasitic mite *Varroa destructor* or chalkbrood (caused by the fungal pathogen *Ascosphaera apis*), and to assess general colony condition by observing the bees' behaviour before and during hive opening. Based on the results of the inspection, the beekeeper decides whether



FIGURE 1.2.1: Deconstructed hive consisting of, from bottom: base board with landing board, two brood boxes, queen excluder, honey super (frames and box), and top cover. Image from Zhu et al. (2024).

a colony requires any intervention, such as queen replacement, disease treatments, or supplemental feeding (Wahl, 2023).

Colonies are typically inspected every 2 weeks, with some recommending weekly inspections during early summer to prevent swarming. Inspections should be kept as short and infrequent as possible, since opening the hive and moving frames disrupts colony homeostasis. Rebalancing the in-hive climate is thought to take up to 2 days, though research on this is lacking (Wahl, 2023).

Experienced beekeepers can perform a hive inspection in less than 15min[1], and commercial beekeepers, who may be responsible for hundreds or thousands of hives,

[1] anecdotally, see for example https://theapiarist.org/inspections/

often perform rapid inspections lasting a few minutes per hive. While this is an efficient way to process large numbers of hives, it creates scope for missing developing problems. Hence, there is potential for new applications of sensor technology and data science to improve colony productivity and survivorship and to increase beekeeper efficiency by reducing the need for inspections.

1.2.2 The honeybee colony

Honeybees live in eusocial colonies made up of three castes: a single queen (the reproductive female), thousands to tens of thousands of (mostly) sterile female workers, and a seasonal population of drones (male reproductives) (Winston, 1987; Seeley, 1989b). Workers have average lifespans of between 2–4 weeks in summer (Nearman and vanEngelsdorp, 2022) and 6–8 months in winter (Amdam and Omholt, 2002), while drones live for ca. 1 month in spring and early summer and up to 3 months in late summer and autumn (Winston, 1987). Queens typically live for 2–3 years, though beekeepers sometimes replace them every 1–2 years to maintain productivity (Winston, 1987; Thompson, 2016).

Division of labour

With the exception of egg laying, which is done by the queen, all colony maintenance tasks are fulfilled by the worker caste. Tasks include nest maintenance, feeding the brood, queen, and other workers, wax production and comb building, food handling, guarding, and foraging for pollen, nectar, water, and plant resins. Most workers progress through these tasks as they age, but they show great flexibility to switch between tasks based on the needs of the colony (Lindauer, 1952; Winston, 1987). This is especially true for physiology-independent tasks, such as cell cleaning, debris removal, comb building, and cell capping. Other tasks require certain physiological specialisations (Johnson, 2003). These chiefly include nursing and foraging.

Workers in the nurse caste are able to digest pollen and produce glandular secretions including royal jelly with their large hypopharyngeal glands (HPGs) to feed the queen, brood, and other members of the colony. Their abdominal fat body contains significant nutrient reserves, including the storage protein vitellogenin, which can be

drawn on during periods of pollen dearth (Amdam and Omholt, 2002; Corby-Harris, Snyder, and Meador, 2019).

The transition from hive bee to forager is mediated by Juvenile Hormone (JH) and accompanied by physiological changes, chiefly the degradation of the HPGs and fat body (Robinson, 1986; Huang et al., 1991; Sullivan et al., 2000). JH biosynthesis is affected by octopamine, a neurotransmitter linked to energetic stress (Kaatz, Eichmüller, and Kreissl, 1994; Salomon et al., 2012), and Queen Mandibular Gland Pheromone complex (QMP), a group of pheromones that signal queen presence and suppress worker fertility (Pankiw et al., 1998; Katzav-Gozansky et al., 2004). Workers typically begin foraging at 18–28 days old (Winston, 1987), but when a colony is food-stressed, workers as young as 7 days can become precocious foragers (Rösch, 1930; Page, 1992). As a result of these physiological changes, foragers can no longer perform brood-rearing tasks but are 40% lighter, enabling energy-efficient flight (Harrison, 1986). This temporal division of labour optimises the use of each worker's lifespan by avoiding the most dangerous tasks, foraging and guarding the hive entrance, until her internal nutrient reserves are depleted (Amdam and Omholt, 2002).

Nutrition

Colony nutrition is a collaborative process in which all ages of workers take part. Floral nectar and pollen are collected by foragers and brought to the hive, where they are accepted by food storers and processed into honey and bee bread. Nurse bees digest bee bread to produce glandular secretions and feed all other members of the colony by trophallaxis, mouth-to-mouth feeding (Crailsheim, 1991; Wright, Nicolson, and Shafir, 2018). Foragers are highly sensitive to the nutritional state of the colony which they perceive via interaction with hive bees (Seeley, 1989a; Camazine, 1993; Schulz, Huang, and Robinson, 1998).

Pollen is the principal source of essential nutrients for brood development, including protein and fat as well as micronutrients such as minerals, vitamins, and essential sterols (Wright, Nicolson, and Shafir, 2018). Pollens differ in micronutrient content (Wright, Nicolson, and Shafir, 2018; Baker, 2024). Nurse bees are able to self-regulate between complementary diets to cover their nutritional needs (Stabler et al., 2021),

and it is assumed that foragers fulfil this function for the colony, though this has not been demonstrated directly.

Honey is the colony's main source of carbohydrates and is consumed directly by hive bees and foragers. Foragers have a higher need for honey since it fuels flight, but they also receive protein and lipids via jelly from nurses (Crailsheim, 1991; Wright, Nicolson, and Shafir, 2018).

1.2.3 Starvation

Colonies can be affected by sugar starvation (lack of nectar/honey) as well as by pollen starvation. A lack of honey shows its effects very quickly: Saltykova, Gaifullina, and Nikolenko (2008) reported 50% mortality after 36 hours in an adult sugar starvation caged assay. Pollen starvation mainly affects brood production and thus has slower, but longer-term impacts on colony health.

Because each worker plays a role in managing and processing colony resources, the nutritional state of individuals has repercussions for that of the entire colony. A weak forager is less likely to return from a foraging flight. Fewer foragers mean fewer resources for nurse bees to feed the queen, brood, drones, and other workers (Wright, Nicolson, and Shafir, 2018). The nutritional state of the queen determines egg size and the proportion of drone eggs laid, affecting colony-level reproduction (Sasaki and Obara, 2001; Amiri et al., 2020). Poor nutrition of worker and drone larvae can affect adult body size, skill in foraging and waggle-dancing, lifespan, and health (Alaux et al., 2010; Scofield and Mattila, 2015).

Signs of Starvation

Because starvation affects each caste of the colony differently, the signs of starvation are varied and differ between castes as well as between types of starvation. The brood and queen are most directly affected by pollen starvation, and workers by sugar starvation, but foragers are sensitive to both types as they regulate the colony's nutritional intake (Fewell and Winston, 1992). Drone larvae that had experienced a 10-hour period of starvation grew into smaller adults with smaller wings (Szentgyörgyi, Czekońska, and Tofilski, 2017). Workers starved as larvae have smaller ovaries, elevated glycogen and juvenile hormone levels, and decreased sugar sensitivity (Wang

et al., 2016). Hungry larvae emit the pheromone E-beta-ocimene which elicits more inspection visits from nurse bees (Huang and Otis, 1991; He et al., 2016). Nurse bees that are starved after emerging have smaller hypopharyngeal glands than their well-fed sisters due to activation of autophagy functions (Corby-Harris, Snyder, and Meador, 2019). Changes to fat storage have also been observed (Mayack, Natsopoulou, and McMahon, 2015).

Foragers from colonies low on nectar and honey accept and advertise sources of less-concentrated nectar and drink nectar more quickly than foragers from well-fed colonies (Seeley, 1989a; Buckemüller et al., 2017). When nectar foragers return from a flight, they transfer their nectar load to waiting food storer bees. A storer bee then looks for a cell to deposit the nectar where it can be processed into honey. When nectar is plentiful, there are a lot of foragers arriving and storer bees have more trouble finding space to store the nectar, leading to longer wait times for foragers. Seeley (1989a) suggested that this wait time can be used as a proxy for the availability of storage space and thus the state of the colony's resources.

Starvation causes colony size to decrease over time (Schulz et al., 2002). This happens via several mechanisms. Firstly, a lack of food stores in the colony leads to increased foraging (Mayack, Natsopoulou, and McMahon, 2015) and precocious foraging, where workers begin foraging at a younger age (Schulz, Huang, and Robinson, 1998; Schulz et al., 2002). As foraging is the riskiest task a worker performs in her lifetime, this leads to higher losses of adult workers.

Secondly, survival of larvae and pupae is reduced because there is not enough food for them to complete their development. Colonies quickly resort to brood cannibalism, where nurse bees kill and consume younger larvae to prioritise feeding older brood, which have received a greater total investment of labour and nutrients (Woyke, 1977; Scofield and Mattila, 2015). Feeding on brood allows nurses to replenish their own stores of protein and fat (Weiss, 1984). Brood cannibalism is thought to mainly target young larvae (Schmickl and Crailsheim, 2001; Crailsheim et al., 2003), but cannibalism of older larvae has also been documented in resource-depleted hives without foragers (Rösch, 1930).

Workers in food-stressed colonies have been seen to harass their queen, biting and stinging her and pulling on her legs and wings, to stop her from laying eggs and

so prevent further drain on colony resources (Gonçalves, 2021). This behaviour has also been observed prior to swarming (Allen, 1955). However, the egg-laying rate of queens in resource-depleted colonies is typically constant (Crailsheim et al., 2003).

Thirdly, poorly fed workers and brood are more susceptible to diseases and parasites and more likely to die as a result of infection (Dolezal et al., 2019; Di Pasquale et al., 2013). Infection of workers with the gut parasite *Nosema* increased hunger, as shown by higher responsiveness to sucrose in PER assays and a decreased willingness to share food with nest mates via trophallaxis (Naug and Gibbs, 2009). Mayack and Naug (2009) also showed that bees infected with *Nosema* die more quickly than healthy bees in a survival assay. Dolezal et al. (2019) demonstrated increased mortality from Israeli Acute Paralysis Virus (IAPV) when bees were fed monofloral pollen diets. Deformed Wing Virus (DWV) titre was also shown to be highest in bees fed only sucrose syrup, and lowest in bees provided pollen or pollen substitute (Dolezal et al., 2019).

These behavioural signs of starvation are easily noticed during hive inspections, but are harder to measure directly using sensors, either due to the cost (eg gas sensors to detect the hunger pheromone E-beta-ocimene) or the challenges of sensor deployment and complexity of data analysis (eg in-hive cameras to monitor brood condition or pollen stocks). By studying how more easily measurable hive metrics (in-hive sound, temperature, and humidity) respond during periods of starvation, we can develop simpler methods for the remote monitoring of nutritional state.

1.2.4 Queen loss

The queen bee is unique in the honeybee colony as the only reproductive female. She lays up to 1000 eggs per day to maintain and grow the colony (Winston, 1987). Given that honeybee workers have lifespans as short as a few weeks, without this supply of new workers, the colony quickly dwindles and dies. To prevent extended periods without brood, honeybee colonies are able to replace their queen if she dies or becomes unproductive.

Queen presence is communicated throughout the hive by several queen-specific

non-volatile pheromones (QMP complex) (Seeley, 1979; Naumann et al., 1991; Naumann, Winston, and Slessor, 1993). Through allogrooming and trophallaxis (grooming and food sharing between nestmates), these pheromones are passed from worker to worker so that a constant level of the pheromones is present throughout the hive (Seeley, 1979). A decline in QMP levels triggers reproductive rescue strategies, such as emergency queen rearing and eventually worker ovarian development (Butler, 1954; Katzav-Gozansky et al., 2004).

Emergency queen cell construction has been reported to begin from 10h (Seeley, 1979) to 12–48h post-queen removal (Fell and Morse, 1984), with the first signs of queen loss recognition observed after 6–12h. This suggests that the colony can become aware of the loss of their queen within hours and often begin rearing a replacement within the same day.

To produce a new queen, young worker larvae are fed liberally with royal jelly secreted from mandibular and hypopharyngeal gland, rather than the less nutritious "worker jelly", a mix of honey and pollen with glandular secretions (Winston, 1987; Crailsheim, 1990; Wright, Nicolson, and Shafir, 2018). Larvae chosen for emergency queen rearing are typically between 1–2 days old and well-fed (Fell and Morse, 1984; Sagili et al., 2018). The nutritionally dense and abundant feed allows them to grow rapidly and fully develop their ovaries.

A colony left without a queen ("queenless") for too long develops laying workers – bees who, in the absence of QMP, have developed their ovaries and now are able to lay eggs (Katzav-Gozansky et al., 2004). However, since these laying workers are unmated, their eggs develop into drones, not workers. Drone production by laying workers can be considered a final "last-ditch" effort at colony reproduction (Utaipanon, Holmes, and Oldroyd, 2019), after which the colony inevitably dies off.

Beekeepers usually monitor queen presence during regular hive inspections, which take place every 1–2 weeks throughout spring, summer, and autumn. While inspecting frames, beekeepers look for the queen herself, easily recognised by her long abdomen and often a paint mark on her thorax. They also look for fresh eggs (standing upright in the centre of the cell floor) to check that the queen is actively laying. The presence of fresh eggs by themselves is usually taken as sufficient evidence of a healthy queen, while the presence of the queen without any fresh eggs can hint at

issues with her productivity. If the queen cannot be found or is determined to be unproductive, a new mated queen can be introduced. Alternatively, the colony can be assisted in rearing a new queen, for example by introducing a frame with young brood from another hive (Thompson, 2016).

Expected behavioural shifts after queen loss

Honeybee workers quickly become aware of queen loss and commence emergency queen rearing within 10–48h (Seeley, 1979; Fell and Morse, 1984). Since queen presence is mainly communicated via pheromones (QMP), which are propagated by contact (Winston, 1987; Katzav-Gozansky et al., 2004), it is unlikely that honeybees use an additional acoustic or vibrational signal. Instead, I expected any acoustic indications of queen loss to result from changes in behaviours with acoustic or vibrational components.

Emergency queen rearing involves constructing queen cells onto existing brood cells with a young larva (1–2d old) (Fell and Morse, 1984). Worker bees select only well-fed larvae as potential queens, since larval nutrition is the main contributor to successful queen development (Sagili et al., 2018; Fell and Morse, 1984). Workers need to evaluate the availability and nutritional state of young brood and select the best candidates for queen rearing. This could lead to a short-term increase in brood nest activity in the first few days following queen loss until queen larvae have been fully provisioned and sealed, typically 3–4d after queen rearing was initiated (Fell and Morse, 1984; Winston, 1987).

Queen rearing is prioritised over brood care and up to 50% of the brood present at the time of queen loss are lost as a result (Winston, 1979; Winston, 1987). Coupled with the lack of new brood, activity in the brood nest (feeding larvae, cell cleaning, and wax working) would likely decrease once queen cells and other brood are capped.

The loss of JH repression by QMP would likely induce many hive bees to transition to foraging (Pankiw et al., 1998), perhaps serving to bring in more pollen for queen provisioning. This would increase foraging and flight activity at the entrance starting several days after queen loss.

In caged assays, small, queenless groups of bees were more active and clustered less tightly compared to queenright groups (Grodzicki, Piechowicz, and Caputa, 2020).

Workers from queenless colonies are also often described as nervous, agitated, and restless (Fell and Morse, 1984; Winston, 1987). However, this agitated state is likely to represent a general stress response rather than being unique to queenlessness.

Overall, we might expect queen loss to cause an immediate increase in activity in the brood nest, followed by a decrease in activity after several days. This might be accompanied by an increase in activity at the nest entrance as more workers transition foraging.

Roaring

Queenless hives are often described as emitting a 'roaring' sound when opened. This sound has been attributed to increased Nasonov scenting (exposing Nasonov glands on abdomen and fanning wings) by workers inside of the hive (Butler, 1954; Fell and Morse, 1984). Fell and Morse (1984) observed an increase in scenting following queen removal, peaking after 12-24h and declining as more queen cells were constructed. Nasonov scenting is typically used as an orientation signal — worker bees use this pheromone to mark the nest entrance, exceptionally rewarding forage sites, and swarm clustering locations (Winston, 1987). Increased scenting following queen loss has therefore been suggested to be an attempt to aid a lost queen in finding her way back to the hive (Fell and Morse, 1984). The decrease in roaring soon after queen cell construction begins would then represent the "realisation" of the scenting workers that the old mated queen is not returning.

A thorough vibrational characterisation of the roar, as well as its evolution throughout the queenless period is lacking. Such a study, using coupled in-hive audio and video recording, would be a valuable future contribution to the literature.

1.3 Sounds & Vibrations in honeybee hives

Honeybees generate substrate-borne and airborne vibrations and sounds in many context, including communication in the darkness of the hive. To extract information on colony state from in-hive sound and vibration, it helps to understand how sounds are generated by bees and transmitted within the hive. Understanding how vibrations and sounds are perceived by bees can help distinguish communication signals from

incidental vibrations. Knowing which behaviours have vibrational components can aid in linking colony sound to the underlying behaviours and physiological states.

In most contexts, airborne and substrate-borne vibrations are not measured concurrently. Therefore, we treat both modes of transmission as equivalent here.

1.3.1 Generation, Transmission & Perception

Honeybees generate vibrations by thoracic oscillation (Hrncir, Barth, and Tautz, 2006). Bees can modulate the fundamental frequency of the vibration by contracting the flight muscles in the thorax (King, Buchmann, and Spangler, 1996). The position of the wings also affects the frequency, with the highest frequencies possible when the wings are folded behind the abdomen (Seeley and Tautz, 2001; Łopuch and Tofilski, 2017; Hrncir, Barth, and Tautz, 2006). The surface area of the wings also plays a role; experimentally clipping wings increases the frequency at which they can beat (Hrncir, Barth, and Tautz, 2006), presumably due to reduced air resistance or a lower resonant frequency of the oscillatory system. While the same muscles are used for sonication and flight, the ratio of muscle activity in the dorsal longitudinal muscles and the dorsoventral muscles is 1.5–4 times higher during sound production in wild bees. This unequal shortening of the dorsal longitudinal muscles may decouple the wings from the indirect flight muscles, allowing the muscles to move more rapidly (King, Buchmann, and Spangler, 1996).

Vibrations are transmitted as sound through the air and as substrate vibrations via contact of their legs with the comb. Waggle dancers' vibrating wings create a 'jet' of air that may aid in directional signal transmission and attract recruits to the dance floor (Michelsen et al., 1987; Kirchner and Sommer, 1992). The substrate-borne component of the waggle dance may also attract potential foragers. Waggle dances usually occur on open comb, which better transmits vibrations, and dances on open comb recruit up to three times as many workers as dances on capped brood comb (Tautz, 1996; Tautz, Casas, and Sandeman, 2001); however, see also Dyer (2002).

Bees perceive substrate-borne vibrations between 150–750Hz via the subgenual organs (Kilpinen and Storm, 1997). Low-frequency substrate vibrations are likely detected with other sensitive mechanoreceptors including campaniform sensilla and hair plate sensilla (Thurm, 1964). Airborne vibrations are perceived with the Johnston's

organs in the pedicel of the antennae (Dreller and Kirchner, 1993; Tsujiuchi et al., 2007). Sensitivity of the Johnston's organs is age-dependent and is highest between 265–350Hz (Tsujiuchi et al., 2007). Bees seem to perceive airborne sound as air-particle oscillations (near-field sound), rather than pressure oscillations as most vertebrates, and microphones, do (Towne and Kirchner, 1989).

1.3.2 Sounds & vibrations as communication

Waggle dance

The waggle dance is used by a returning forager to communicate the location and quality of a food source to other foragers. It involves the dancer rapidly waggling from side to side while moving her body forward, before returning in a half circle to her starting point. The angle between the "waggle run" and a vertical line corresponds with the angle between the food source from the hive and the sun. The length of the waggle run corresponds with the distance to the food source (Dyer, 2002).

During the waggle run, the dancer often emits 20ms "wing-beating buzzes" at 250–300Hz ((Michelsen et al., 1986; Esch, 1961a)). Sounds are mainly transmitted to the air via the wings (Esch, 1961a; Michelsen et al., 1987; Michelsen, 2003; Kirchner and Sommer, 1992). Dancers with experimentally shortened wings were much less successful in recruiting other foragers than dancers with intact wings (Kirchner and Sommer, 1992). The wing-beating buzz usually occurs whenever the bee is fully displaced to one side during the waggle run (Łopuch and Tofilski, 2017), which may aid in transmitting vibrations via the legs to the comb surface (Tautz, 1996; Tautz, Rohrseitz, and Sandeman, 1996; Tautz, 2009; Hunt and Richard, 2013). The pulse duration, amplitude, and duty cycle of these dance sounds correlate with the sugar concentration and flow rate of a nectar source. This suggests that waggle dances not only communicate the location of a food source, but also encode information about its profitability (Hrncir et al., 2011).

In stingless bees (*Melipona*), pulsed thoracic vibrations by returning foragers appear to communicate the energy budget associated with a food source, having a higher pulse duration and shorter pauses between pulses when the sugar concentration at a feeder increased (Hrncir et al., 2004). Rather than communicating either the distance

or sugar concentration directly, it reflects the trade off between both variables. Because the communication system of honeybees has specialised to communicate distance and direction separately, the thoracic vibration appears to now reflect the profitability more directly than in the stingless bees' compound signal.

Thoracic vibration

Every worker produces a low-frequency latent thoracic vibration that may act as an indicator of activity level and motivation. Hrncir, Maia-Silva, and Farina (2019) recorded thoracic vibrations of foragers, active hive bees, and resting hive bees with a laser vibrometer aimed at the bee's thorax. Frequency and vibration velocity both increased with activity (active hive bees during short interruption: 30 Hz; active hive bees: 36 Hz; foragers: 43Hz) and, for foragers, with increasing quality of an offered sugar source (Hrncir, Maia-Silva, and Farina, 2019). Communicative intent of these latent vibrations has not been directly demonstrated — they could be a byproduct of metabolic activation, or an intentional activity signal. However, they do appear to be reliable cues of activity state, and could communicate this to other bees. Overall, thoracic vibrations seem to indicate the motivational state.

Shaking signal

In the shaking signal first described by Allen (1956), a worker grasps another bee's thorax with its legs and shakes its body for ca. 1.2s at 16.3 ±5.8Hz (measured from video recordings by Gahl (1975)). It is currently unclear whether the shaking signal has a significant sound component. The signal is thought to have a motivational function and may have originated through ritualised aggression (Skaggs et al., 2014).

Shaking signals are typically performed by single workers that move through the nest and shake between a dozen and a hundred bees in series (Seeley, Weidenmüller, and Kühnholz, 1998) and are usually directed at inactive workers (Lewis, Schneider, and Degrandi-Hoffman, 2002). While workers of all ages can receive and perform shaking signals, the response to the shaking signal is age-dependent. Young hive bees who receive the shaking signal spend more time moving through the hive, inspecting cells, and exchanging food with other workers (Cao et al., 2007). Bees of foraging age

are more likely to move closer to the entrance where they can be recruited by waggle dancers while younger workers engage in in-hive maintenance tasks (Nieh, 1998). These activities allow the shaken worker to gather information about the colony's needs which may improve organisation of labour. The queen is also targeted with shaking signals in preparation for swarming, and signals become more frequent until finally the queen is being shaken every 30s up to the swarm's departure (Allen, 1956).

The bee performing the signal only shakes less than half of the bees that she makes contact with (Lewis, Schneider, and Degrandi-Hoffman, 2002). It is possible that the shaking bee perceives the latent thoracic vibration (see section 1.3.2) of each worker she contacts. This activity indication could allow her to distinguish between truly inactive workers, in which case she may perform the shaking signal, and active workers taking a short rest.

Queen piping

Queens and workers both produce sounds known collectively as 'piping'. Like other sounds, this is generated by activating the wing musculature. During piping, the thorax is pressed against the comb, resulting in substrate-borne vibrations as well as airborne sound (Simpson, 1964; Michelsen, Kirchner, and Lindauer, 1986; Kirchner, 1993).

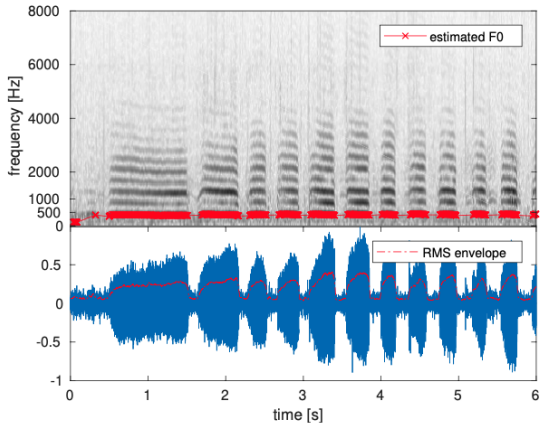
Queens pipe to communicate with each other and their colony during queen rearing and swarming. Queen piping is divided into 'tooting' and 'quacking' sounds, which are distinguished by their temporal structure and fundamental frequencies (Michelsen et al., 1986).The fundamental frequency of queen piping varies with age, with toots ranging between 300–515Hz and quacks ranging from 215–320Hz. The fundamental frequencies of these signals stabilise 2d post-emergence for quacking and 4d post-emergence for tooting (Michelsen et al., 1986).

Both tooting and quacking are poly-syllabic signals, consisting of a sequence of similar syllables (Fig. 1.3.1). A tooting bout commonly presents with two long syllables, followed by several shorter syllables. The first few syllables of a quacking bout are often shorter, but can occasionally be longer than subsequent syllables. The initial syllables of a tooting bout go through a frequency sweep from 340–500Hz (Michelsen

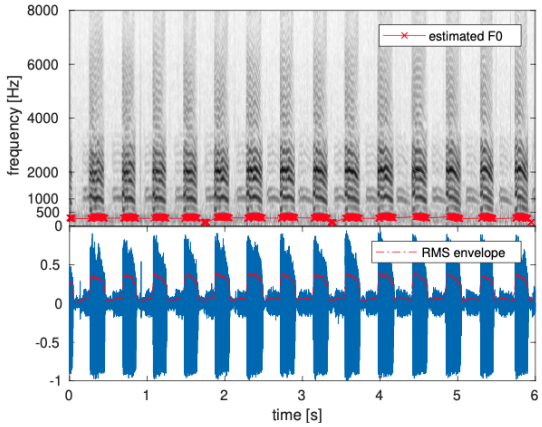
et al., 1986). Consecutive tooting syllables are similar in amplitude. In contrast, quacks are consistent in frequency but increase in amplitude over the first 5–8 syllables.

The established queen can pipe for several days leading up to swarming (Allen, 1956). The first virgin queen to emerge begins tooting, which elicits quacking from the queens still in their cells. The tooting signal only elicits quacking if it is loud enough, and the distance within which tooting will pass the quacking release threshold was estimated at 10cm (Michelsen et al., 1986). The audible distance is greater than this, and queen piping can be heard easily during hive inspections.

The definitive functions of queen piping are unknown. During queen replacement, the first queen to emerge can choose to depart with a secondary swarm or to take over the current nest, in which case she will need to kill the other virgin queens. Piping back and forth may allow rival queens to gauge each other's strength. Quacking



(A) Tooting



(B) Quacking

FIGURE 1.3.1: Queen tooting (A) and quacking (B). Spectrograms from Fourer and Orlowska (2022)

has been suggested to attract nurses that feed and defend confined queens (Kirchner, 1993), and playback of tooting sounds has been shown to delay the emergence of queens even in the absence of an emerged queen (Grooters, 1987). Workers freeze in response to queen piping (Michelsen et al., 1986), but the function of this behaviour is not known. A potential use could be to reduce movement-related noise on the comb, which appears to be the principal channel of communication for vibration signals. An inconsistency in the freezing threshold reported by Michelsen et al. (1986) and Michelsen, Kirchner, and Lindauer (1986) seems to support this hypothesis. Both studies reported freezing thresholds in response to comb vibration, but the threshold was 5 times higher in Michelsen, Kirchner, and Lindauer (1986) where background noise on the comb was also 5 times higher than in Michelsen et al. (1986). The authors suggest that bees may adapt their reaction to signals by accounting for background noise levels.

Worker piping

Worker piping has been described as a 'begging signal' (Frisch (1967) cited in Kirchner (1993)) or a 'stop signal' by Gould (1976). The signal is described as a 100ms sound pulse at 320Hz (Michelsen, Kirchner, and Lindauer, 1986). In Thom, Gilley, and Tautz (2003), piping occurred primarily during tremble dances, which are a type of dance performed by returning foragers who wander through the hive while shaking and repeatedly changing direction. The tremble dance is thought to recruit workers as food storer bees and is often performed by returning nectar foragers that cannot find a worker to accept their nectar load (Seeley, 1992; Thom, Gilley, and Tautz, 2003). Thom, Gilley, and Tautz (2003) were able to elicit worker piping by increasing nectar foraging, supporting Nieh (1993)'s suggestion that piping in this context can aid in food storer recruitment, as well as interrupting recruitment to new nectar sources, and so rebalances the colony effort spent on collection and processing of resources.

Worker piping also plays a role in swarm organization. Once a nest site has been chosen for a swarm, scouts begin to move quickly through the cluster while rapidly emitting ca. 0.8s long sound pulses rising from 100–200Hz to 200–250Hz over time (Seeley and Tautz, 2001). The timing and buildup of piping match that of the swarm warming to flight temperature before liftoff (Seeley and Tautz, 2001). Seeley & Tautz

point out that scouts pipe with their wings together, while workers in hives pipe with their wings held apart. Due to the change in resonant frequency of the oscillatory system, this likely modulates the frequency at which the piping signal is audible. Egg-laying workers in queenless colonies have also been observed to make piping sounds similar to queen quacking (Ohtani and Kamada, 1980). Guard workers may pipe in response to wasp attacks, and their piping has been compared to queen tooting (Pratt et al., 1996; Ohtani and Kamada, 1980).

Other in-hive vibrations

Other than communication signals, most behaviours likely have vibrational components, but their frequencies are rarely recorded. Walking on the comb, cleaning cells, and moving debris could create scraping sounds while wax working and food processing, like ripening honey between mandibles or storing pollen, might create sounds similar to chewing. Each individual bee is likely too quiet to be recorded, but thousands of bees together could audibly alter the soundscape.

Audio signatures of behaviours linked to wing beating have been recorded. Fanning, used to ventilate the hive and evaporate moisture during honey ripening, has a fundamental frequency of ca. 120Hz (118.6 $\pm$ 9.8Hz, Junge (2006)). Flight has a fundamental frequency between 200–250Hz (226.8 $\pm$ 12.8Hz, Vance et al. (2014); 222.4 $\pm$ 21.4Hz, Rodríguez Ballesteros et al. (2024)).

The fundamental frequency ranges of all behaviours discussed here are summarised in Table 1.3.1. How frequently each of these behaviours is performed changes with the colony's current state. If these individual changes compound to affect the collective soundscape of the hive, Machine Learning algorithms could be trained to detect colony state from hive sounds.

TABLE 1.3.1: Fundamental frequencies of in-hive sounds and vibrations. Frequencies given are ranges or mean ± standard deviation if available.

Behaviour		Frequency	Reference
Shaking signal		16.3 ± 5.8Hz	Gahl (1975)
		16 – 19Hz	Łopuch and Tofilski (2019)
Thoracic vibration	Pausing hive bee	30Hz	Hrncir, Maia-Silva, and Farina (2019)
	Active hive bee	36Hz	Hrncir, Maia-Silva, and Farina (2019)
	Forager	43Hz	Hrncir, Maia-Silva, and Farina (2019)
Fanning		118.6 ± 9.8Hz	Junge (2006)
		120Hz	Evans (2021)
Flight		220Hz	Evans (2021)
	Free-flying	222.4 ± 21.4Hz	Rodríguez Ballesteros et al. (2024)
	In flight chamber	226.8 ± 12.8Hz	Vance et al. (2014)
	Tethered	250Hz	Hasegawa and Ikeno (2011)
Waggle dance		250Hz	Esch (1961b)
		265Hz	Hasegawa and Ikeno (2011)
		250 – 300Hz	Michelsen, Kirchner, and Lindauer (1986)
Worker piping	In swarm, early	177 ± 22Hz	Seeley and Tautz (2001)
	In swarm, late	236 ± 20Hz	Seeley and Tautz (2001)
	"Begging" signal	320Hz	Michelsen, Kirchner, and Lindauer (1986)
	Modulation	270 – 540Hz	Thom, Gilley, and Tautz (2003)
	Queenright	330 – 430Hz	Pratt et al. (1996)
	Piping bee 1	337 ± 15Hz	Pratt et al. (1996)
	Piping bee 2	424 ± 16Hz	Pratt et al. (1996)
	Queenless	350Hz	Ohtani and Kamada (1980)
	Queenright	300 – 400Hz	Esch (1964)
	Guards	500 – 700Hz	Ohtani and Kamada (1980)
Queen piping	Quacking	306.6 ± 24.0Hz	Fourer and Orlowska (2022)
	Quacking (16d old)	245.7 ± 32.2Hz	Michelsen et al. (1986)
	Quacking (19d old)	306.9 ± 14.2Hz	Michelsen et al. (1986)
	Tooting	383.0 ± 61.5Hz	Fourer and Orlowska (2022)
	Tooting (16d old)	336.9 ± 34.7Hz	Michelsen et al. (1986)
	Tooting (19d old)	488.3 ± 25.0Hz	Michelsen et al. (1986)



1.3.3 Measuring in-hive vibrations & sounds

Vibrations in the beehive have been measured in a variety of ways, including oscillographs (Esch, 1967), accelerometers (Bencsik and Newton, 2018; Ramsey, Bencsik, and Newton, 2017; Smith and Chen, 2017), laser-doppler vibrometry (Michelsen, Kirchner, and Lindauer, 1986; Nieh and Tautz, 2000; Tautz, Casas, and Sandeman, 2001; Hrncir, Maia-Silva, and Farina, 2019), microphones (Favre, 2011; Ferrari et al., 2008; Tan et al., 2016), and high-speed video (Łopuch and Tofilski, 2017). Taking accurate recordings is made more complicated by a range of factors, including the presence of background noise (such as environmental noise and the comparatively loud sounds of flight and fanning), the difficulty of operating in or near the hive without disturbing the colony, data transmission and storage constraints, and the tendency of honeybees to cover any foreign objects in the hive with wax and propolis. Signal characteristics can also vary with the emitter's phenotype, including age (Michelsen, Kirchner, and Lindauer, 1986) and wing length (Kirchner and Sommer, 1992).

Each method presents a trade-off between different aspects. Laser-doppler vibrometry allows for contactless point measurements of substrate vibrations with very high accuracy, avoiding modification of the resonant properties of the measured object. However, the required equipment is expensive, and the method is highly sensitive to externally introduced vibrations. The experimental setup must therefore be closely controlled. Additionally, it is not possible to follow a bee for long distances across the comb. Therefore, this method is most suitable for recording comb vibration, or bee vibrations over short time periods when they are not moving quickly.

Video recording allows linking vibrations directly to the emitter (including of several emitters simultaneously) and provides valuable behavioural context to the recording. However, at the high speeds necessary, this method generates large amounts of data, often limiting the recording time (Łopuch and Tofilski, 2019).

Less storage-intensive methods such as accelerometer and microphone recording can generate long-term time series or even continuous recordings, but do not allow for the same resolution and attribution of recorded signal to emitter. Instead, they aggregate sounds within a radius around the sensor. Microphones capture airborne sound, while accelerometers measure substrate vibration.

1.4 Hive Monitoring

The wide range of behaviour-linked vibrations present in honeybee hives and tightly regulated in-hive microclimate potentially contain information about a range of colony conditions. Hive monitoring aims to exploit this information using in-hive sensors (IHSs) and AI to increase colony outputs (pollination and honey production) and improve survival. Sensor-based hive monitoring applications vary in target client (hobby beekeepers to large-scale commercial apiarists) and monitoring aims, from insights into hive wellbeing through tracking honeyflows to quality assurance for growers during pollination contracts.

Hobbyists typically care for their hives as individuals and will be interested in ways to gather information on their condition and wellbeing between inspections. Sensor-based hive monitoring can provide detailed data for human interpretation without using predictive models (operational monitoring sensu Zaman and Dorin (2023)). This style of monitoring offers access to plain data and simple graphs, allowing beekeepers to track key hive metrics such as in-hive temperature and weight. This allows the beekeeper to connect with and watch over their hives even from outside of the apiary.

Commercial beekeepers have a less individualised approach. Inspections have particular targets depending on the time of year. If a hive meets a few criteria, it is treated in one way; if it does not, it is treated another way. Interventions, such as anti-mite treatments or supplemental feeding, will generally be distributed to all colonies at the same time, or to all colonies meeting simple criteria. Here, statistical or ML models that predict different aspects of colony health from time series of sensor data (predictive hive monitoring sensu Zaman and Dorin (2023)) have the potential to save material and labour costs. Recent decades have seen much effort in academic research and the private sector towards predictive hive monitoring and precision apiculture. If sufficiently accurate and applicable across hives and apiaries, such predictions could streamline colony maintenance by flagging hives requiring intervention.

Since the length of commercial inspections has already been reduced to a minimum, the most useful models would detect conditions which take longest to assess.

Two such targets are queen presence and food stores. Both typically require removing several frames to check for the presence of fresh eggs and for sufficient amounts of pollen and honey. Commercial beekeeping practice favours quick decision making, and therefore may overcorrect these conditions — requeening hives or providing supplemental feeds when it is not necessary. Feed, including honey and pollen supplements or substitutes, is by far the highest area of expenditure for US beekeepers, exceeding $46m in 2024 (NASS, 2024b). A more streamlined approach to feed distribution could reduce this expense and lead to substantial savings for commercial beekeepers.

1.4.1 Sensor technology

Overall, many commercial and research applications of hive monitoring utilise a similar set of sensors. Most systems record in-hive temperature since it is directly regulated by bees to 34-36° C, the ideal range for brood development (Kronenberg and Heller, 1982; Fahrenholz, Lamprecht, and Schricker, 1989). Most systems also include a humidity sensor. Bees actively regulate hive humidity, likely by fanning and water evaporation, and excessively high or low humidity can indicate that the colony is not able to maintain an optimal in-hive environment (Li et al., 2016).

Another common sensor type are microphones, which record sounds at the entrance or inside of the hive. Audio is a rich data source and may hold information about a variety of colony states, since honeybees are known for their sophisticated vibrational communication (section 1.3.2). Audio is recorded as a one-dimensional time series of air pressure changes, but often represents a complex soundscape with multiple overlapping sources of sound that can each vary in time. Data processing and feature engineering are required to decompose recordings into frequency components and capture meaningful properties.

Some systems also include hive scales. By measuring hive weight, algorithms claim to detect honeyflows (periods of high nectar availability), nectar dearth, and robbing. More nuanced assessments of colony condition are difficult, however, since the weight of colony stores cannot easily be separated from the weight of the population. Furthermore, the weight of colony stores only gives limited insight into the colony's nutritional state as it contains no information on the nutritional value of the

incoming nectar, nor about pollen amount or quality. Nevertheless, heavy colonies tend to be in better condition than lighter colonies, so weight can be considered a key metric for hive monitoring (McMinn-Sauder et al., 2024).

Other systems use cameras for targeted surveillance of the hive entrance, eg to detect *Varroa* mites or measure foraging traffic. Foraging activity can also be measured with entrance sensors that funnel bees through infrared-gated tunnels, allowing estimation of the number of in- and outgoing bees (Evans et al., 2024; *iPollinate* 2025).

Hive monitoring sensor technology has been comprehensively reviewed in Zaman and Dorin (2023).

This project utilised proprietary IHS provided by my industrial partners BeeHero and Canetis SRL. These Internet of Things (IoT) sensors include two microphones as well as temperature and humidity sensors alongside other data loggers. The frequency response of the microphones was not tested, since the aims of my project did not include directly measuring the amplitudes of certain frequencies. Instead, I assumed that the frequency response would be uniform across different IHSs, so that the recorded spectra would be comparable.

IHS sampling is directed by internet-enabled 'gateways', which prompt the IHS to take recordings at a set schedule and sampling parameters and upload samples to Amazon Web Services (AWS) cloud storage. This reduces the need for local storage, facilitating long-term recording, but can lead to loss of data when upload bottlenecks occur due to the small onboard memory of the IHS. The microphones are fully enclosed and therefore minimally affected by propolis or humidity.

1.4.2 Prediction targets in hive monitoring

Honeybee colony health can be broken down into interacting aspects including:

- Population size and foraging strength
- Queen condition and productivity
- Brood production
- Parasite and disease load
- Honey and bee bread stores

- Size of nest cavity compared to colony size

While each of these factors can be measured during inspections, some are harder to manipulate experimentally than others. Capturing sensor data from colonies with each possible combination of health factors in a controlled way with sufficient replication is close to impossible. Therefore, hive monitoring studies tend to investigate one or two factors while aiming to control for the others. This is where a lack of unification in the concept of hive health may arise — many people are claiming to measure and monitor hive health, but using different sensor systems and predictive targets (Zaman and Dorin, 2023).

Swarming is a popular target for monitoring studies using audio data because this process has both a high impact on apiary profitability and a well-defined sound component. Murphy et al. (2015) reported an increase in audio magnitude prior to swarming, while Zacepins et al. (2016) used single-point temperature monitoring to detect the ca. 20 minute warm-up phase immediately before swarming. Ramsey et al. (2020) claim that their ML algorithms can discriminate sounds from swarming and non-swarming hives up to 30 days before a swarming event. Queenlessness is another common prediction target, and recent studies in this area are discussed in the following section.

Other studies focus on the technical aspects of hive monitoring, like Paffhausen et al. (2021), who recorded the electromagnetic oscillations caused by bees' electrically charged bodies during the waggle dance, Terenzi et al. (2019), who demonstrate several techniques for sound feature extraction and their use in detecting the queenless state, or Chazette, Becker, and Szczerbicka (2016), who proposed a computer vision-guided laser for *Varroa* mite control. Hive monitoring studies are comprehensively reviewed in Meikle and Holst (2015) (for work up until 2014) and Abdollahi, Giovenazzo, and Falk (2022) (for 2012–2021).

Cross-hive generalisation

Many studies have attempted to develop methods for sound-based or multi-modal honeybee colony monitoring. Most of them report high prediction accuracy, but lack rigour in experimental design and model testing. The most common problems include

small sample sizes (number of colonies), not controlling for time-linked changes in colony state, and using a random test set. For example, Liang (2024) incorporated data from 25 colonies into the development of their sound- and video-based colony condition detection model. Despite the large number of colonies available, the model was tested on a random subset of recordings across all colonies, achieving an accuracy of over 90%. However, this type of test is sensitive to temporal autocorrelation within the dataset and does not reflect the model's real-world performance.

Very few studies have examined the performance of their models on data from hives outside of their training set (out-of-group (OOG) generalisation). Kulyukin, Mukherjee, and Amlathe (2018) tested their models on unseen hives, as well as hives at a different location, demonstrating generalisability to new hives as well as new apiaries. However, their prediction target was sound type, classifying recordings as bee sounds, cricket sounds, or ambient noise.

Orlowska et al. (2021) attempted queen detection in a combined dataset from 6 hives. Using data augmentation, they achieved a maximum accuracy of 96% on unseen hives[2]. While the technical approach and attempt to test on unseen hives were promising, the underlying dataset was highly imbalanced both across hives and across classes. In one test set, the minority class made up only 0.4% of the majority class size[3]. A biased model could achieve very high accuracies on this test set without any true ability to generalise, simply by predicting one output value no matter the input data. This, combined with the failure to report mean accuracy metrics, puts the reliability of the results into question.

Nolasco et al. (2019) achieved up to 94% Area Under Curve (AUC) when testing their queen detection models on unseen hives. Work on the same data and task by Terenzi et al. (2022) achieved a mean AUC of 77%. The dataset was limited in scope, containing data from only two hives and across 4 days, so it is not possible to draw conclusions about the wider applicability of the model.

Ruvinga et al. (2023) also aimed to detect queenlessness using a dataset spanning 5 hives across two sites. Their models were trained on 4 Italian hives (2 of each subspecies *ligustica* and *carnica*) and tested on a fifth hive kept in Surrey. Crucially, this

[2]The authors do not report the mean accuracy of their models.

[3]In the three other test sets, the minority class had 10.0%, 54.9%, and 57.2% of the samples of the majority class, respectively.

test hive always had a mated queen present throughout the recording period. Since binary classification models cannot be tested meaningfully on a test set lacking one of the classes, the high accuracy reported by the authors unfortunately is not evidence of cross-hive generalisation.

Most recently, Bricout et al. (2024) were unable to achieve cross-hive generalisation with their queen state prediction models. Instead, they proposed a model trained to detect a relative change in state by comparing new audio recordings to a reference recording.

Generalisation of colony state prediction models to unseen hives is clearly an ongoing challenge which will only be overcome with robust experimental methodologies, large multi-hive datasets, and rigorous approaches to model testing.

1.5 Outline

To work towards functional hive monitoring, we need to characterise the barriers to cross-hive generalisation. In this thesis, I use two novel multi-hive datasets of in-hive sound and microclimate data, collected by me and my industrial supervisors at Canetis SRL, to train ensembles of neural networks for colony state detection. I then investigate in-depth the performance of these models and use feature importance metrics to link colony behaviour and model outputs to answer the questions:

- What factors influence the performance of colony state prediction models?
- What constitutes best practice in hive monitoring, including experiment design, data collection, and model testing?

In Chapter 2, I train neural networks for queen loss detection and evaluate the impact of different parts of the model pipeline on prediction accuracy. I address how training set size and diversity, microphone position, feature extraction method, and hive identity affect model performance. Further, I demonstrate a stark difference in model accuracy between within-group (WG) and OOG tests, showing that the performance of most existing models likely has been overestimated. These results show how different parts of the hive monitoring pipeline impact model performance.

In Chapter 3, I analyse in-depth the performance of the best models from Chapter 2. I show how soundscapes vary between hives and with hive metrics such as colony strength and brood counts. When testing different methods for improving prediction accuracy, such as thresholding or subsetting by time of day, I find that their efficacy varies substantially between hives. I use SHAP analysis to extract feature importance information, linking model predictions to colony-level behaviour changes during queen loss and queen rearing. My results show how post-processing of predictions can improve model performance, and how feature importance information can inform feature choice for future models.

In Chapter 4, I present a large multi-hive dataset of sound and microclimate data from colonies affected by nutritional and disease stress. The dataset includes precise phenotypic annotations from inspection data, including brood counts and colony strength assessments. I analyse the effect of both sources of stress on brood production and develop a colony nutritional state label based on brood counts. I then use neural networks to detect colony strength and nutritional state from sensor data and evaluate model performance. These results demonstrate the impact of prediction task choice on model performance and inform best practice in experiment design and dataset generation.

Finally, I draw on my results across chapters to identify factors influencing model performance, both regarding model specification and prediction task design. I also develop a set of guiding principles and best practice in honeybee hive monitoring and propose various new avenues of research. With a deeper understanding of how aspects of dataset generation, model design, and task choice impact model performance, we will be able to develop models with greater capacity for generalisation and potential for real-world applications.

Chapter 2

Robust design and testing of hive monitoring models

Contributors:

Stella Marie Felsinger (study design, data analysis, interpretation of results, writing),

Huw Evans (experiment design, data collection),

Sandra Evans (experiment design, data collection),

Beth Mortimer (supervision, interpretation of results),

Geraldine Wright (supervision, interpretation of results).

2.1 Abstract

Acoustic hive monitoring is a growing field of research with the potential to improve beekeeper efficiency and honeybee welfare. Despite increasing complexity in sensor data processing pipelines and ML model architectures, models are rarely tested for their ability to generalise to unseen hives. Factors influencing model generalisability are poorly understood and models are insufficiently tested, currently limiting real-world applicability.

I developed models for detecting queenlessness in a novel dataset of in-hive sounds from 10 honeybee colonies, 8 of which experienced queen loss. I tested the effect of feature extraction method, sensor position, number of hives in the training set, and test hive identity on model performance on unseen data from training hives (WG test) and on data from fully unseen hives (OOG test). My results reveal significant hive dependence in how sensor position and feature set affected model performance. Models

trained on data from multiple hives (multi-hive models) performed better OOG, but worse WG, than models trained on single hives (single-hive models), showing that increased training set size and diversity improved cross-hive generalisability.

These results suggest that in-hive soundscapes differ substantially between hives, possibly mediated by hive organisation (chiefly the location of the brood nest) or aspects of colony health, such as colony strength or brood rearing capacity. They also show that the performance of existing models has been vastly overestimated. Further research into how these factors impact model performance is essential to develop models that are robust across apiaries and seasons.

2.2 Introduction

As sensor components and data storage become more affordable, biological research increasingly captures bigger and more complex datasets. Storage-hungry data types, such as video and audio, have especially benefitted from these technological advances, enabling significant expansion of spatial and temporal recording coverage (eg Roe et al. (2021)). The advent of deep learning has transformed the exploitation of these big datasets, offering performance gains over traditional methods (Tang et al., 2019).

First proposed in the mid-19th century, Artificial Neural Networks (ANNs) are now a fundamental method within Machine Learning (Thompson et al., 2021; Sarker, 2021). ANNs are based on 'nodes' or 'neurons', which are effectively functions combining multiple weighted inputs into a single output. By combining nodes into layered networks, an almost infinite range of functions can be approximated. Among other applications, ANNs are well suited to pattern matching or classification tasks, where samples are assigned to one of a small set of target classes (Vinuesa et al., 2014).

By passing training data with known output labels through the network, the ANN can 'learn', ie each node's input weights are tuned so that the network produces the correct output (supervised learning). Training occurs in multiple rounds and each round, or pass of the training data through the network, is called an epoch. After training for enough epochs, large networks can learn to replicate almost any function, from simulating ecosystem dynamics to recognizing faces or generating human-like text (Bonnaffé, Sheldon, and Coulson, 2021; Wang and Deng, 2021; Lund et al., 2023).

The size of the network can be varied to suit the data structure and task complexity. Single-layer perceptrons (SLPs) and multi-layer perceptrons (MLPs) are two subtypes of ANN architectures distinguished by the number of layers between input and output layers (hidden layers). Increasing the number and size of hidden layers typically improves the ability of a model to learn, but also increases training time. In addition, larger networks can be more prone to overfitting, where models make highly accurate predictions on training data but fail on new data. Training length (ie number of epochs) similarly presents a trade-off between optimising predictions on the training set and generalisation to unseen test data, which attempts to measure the model's real-world applicability. If training stops too early, model performance on both training and test set will be low; if training runs for too long, the model will perform well on the training set but will fail to generalise. To prevent overfitting and accurately assess model performance, datasets are split into training, validation, and test sets. Dividing the dataset in this way allows us to monitor training progress using the validation set and to evaluate the model on unseen data using the test set.

Model development requires designing the network architecture, as well as making choices about how the model should be trained. Together, these settings are referred to as hyperparameters (Tab. 2.2.1). Identifying a good combination of hyperparameters via grid search or iterative changes is a prerequisite for successful model training.

Other hyperparameters are domain-specific. In acoustic monitoring of honeybee colonies, they involve choices about microphone placement, how to account for hive dependency during train-validation-test splitting, and how to extract features from audio samples. The effects of these choices have largely gone unaddressed so far. Other unresolved questions surround the effect of hive individuality on the success of model training and testing.

Recent decades have seen advances in building models for acoustic hive monitoring (reviewed in Abdollahi, Giovenazzo, and Falk (2022)). While many studies report high model accuracies, their datasets usually only include a few colonies and neglect to test models on unseen hives, failing to demonstrate generalisability (Bricout et al., 2024; Abdollahi, Giovenazzo, and Falk, 2022). A handful of studies do evaluate queen detection models on unseen hives with varying success (Nolasco et al., 2019; Orlowska

TABLE 2.2.1: Common hyperparameters.

Term	Definition
Node/Neuron	Non-linear function applied to a weighted sum of inputs (features or outputs of previous layer)
Layer	Set of parallel nodes
Layer size	Number of nodes in a layer
Hidden layer number	Number of layers between input and output layer
Loss function	Calculates the difference between model output and ground truth
Optimizer	Algorithm for adjusting weights based on loss
Learning rate	Determines how much weights are changed at each step
Batch size	Number of training examples passed to the model at the same time
Epoch	Single pass of all training examples through the network
Max. epochs	Maximum number of epochs after which training is terminated
Patience	Number of epochs after which training is terminated if performance on validation set does not improve

et al., 2021; Terenzi et al., 2022; Ruvinga et al., 2023).[1] However, these studies were limited by small sample sizes taken over short intervals, which limits their power to make predictions over a wider population.

In this chapter, I test five key hypotheses concerning the impact of data collection and model development on model performance.

Hypotheses

H1 Test set — **Model performance is lower when tested on an unseen hive, compared to testing on unseen recordings from training hives.**

The majority of existing hive monitoring studies test their models on a random sample of recordings held-out during training. I assumed that this practice strongly inflates model performance over what could be expected in a realistic deployment scenario.

[1] Some of these studies call their testing scheme 'hive-independent' — a misleading term, since the data is in fact split in a hive-dependent manner.

H2 Sensor position — **Recordings from the centre of the hive carry the most information about hive state.**

Density of workers and behaviours performed differ throughout the hive. Depending on the predictive task, the most relevant location in the hive will vary. For tasks concerning the internal state of the hive, such as queen presence, I assumed that the information content will be highest in the centre of the hive where the cluster and brood nest are located.

H3 Number of training hives — **Models trained on many hives generalise better than those trained on a single hive.**

Many datasets in the literature contain data from very few hives. Training on one or a few hives could produce models with power to generalise to new hives if in-hive soundscapes don't differ substantially between colonies. Conversely, if in-hive soundscapes are highly heterogeneous, a large number of training hives would be necessary to achieve generalisation. I assumed that in-hive soundscapes differ between colonies and therefore that including a greater number of hives in the training set would improve performance on unseen hives.

H4 Feature engineering — **ResNet encodings capture more information from spectrograms of raw audio data than frequency and waveform-derived features.**

Spectrograms resolve an audio recording in both the frequency and time domains, whereas Power Spectral Density (PSD) represents only the frequency domain, averaging over the time dimension. I assumed that ResNet encodings of two-channel spectrograms are more information-rich and therefore more suitable as features, compared to features derived from waveform and frequency representations lacking the time domain (PSD at 4Hz resolution, zero-crossing rate, and spectral centroid).

H5 Hive individuality — **Model performance varies between hives.**

Honeybee colonies are superorganisms made up of thousands of individuals. Their genetic makeup, together with environmental and hive health factors such as disease, queen presence, and nutrition shape the emergent properties of the colony (Winston, 1987). This can lead to marked differences in the temperament and behaviour of hives within the same apiary (Wray, Mattila, and Seeley, 2011).

I assumed that these hive characteristics would affect in-hive sound, leading to differences in model performance depending on the test hive.

Each of these hypotheses is addressed in turn below (Section 2.4).

2.3 Data analysis

2.3.1 Data collection

In-hive sound was recorded from 10 honeybee hives (*A. mellifera ligustica* hybrids) between 27. July – 26. September 2023 by Huw and Sandra Evans (Canetis SRL) in Tuscany, Italy. The study site was far from urban centres and roads, limiting ambient noise other than nature sounds. All colonies were housed in Langstroth hives with 2 deep boxes with 10 frames per box, no honey supers and no queen excluder. Hives were arranged on 4 pallets (A-D), each holding up to 4 hives (1-4). 10 healthy hives with comparable population sizes were selected for the experiment.

A BeeHero IHS was placed on every second frame of the bottom box so that each inter-frame space was recorded from by one microphone (Fig. 2.3.1). The recordings from both microphones of one sensor were treated as being in the same position. Sensor positions were named to reflect the location within the hive ('Outside', 'Outside 2', 'Centre', 'Inside 2', and 'Inside'), with the 'Outside' position closest to the hive entrance. IHS were placed into hives on 22. July to allow for habituation and incorporation before recording began. The IHS were set to sample 16s of in-hive audio every 15min at 1kHz.

Population size, brood area, and honey and pollen stores were recorded fortnightly using Liebefeld grading (Imdorf et al., 1987; Dainat et al., 2020). This method estimates the area covered by each cell type and by worker bees to a precision of 1/8th of a frame.

Queens were removed on 5. August from all but 2 control colonies (D1 & D3), following which all queenless colonies began rearing replacement queens. I used a continuous subset of recording days spanning 27. July – 17. August to include recordings from the week prior to queen removal and the entire queen rearing period until the day before queen emergence.



FIGURE 2.3.1: 5 BeeHero in-hive sensors in bottom box of hive. The hive entrance is at the right end of the flyboard at the bottom of the image. The rightmost IHS was designated the 'Outside' sensor and the leftmost the 'Inside' sensor. (Photo by Huw & Sandra Evans)

2.3.2 Feature selection

An important early step in ML model development is choosing how to present raw data to the model. Digital audio recordings are typically represented as 'waveforms', strings of binary numbers encoding the changes in air pressure caused by an acoustic sound wave. While this waveform can be analysed directly, typically it is pre-processed into a more compact data representation (Okawa et al., 2019). Common representations include frequency spectra (frequency x amplitude) or spectrograms (time x frequency x amplitude), using transforms like the Short-Time Fourier Transform (STFT) or Continuous Wavelet Transform (CWT) (Sejdić, Djurović, and Jiang, 2009). These representations can be used as features directly or further processed by hand, algorithm, or deep learning. Handcrafted features are usually quantities that can be linked to biological or physical meaning. This makes them human-interpretable, but also limits them to the scope of human imagination and conceptualisation. By contrast, deep-learned features, often extracted using Convolutional Neural Networks (CNNs), can translate both macro- and micro-patterns in the data into higher-level

features, which are more information-dense than the original representation. Deep features can rely on multiple pixel-level differences across the image that would be imperceptible or meaningless to the human eye. These contrasting approaches to feature engineering broadly represent a trade-off between interpretability (favouring handcrafted features) and feature complexity (favouring deep features) (Lin et al., 2020).

In order to test which of these feature engineering paradigms is more suitable to the task of detecting queen state from in-hive audio, I trained models on two feature sets (see H3, section 2.2). A handcrafted feature set included frequency- and waveform-based features, namely the zero-crossing rate, spectral centroid, and power spectral densities at ca. 4Hz intervals, derived from single-channel recordings. A second, deep feature set was derived from two-channel spectrograms, chosen to provide a more complex and information-dense representation of in-hive sound including the time dimension. The spectrograms were processed using a pre-trained convolutional neural network for image analysis, and model outputs were treated as linear encodings of spectrograms.

Both feature sets reflect dimension-reduced vector representations of the raw audio data, with 135 and 512 features respectively. Below, I detail the processing steps to arrive at each representation. All data cleaning, analyses, and visualisations were performed in Python v.3.10.9 (Van Rossum and Drake, 2009). The main libraries used were `pandas v2.0.3` (The pandas development team, 2020), `numpy v1.26.4` (Harris et al., 2020), `librosa v0.10.1` (McFee et al., 2023), `scipy v1.10.0` (Virtanen et al., 2020), and `sklearn v1.0.2` (Pedregosa et al., 2011) for data processing and feature extraction; `keras v3.0.5` (Chollet et al., 2015), `pytorch v2.4.1` (Paszke et al., 2019), and `tensorflow v2.16.1` (Abadi et al., 2015) for model training and testing; and `matplotlib v3.7.1` (Hunter, 2007) and `seaborn v0.13.2` (Waskom, 2021) for visualisation.

Handcrafted features

My first feature set consisted of hand-engineered frequency- and waveform features. All audio files recorded during the experiment were saved to a folder with a subfolder for each sensor. The date and time of the recording (in UTC) and the IHS MAC address were extracted from the filenames using regular expressions and these

metadata were saved as a separate dataframe. The two concurrent recordings from each sensor were treated as separate samples for this feature set. From each sound file, the averaged PSD (`scipy.signal.welch()`), the per-second zero-crossing rate (ZCR) (`librosa.feature.zero_crossing_rate()`) and the per-second spectral centroids (`librosa.feature.spectral_centroid()`) were calculated. The PSD was extracted using Welch's method by calculating periodograms over several overlapping segments of the recording. The segment length was left at the default setting of 256 and the sampling frequency was set to the native sampling rate of the recordings (1000Hz). These settings resulted in a resolution of $\Delta f = 3.9$Hz for the PSD. To obtain per-second spectral centroid and ZCR features, the frame length was set to the sample rate and the hop length to half of the sample rate. This way, the ZCR and spectral centroids were calculated on 1s frames with an overlap of 0.5s between consecutive frames. The three groups of features were extracted from each sound file, labelled with the matching metadata, and saved to a new audio features dataframe including 33 ZCR columns, 33 spectral centroid columns, and 129 PSD columns covering the range of 0–500Hz ($\Delta f = 3.9$Hz). I calculated the mean, standard deviation, variance, maximum, minimum, and range of the ZCRs and spectral centroids for each sound file and saved these as additional columns. The dataframe was merged with IHS serialization data to label each sound file's feature vector with the sensor position, hive ID, and hive location. I calculated feature cross-correlation to choose which summary statistics of the per-second ZCRs and spectral centroids to include in the final handcrafted feature set.

Deep features

As the second feature set, I used ResNet encodings of spectrograms. After extracting the metadata from the audio file names as described above, each recording was transformed into a Mel spectrogram using the function `feature.melspectrogram()` in the `librosa` package (n_mels = 64, n_fft = 1024, and hop length = 256.) The mean and standard deviation of each pixel value were calculated across all recordings. These statistics were used to normalise the spectrograms.

The spectrograms were resized to 224x224 pixels. Two simultaneous recordings

taken by each microphone, one from each leg of the sensor, were treated as two channels of the same sample and stacked with a third, blank channel. The blank channel was added to create a 3x224x224 array, conforming to the data shape expected by the pretrained ResNet.

I used the smallest version of ResNet (ResNet18) with pre-trained weights from ImageNet1K_v1 available in `PyTorch` (Paszke et al., 2019; Deng et al., 2009). The final 1,000–neuron classification layer of ResNet18 was replaced with an identity layer copying the previous 512–neuron layer. After passing a spectrogram through the ResNet, the outputs of the final identify layer were used as the spectrogram's deep feature representation.

2.3.3 Model design and testing

Model architecture and hyperparameters

I used a SLP architecture with a single hidden layer. Models were built and trained using the `keras` API for `TensorFlow`. I used the Adam optimiser and binary crossentropy as the loss function. I performed preliminary testing at different batch sizes, patience settings, and hidden layer sizes[2]. Preliminary testing showed that networks with 128 or fewer nodes in the hidden layer produced a poor fit to the data. Therefore the hidden layer size was increased to 256.

Batch size and patience parameters had little effect on convergence speed or model performance. I used patience = 5 and batch size = 32, which are standard settings. If the early-stop criterion was not triggered, training was capped at 300 epochs. Once this set of suitable hyperparameters had been established, they were held consistent throughout to allow for direct comparisons between different sensor positions, feature sets, and training set sizes (H2, H3, and H4, section 2.2).

Learning rate

I trained models across five learning rates from 1e-2 to 1e-6. Learning rate was not found to affect training or performance during preliminary testing (Fig. A.1.1). 1e-1

[2]Refer to Table 2.2.1 for definitions of these terms.

was also tested, but these models often failed to converge within the 300 epoch limit. Therefore, this learning rate was dropped from the range.

Test metric

I used Area Under Curve (AUC) to assess the performance of my classification models. This metric is the integral of the Receiver Operating Characteristic (ROC) curve (Fig. 2.3.2). The ROC curve is constructed by calculating a model's true positive and false positive rates while varying the detection threshold, the threshold used to binarise the continuous model output and discriminate between positive and negative predictions. The true positive rate is then plotted against the false positive rate at each threshold setting.

Good classifiers have an ROC curve rising steeply to a plateau close to $y = 1$ with an integral close to 1, while the ROC curve of a random classifier would plot a diagonal from (0,0) to (1,1) with an integral close to 0.5. AUCs of 0.5 represent a classifier with the same performance as a random coin toss, while an AUC of 1 would indicate a perfect classifier.

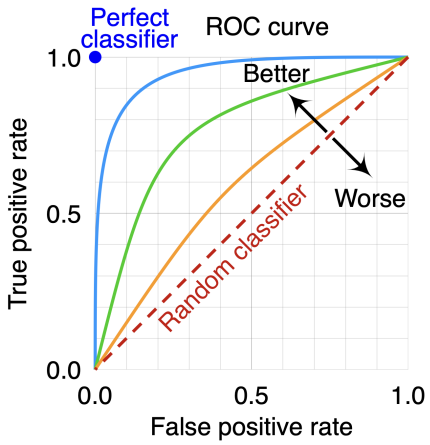


FIGURE 2.3.2: Illustration of ROC curves of models with random and good performance. Red dashed line indicates random classifier with 50% AUC and blue line indicates a very good classifier approaching perfect AUC. Figure "Roc curve.svg" reused from "Receiver operating characteristic" on Wikipedia. Authors: cmglee, MartinThoma. Date: 24 June 2018. Licence: Creative Commons Attribution-Share Alike 4.0 International. Retrieved on 6 Nov 2024.

2.3.4 Hypothesis testing

H1: Test set

To estimate the performance of a model under field conditions, it is important to avoid information leakage between the training and test sets. Due to a high potential for autocorrelation, data from the same hive being present in both the training and test set constitutes such a source of contamination. To better understand the effect of test set contamination on model performance, every model was tested both within-group (WG) and out-of-group (OOG).

WG tests were conducted using a random test set from the same hive or set of hives used to train the model. For single-hive models (SHMs), a random 20% of recordings were sampled as the test set, and 20% of the remaining data were used as a validation set to monitor model convergence (Fig. A.1.2b). For multi-hive models (MHMs), 20% of the pooled recordings from the 9 training hives were evenly split into a test and validation set. To maintain the same balance of queenless and queenright samples in each set, all splits were stratified by class (using the `stratify` keyword in `sklearn.model_selection.train_test_split()`).

OOG tests were conducted on data from an unseen hive. Using a hold-one-out strategy for MHMs, each model was tested on one unseen hive (Fig. 2.3.3). SHMs were tested on all other hives, including control hives. MHMs could only be tested on treatment hives containing both pre- and post-queen removal samples, since both control hives (D1 & D3) were used during training.

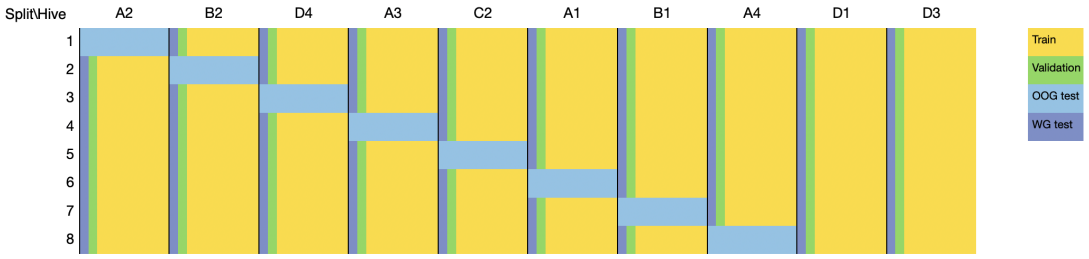


FIGURE 2.3.3: Scheme for assigning data to train, validation, and WG/OOG test sets for multi-hive models. In each split (rows), one hive (columns) was held out as an OOG test set (light blue). From all other hives, 10% of all data were used as WG test set (dark blue) and another 10% as validation set (green). The remaining data were used as the training set, including all data from control hives D1 & D3. The size of WG test set and validation set is equivalent to column proportions.

H2: Sensor position

I tested 4 sensor position settings: 'all sensors', 'Inside', 'Centre', and 'Outside'.

Preliminary testing with handcrafted features and a smaller network architecture (SLP with 128 nodes) showed that model performance was variable within each position but did not differ significantly between positions (see Fig. A.1.1). Since I expected the soundscape at the two intermediate positions 'Inside 2' and 'Outside 2' to be rather similar to the neighbouring positions, I chose to leave them out. The data from these positions were nevertheless included when testing the 'all sensors' position setting.

H3: Number of training hives

To evaluate the effect of hive diversity within the training set, I trained two types of models. Single-hive models (SHMs) were trained on data from a single hive, while multi-hive models (MHMs) were trained on data from 9 hives (section 2.3.4). I then compared the performance of both model types on data from unseen hives.

H4: Feature selection

Models were trained and tested with handcrafted features and deep features (see section 2.3.2) to determine the effect of feature representation on model performance.

H5: Hive identity

I compared the performance of SHMs and MHMs across different test hives to understand how hive identity and the previously mentioned parameters would affect the ability of a model to generalise to unseen hives.

2.3.5 Data processing

I created custom dataset classes to handle subsetting, folding, and preprocessing uniformly across all model types and variable settings. One dataset class was created for handcrafted features and another for deep features to adjust to the differences in feature number and data representation, however the procedures for subsetting and folding were almost identical.

If required, the data were subset to the sensor position used. Data were then filtered by hive, and the appropriate features were selected. Since the class imbalance within each hive was not strong, I chose not to balance single-hive datasets and allow for some unevenness in the number of samples from each class in order to include more samples. For multi-hive models, both classes and hives were balanced by finding the number of samples in the smallest hive-class subset and randomly sampling the same number of samples from all other hive-class subsets.

When using handcrafted features, I fitted a standard scaler to the training data and scaled the features of the validation and test sets using the same fitted scaler. This was done after splitting the dataset to avoid information leakage between sets.[3] The deep features were already scaled to between 0–1.

For each of the 2 feature sets, 8 queenless hives, 4 sensor position subsets, and 5 learning rates, 3 replicate models were trained to assess repeatability. For each combination of variables, I trained a single-hive model as well as a multi-hive model. I therefore trained 15 models per hive, 120 models per sensor position, 480 models per feature set, and 960 models per model type, for a total of 1920 models trained and tested.

Any models that failed to converge within 300 epochs were discarded. Of 1920 combinations of hyperparameters and replicates, there were 1319 converged models.

[3] If features had been scaled prior to splitting, information from the validation and test set would have been used to scale data in the training set.

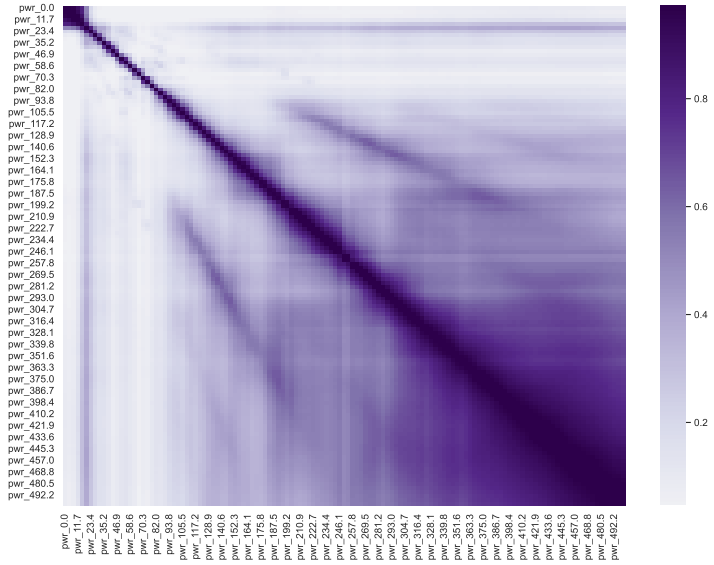
2.4 Results

2.4.1 Feature correlation

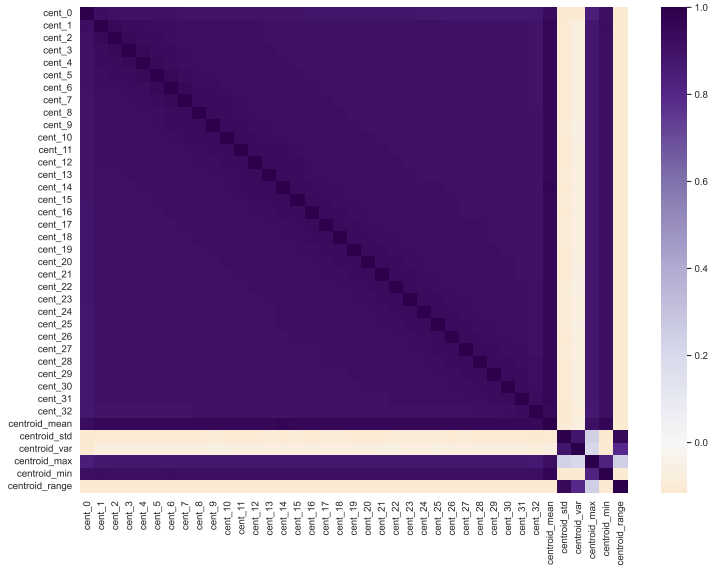
I calculated cross-correlation coefficients (using the `.corr()` method in `pandas` and Pearson correlation) between the features within each subset of handcrafted features. Data from all hives and all hive states were compiled to ensure that the selected features held relevance across the full dataset, helping to achieve generalisation (Fig. 2.4.1).

Power Spectral Density (PSD) cross-correlation strength increased with frequency (Fig. 2.4.1a). PSDs at very low frequencies up to ca. 16Hz were highly correlated with each other. PSDs at 16–24Hz appeared slightly correlated with frequencies above ca. 100Hz. Correlation strength between neighbouring features (the left-right descending diagonal in Fig. 2.4.1a) also increased with frequency. Between 100–200Hz, the diagonal was flanked by two diverging bands of higher correlation strength, likely representing harmonics of these frequencies. I chose to include all frequency features in the analysis.

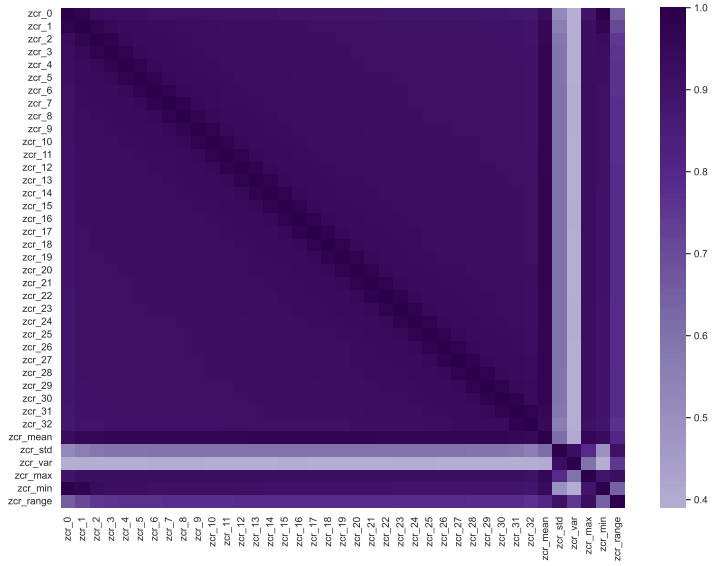
Conversely, per-second spectral centroid features (Fig. 2.4.1b) and per-second ZCRs (Fig. 2.4.1c) were highly correlated with each other and the sample mean. Therefore, I considered the mean a suitable representation of the full sample and decided to omit all per-second features. Among spectral centroid features, the standard deviation and range were least correlated with the mean, while for ZCR features, all other metrics were correlated with each other fairly strongly. To include the same derived features from each set, the mean, standard deviation, and range of per-second ZCRs and per-second spectral centroids were kept as features.



(A) Power Spectral Density features.



(B) Spectral centroids.



(C) ZCR features.

FIGURE 2.4.1: Feature correlation matrices for power spectral densities, ZCR features, and spectral centroid features. Correlation coefficients were calculated using Pearson correlation (`.corr` method in `pandas`) across the full dataset.

2.4.2 H1: Performance on unseen hives is lower than performance on unseen data from training hives.

Each model was tested in two ways: on a random held-out test set from the same hives on which the model was trained (WG test), and on new hives (OOG test). As expected, models performed more poorly OOG than WG (Fig. 2.4.2). The overall range of test AUCs is larger in OOG tests than in WG tests, though both distributions have similar interquartile ranges.

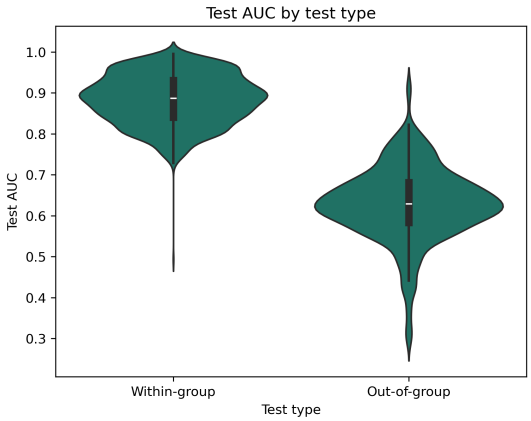


FIGURE 2.4.2: Test AUC by test type. Results from all sensor positions, feature sets, model types, and hives were combined.

2.4.3 H2: Sensor position does not affect model performance.

Models were trained on 4 different subsets of data defined by sensor position: 'all', 'Outside', 'Centre', and 'Inside'. The 'all' subset contained recordings from all sensor positions, including 'Outside 2' and 'Inside 2' which were not tested on their own. Contrary to my expectation, there were no consistent or marked differences in model performance between sensor positions (Fig. 2.4.3a). Mean test AUCs and interquartile ranges were similar, and all distributions were bimodal or trimodal.

When considering WG and OOG test performance separately (Fig. 2.4.3b), the sub-distributions at each sensor position also broadly overlapped. Means and interquartile ranges were similar between positions for WG test AUC, but varied slightly more for OOG test AUCs. The highest mean OOG performance was achieved by models trained on all sensor positions. While OOG test AUC distributions were skewed but

fairly normal, WG distributions in Fig. 2.4.3b were still bimodal, possibly due to the inclusion of both SHM and MHM model results (cf. Fig. 2.4.2; see also section 2.4.8).

2.4.4 H3: Including more hives in the training set improves performance on unseen hives.

I compared the performance of SHMs, trained on one hive, and MHMs, trained on nine hives. SHM test AUCs showed a strongly bimodal distribution. MHM test AUCs were also bimodally distributed, but the peaks were closer together and fell between the two modes of the SHM distribution. The mean test AUC was higher for MHMs (Fig. 2.4.4a.)

The bimodality of both distributions was largely resolved when splitting results by test type (Fig. 2.4.4b). While the WG test AUCs of SHMs were higher, the SHM OOG test AUCs were lower than for MHMs. In addition, the relative drop in performance from WG to OOG was lower. This implies that the added variance in the training data provided better potential for generalisation, aligning with the slight increase in mean test AUC when training on data from all sensor positions as opposed to a single position (Fig. 2.4.3b).

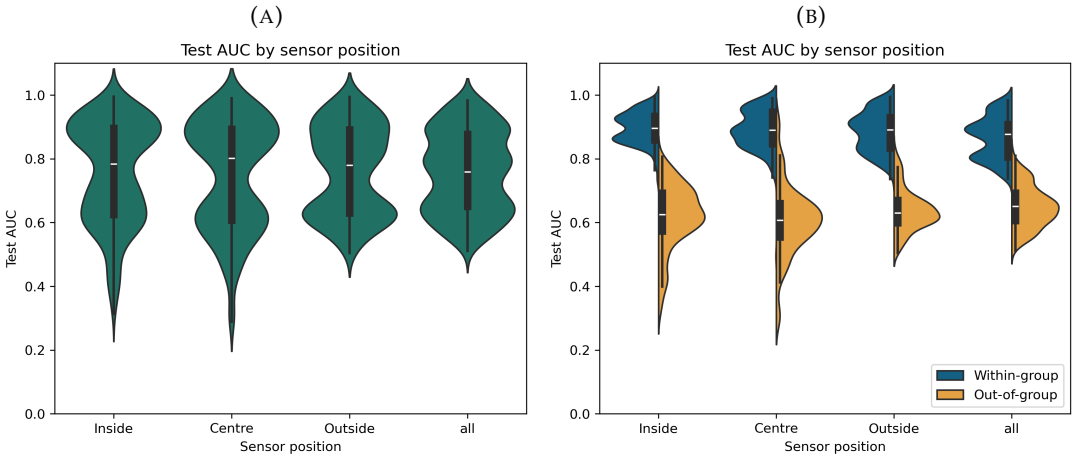


FIGURE 2.4.3: Test AUC by sensor position. Results from all model types, hives, feature sets, and test types together (A) and split by test type (B).

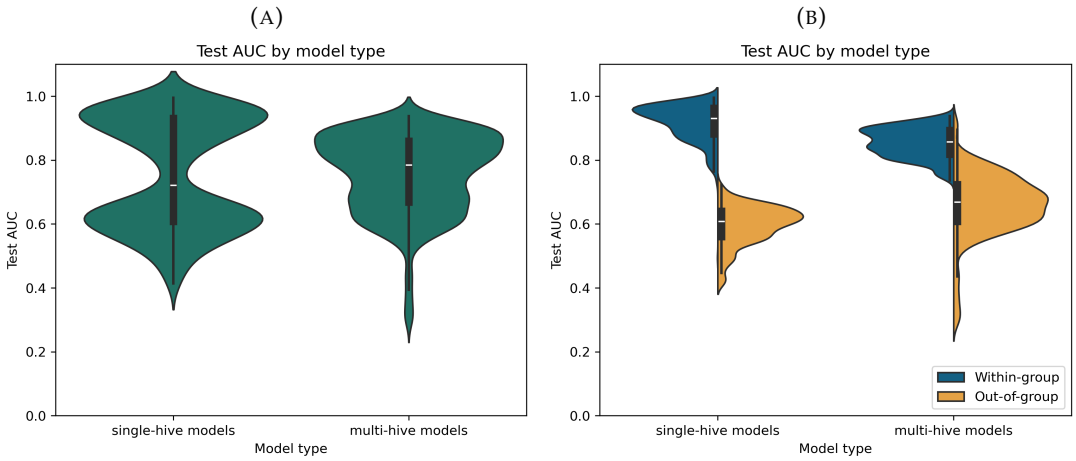


FIGURE 2.4.4: Test AUC by model type. Results from all sensor positions, feature sets, hives, and test types together (A) and split by test type (B).

2.4.5 H4: Handcrafted features outperform deep-learned features.

Models were trained with two different feature sets extracted from the same data: A handcrafted, interpretable feature set (see section 2.3.2) and ResNet-derived deep features (see section 2.3.2). I expected that the increased feature complexity in the deep feature set would be more advantageous for model learning. Contrary to expectation, the handcrafted features had a higher mean test AUC and a higher density of models in the upper end of the distribution (Fig. 2.4.5a). Both distributions were bimodal.

When splitting results by test type (Fig. 2.4.5b), both the mean WG and OOG test AUC were higher for handcrafted features than for deep features. The bimodality of the distributions in Fig. 2.4.5a was well resolved by the split into WG and OOG tests.

2.4.6 H5.a: Model performance varies between hives.

Even when controlling for hive configuration and population size, honeybee hives vary in temperament, resistance to stress, and behaviour (Wray, Mattila, and Seeley, 2011). Colony state detection models must be able to handle this variation between individuals. In my tests, single-hive model performance differed very little across hives in WG tests (Fig. 2.4.6a), while multi-hive model performance varied more, especially OOG (Fig. 2.4.6b).

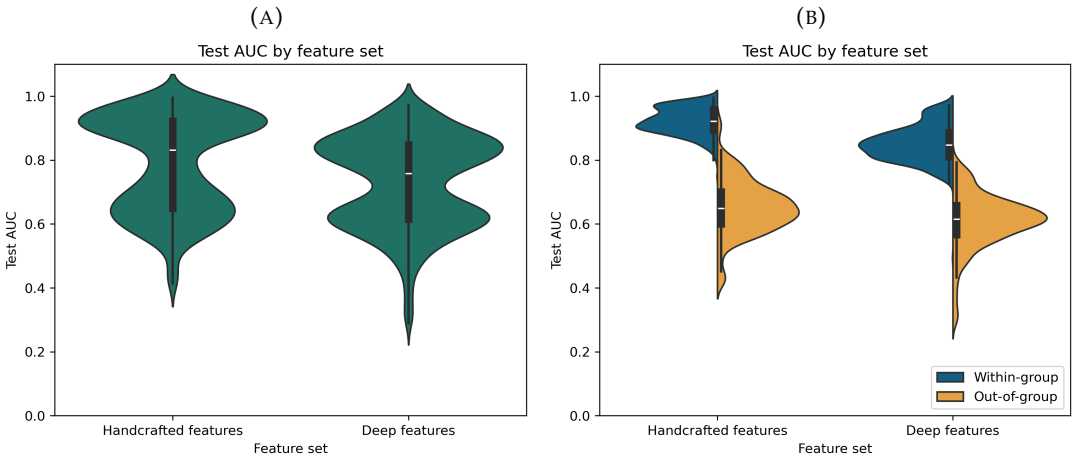


FIGURE 2.4.5: Test AUC by feature set. Results from all sensor positions, model types, treatment hives, and test types together (A) and split by test type (B).

SHM WG test AUCs were similar across all training hives[4] (Fig. 2.4.6a). Mean AUCs varied between 85–95% and interquartile ranges were often narrow and overlapping. For multi-hive models, individual hive distributions shown in Fig. 2.4.6a were almost identical since all models' pools of training hives strongly overlapped. For single-hive models, distributions differed more strongly in AUC range. Some hives, eg D4, had WG test AUCs consistently above 90%, whereas for others, eg C2, AUCs ranged down to 70%. The bimodality of some SHM subdistributions might be explained by performance differences depending on feature set or another variable not included in Fig. 2.4.6 (see also section 2.4.8).

OOG test AUC distributions were quite disparate when split by test hive[5], with means varying between 50–75% and shapes ranging from long-tailed normal to bimodal (Fig. 2.4.6b). While among WG tests, single-hive model mean AUCs were consistently higher than MHM means, the inverse was true for OOG tests. Multi-hive models showed higher performance on unseen hives compared to SHMs for every hive, with increases between 5–15%. While OOG performance could vary strongly between individual models as well as hives, models with more varied training sets

[4]Training hives were identical to test hives for SHMs. For MHMs, where training sets included multiple hives, results were averaged so that each hive distribution included results from all models whose training sets contained that hive.

[5]For MHMs, OOG test sets included only single hives. In contrast, SHMs were tested on all unseen hives not used during training, giving 7 OOG test results per model. To avoid overrepresenting SHM results in graphs, the OOG test AUC was averaged for each test hive (averaging across training hives while grouping by test hive, learning rate, sensor position, feature set, model type, and repeat).

consistently generalised better.

2.4.7 H5.b: Test hive type affects SHM performance.

Test performance of single-hive models decreased with increasing dissimilarity between training and test hives.

SHMs showed high mean AUCs when tested WG, but performance dropped when tested on other treatment hives[6] (mean AUC >81% on diagonal vs. 40 % <AUC <75% off diagonal in Fig. 2.4.7). The drop in AUC occurred to a similar degree with both feature sets (cf. Fig. 2.4.7a and Fig. 2.4.7b).

Performance degraded even more when SHMs were tested on hives not sharing the same pattern of state changes as the training hive (cf. left and right blocks in Fig. 2.4.7). On control hives, mean AUC/accuracy decreased to 15–35% for handcrafted features and 17–46% for deep features. While SHMs performed better than random on hives with the same state pattern as the training hive, performance dropped dramatically when tested on hives with a different pattern.

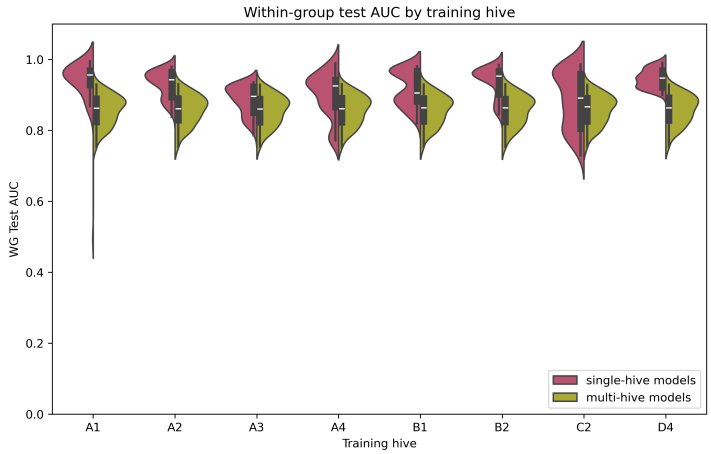
2.4.8 H5.c: The effect of sensor position on MHM performance is not consistent between hives or feature sets.

Fig. 2.4.6b showed strong remaining bimodality in most hives' MHM AUC distributions. This bimodality was resolved when splitting results by feature set and sensor position (Figs. 2.4.8, A.2.1). The variance was often low within these subsets, indicating a high degree of retrainability independent of the learning rate when feature set and sensor position were held stable.

Mean OOG test AUCs of models from the same hive, differing only by sensor position, varied by up to 30 percentage points (cf. positions 'Centre' and 'all' for hive A4 with handcrafted features in Fig. 2.4.8 or the same positions for hive B2 with deep features in Fig. A.2.1).

[6]Test performance on control hives was measured as accuracy (proportion of correct predictions over all samples) because AUC cannot be calculated when the test set contains only a single class. This was the case for control hives, which were queenright throughout the experiment. For treatment hives, where both AUC and accuracy were calculated, the two metrics usually differed by 5–10%. Accounting for a potential 10% underestimation of model performance on control hives, the drop in performance compared to treatment hives was still significant.

(A)



(B)

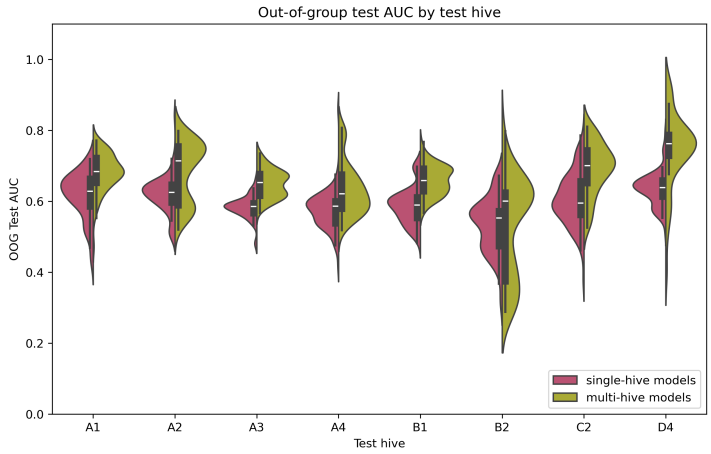
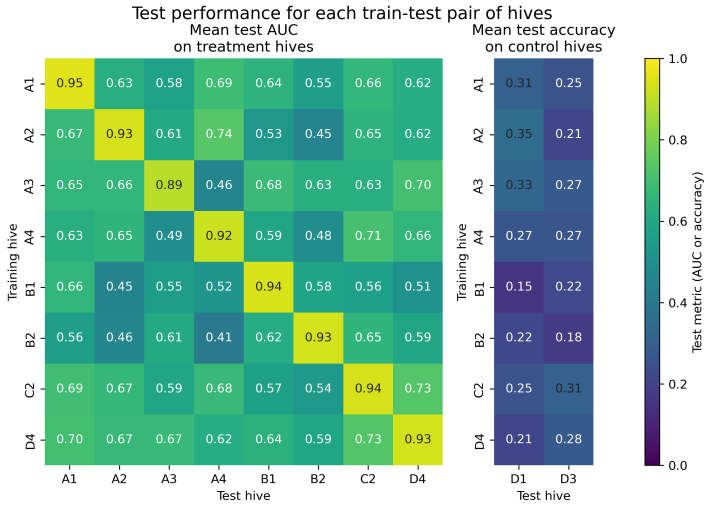


FIGURE 2.4.6: Test AUC by hive. Results from all sensor positions, feature sets, model types, and test types were combined. For subfigure (A), results were averaged by training hive among MHMs with the same learning rate, feature set, model type, sensor position, and repeat. For subfigure (B), results were averaged by test hive among SHMs with the same hyperparameters as for (A).

Additionally, the effect of sensor position was not consistent between feature sets. Mean AUCs of models from the same sensor position also differed by up to 30 percentage points (cf. position 'Outside' between handcrafted and deep features for hive A2 in Fig. 2.4.8 or position 'Inside' for hive B2 in Fig. A.2.1).

Finally, whether or not sensor position affected model performance, as well as which position was best, varied between hives. For example, while the 'Centre' position appeared most advantageous for hive A4, it was the worst position for hive A2 (Fig. 2.4.8).

(A) Handcrafted features



(B) Deep features

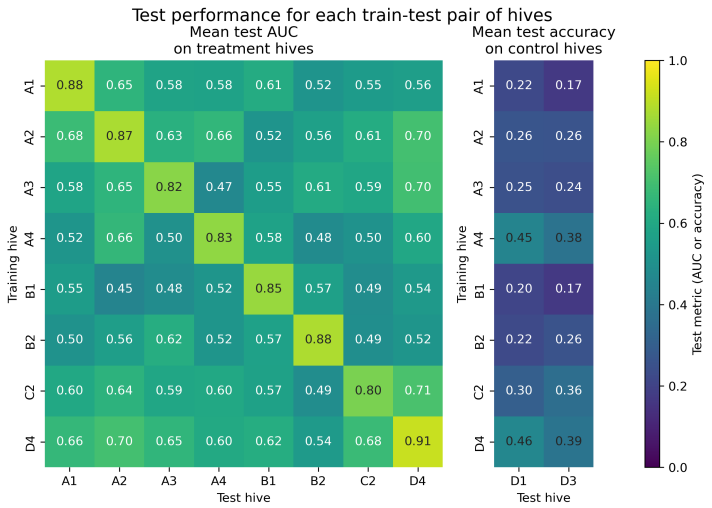


FIGURE 2.4.7: Cross-test accuracy of single-hive models on treatment and control hives using handcrafted (A) and deep features (B).

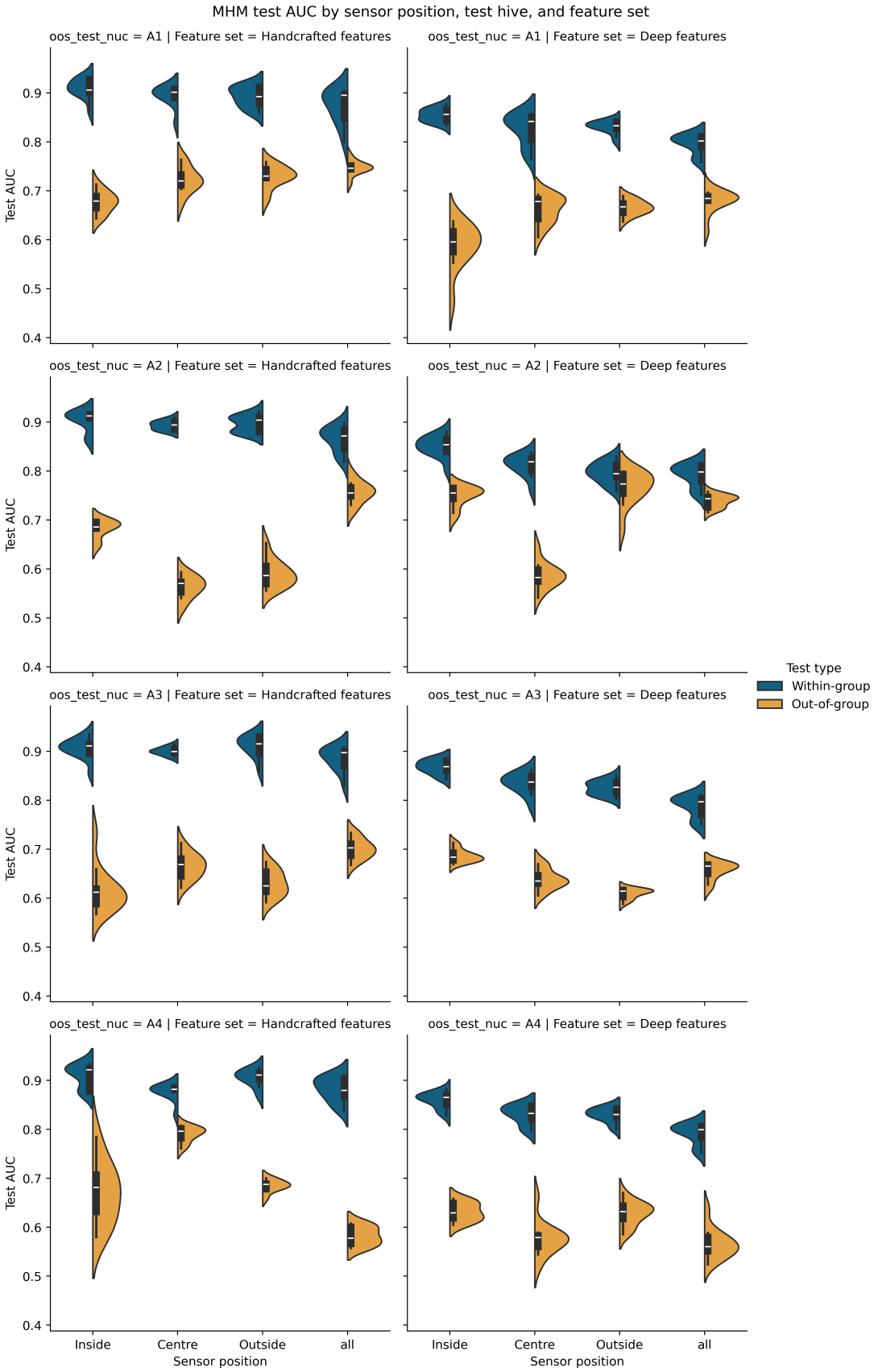


FIGURE 2.4.8: OOG AUC of MHMs by feature set (columns), hive (rows), and sensor position (x-axis). Model performance differed strongly by sensor position within each hive-feature-set subset, but differences were not consistent between subsets.

2.5 Discussion

2.5.1 Summary of results

H1 Irrespective of training parameters, models performed better when tested on unseen data from training hives than on fully unseen test hives.

H2 Sensor position can affect a model's ability to generalise, but the impact was highly inconsistent between hives and feature sets.

H3 Including more hives in the training pool reduced within-group performance, but improved out-of-group generalisation.

H4 Handcrafted features outperformed deep features both in within-group and out-of-group tests.

H5.a Both within-group and out-of-group performance varied between hives.

H5.b Single-hive model out-of-group performance was significantly worse when the test hive was dissimilar from the training hive.

H5.c Feature set, sensor position, and hive identity interacted unpredictably to affect model performance.

2.5.2 Sensor position affects model performance inconsistently

Honeybees perform a wide range of vibration-linked behaviours, some of which are localised to particular regions of the hive (Winston, 1987). For example, ventilation fanning, foraging flights, and waggle dancing for forager recruitment mainly take place at or near the hive entrance (Lindauer, 1952; Tautz, 1996; Tautz and Lindauer, 1997). The brood nest is located in the centre of the hive. This is where brood rearing and queen care take place, as well as wax-working to build comb and cap or uncap brood cells (Tautz, 2009). Thermoregulation behaviours, such as fanning and clustering, are also performed in the brood nest to maintain stable conditions for the developing brood (Kronenberg and Heller, 1982). The brood nest is surrounded by food storage, and the frames furthest from the entrance are typically dominated by honey (Tautz, 2009). Localised behaviours include food storing and fanning to refine honey (Camazine, 1991).

Most of the behaviours listed above either have a documented vibrational or acoustic component (see also section 1.3.2). or are likely to have one (eg noise of bees rubbing together in a cluster). I expected these sounds to attenuate with increasing distance, giving each hive region a distinct soundscape. Sensors placed closest to where a relevant behaviour is performed would then have the highest signal-to-noise ratio (SNR) for that behaviour.

The queen spends most of her time in the brood nest. Her loss causes a decrease in brood rearing and triggers rearing of a replacement queen. Therefore, I expected the quickest behaviour shifts upon queen loss to occur in the middle of the hive, and to find the highest SNR and consequently best model performance at the 'Centre' IHS.

However, my trained models showed the same performance no matter which set of positions was used. If soundscapes differed between hive regions, then models may have been able to exploit different but equivalent sources of acoustic information. Equally, models could have relied on sounds linked to behaviours that are not localised to a hive region, for example regulatory signals like tremble dancing (Seeley, 1992; Nieh, 1993)[7]. A third possibility is that the relevant sounds do not attenuate over the distances in a hive, creating a uniform in-hive soundscape where SNR does not differ strongly between regions. In this case, models trained at one position would be freely transferrable to another.

When splitting results by feature set and hive, sensor position often (but not always) had a significant effect on model performance (Figs. 2.4.8, A.2.1). This suggests that in these hives, the SNR varied by hive region. In other hives, there was no effect of sensor position, indicating that in these hives the SNR was uniform throughout the hive. Whether models for these hives exploited different sources of information in each region or whether they relied on the same sounds is unclear and should be explored in the future.

There could be differences in model functioning between positions. To test whether models from different positions used different sound information, models could be tested on data from a different sensor position. In addition, comparing the most commonly used features between positions, eg using SHAP values, could show whether

[7]The tremble dance appears to be performed throughout the brood nest as well as the dance floor area close to the hive entrance.

models exploit similar cues.

The inconsistency in optimal sensor position may be due to variation in how hives are organised. Typically, the brood nest is located in the centre of the hive (Camazine, 1991), but its location and volume can vary depending on the amount of brood and food stores, shifting the centre of the nest and the cluster of bees backward or forward. Further work could record the spatial organisation of the nest at each inspection and attempt to relate the variation in best sensor position with these characteristics.

Models trained on data from all sensor positions together did not perform significantly better than models at single sensor positions, despite having been trained on more samples as well as a greater variety of samples. This suggests that the size of the training set in itself was not a major contributor to model performance.

My results underline the complexity of developing generalisable colony state prediction models. The optimal sensor position varied by hive even within a closely controlled experiment, where all hives shared the same location, configuration and beekeeping treatments. In light of these results, varying configurations or climates seem almost insurmountable hurdles to model transferability.

As it stands, my recommendation for gathering sensor data from within hives is to place sensors anywhere within the hive, but to keep sensor position consistent between hives and over time. Future work should test the impact of varying sensor position within the experiment dataset and establish whether existing models can be applied to new datasets with a different sensor position. The benefits of conjugating simultaneous recordings from different positions within the hive should also be investigated.

2.5.3 Frequency features outperform deep features

I tested the importance of the feature set for the development of generalisable colony state prediction models. I contrasted a handcrafted feature set containing frequency and waveform features against a deep feature set of spectrogram encodings. I expected the deep features to extract more detailed information from the raw data (including about the time dimension of the sample), leading to better model performance.

Handcrafted features are generally less complex than deep features, since they are conceived through human intuition. While these features do not allow models to learn from minute details in the data, they could have a higher potential to generalise. Simpler features are likely to apply more broadly, since small details in the data, which could lead to overfitting, are ignored. In data regimes with low SNR, deep features could be especially prone to overfitting (Ying, 2019), and handcrafting features could be an essential step to improving inference.

I demonstrated that deep features did not improve the ability of my models to generalise to unseen hives compared to handcrafted frequency and waveform features. This is unusual, as deep-learned features are often considered superior to handcrafted features. When evaluating the quality of liver images, handcrafted features outperformed deep features (ResNet-50 encodings) at small sample sizes, while deep features performed better on larger training sets (Lin et al., 2020). In a fingerprint classification task, deep features were reported to generalise better across sensors, while handcrafted features generalised better across datasets (Agarwal, Rattani, and Chowdary, 2021). However, similar to my own results, model performance varied strongly within each subtask depending on which sensor or dataset models were trained and tested on. For example, one handcrafted feature set produced models with a mean accuracy of only 56% when trained on a dataset from 2015 and tested on one from 2017 — when datasets were swapped, the same features produced models with 78% mean accuracy.

Both Lin et al. (2020) and Agarwal, Rattani, and Chowdary (2021) fine-tuned their pre-trained ResNets, whereas I opted not to. Fine-tuning the encoder by training it on labelled data is an optional step which can improve the encoder, but is also time- and compute-intensive. It is possible that fine-tuning would have produced deep features with higher discriminatory power and possibly better performance. However, the integration of fine-tuning with the factorial design of my experiment would likely have been time-prohibitive. To avoid test set contamination, the encoder can only be fine-tuned on the data used to train the predictive model later on. Consequently, I would have needed to train 64 separate encoders: one for each combination of 4 sensor positions and 8 single-hive and 8 multi-hive sets. However, the impact of fine-tuning could be explored on a smaller set of models and variable combinations.

It is also possible that the ResNet architecture I used was too small to encode the spectrogram effectively. Both Lin et al. (2020) and Agarwal, Rattani, and Chowdary (2021) used ResNet-50, while I used ResNet-18, the smallest variant. A larger architecture may have been more effective at extracting useful features from the spectrograms. Similarly, a MLP might have been better at handling the larger number of deep features.

A further possible reason for why handcrafted features performed better could be the difference in the effective size of the data set. The same amount of data was used to construct both feature sets, but to obtain deep features, two concurrent recordings from the same sensor were layered to create a multi-channel spectrogram. This effectively halved the number of samples and theoretically could have led models to learn more slowly since, overall, they saw fewer different examples of each target state. However, the amount of information going into the models was the same — the information was simply aggregated and transformed in a step prior to being fed into the neural network. Furthermore, I used an early-stop criterion to monitor convergence, instead of training all models for a fixed number of epochs. This measure should have avoided overfitting and poor test performance. Also, given the large number of recordings available (ca. 180,000 in total), halving the effective size of the dataset should have left enough samples to allow for successful training.

While I had more success with handcrafted features, deep features can still be viable for colony state prediction. In addition to the factors explored above, other changes, such as increasing layer size or using a multi-layer perceptron, could improve inference with deep features beyond what I achieved. On other tasks, using handcrafted and deep features in combination has been shown to outperform both feature sets in isolation (Lin et al., 2020). This remains a promising strategy that is so far unexplored in hive monitoring.

2.5.4 Cross-hive generalisation

Every model was tested on data from training hives (WG), as well as at least one unseen hive (OOG). My results demonstrate stark differences in model performance between hives. Model performance is very stable *within* a hive, but significantly differs *between* hives when accounting for feature set and sensor position. The features that

allow one model to distinguish between samples from each queen state do not seem to work for another hive (cf. Fig. 2.4.7), or even a different sensor position within the same hive (Fig. 2.4.8).

The gulf between WG and OOG test AUC in SHMs indicates that individual hives' soundscapes differ enough to affect model performance. While MHMs are better at making predictions on unseen hives, their success also varies by hive despite the models' increased exposure to between-hive variation. This suggests that even a comparatively large training set of 9 hives may not be sufficiently varied to support full generalisation to unseen hives. In addition, the improvement is moderate at 5–15 percentage points depending on the test hive (Fig. 2.4.6b). Further work could test models with intermediate numbers of hives in the training set and characterise the impact of training set diversity on model performance in more detail.

The failure of SHMs to make accurate predictions on control hives (Fig. 2.4.7) could indicate that these models are utilising secondary information to make predictions. Environmental cues that are audible from inside all hives (eg shifts in temperature leading to more or less fanning, rain, etc.) could contain transferable information about the time within the experiment. If models learned to exploit these cues, this could explain the observed drop in prediction accuracy in control hives. While exploiting time information would yield reasonable predictions for other treatment hives, where the schedule of queen removal was identical to the training hive, this strategy would not translate to hives with a different queen state pattern. It is likely that these SHMs would perform even more poorly on data from hives in other locations or at different times of year.

The added variation in the training set of multi-hive models simultaneously appeared to increase OOG performance, but to reduce WG performance. The multi-hive training set may have provided less scope for exploiting hive-specific indicators of queen presence or time-within-experiment which would not translate to other hives. The inclusion of control hives in the training pool for MHMs is another likely cause for their improved performance on unseen hives, since their presence in the validation set precluded multi-hive models from relying on external cues to make predictions.

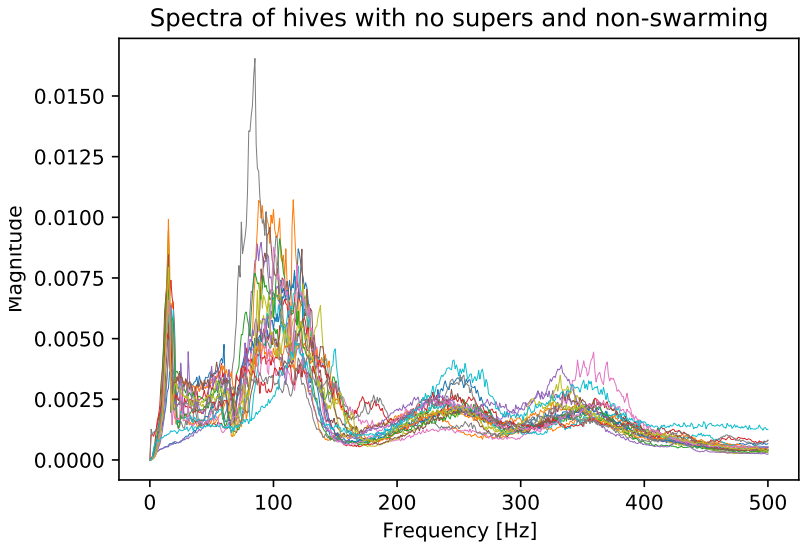
A major contributor to sound differences between hives is the physical configuration of the hive. The presence or absence of additional brood boxes or honey supers

fundamentally changes the frequency spectrum within the hive, since sounds resonate differently depending on the shape and size of the in-hive volume (Fig. 4.3.2). In this experiment, all hives had the same configuration, so between-hive differences could be linked to variation in colony behaviour and/or spatial organisation (see discussion in section 2.5.2).

Physical properties of frames can also affect how sounds resonate and propagate within the hive. Vibrations were attenuated on comb attached to a frame, but amplified on smaller, unattached pieces of comb (Sandeman, Tautz, and Lindauer, 1996). Forager recruitment was three times higher to waggle dances on empty, uncapped cells compared to dances on capped brood comb (Tautz, 1996), likely due to the amplifying properties of empty cells (Tautz, Casas, and Sandeman, 2001). Honeybees themselves have also been shown to passively regulate background noise on the comb, their bodies damping vibrations when standing across multiple cells (Smith and Chen, 2017). To my knowledge, the effect of honey, pollen, and developing brood on comb resonance and vibration transmission has not yet been investigated. However, it stands to reason that comb filled with a dense, viscous mass like honey will resonate differently than comb containing less dense material like brood, or containing nothing at all. In fact, a honeycomb soundproofing board was found to be over 50% more effective when filled with straw or hollow beads than when empty (Song et al., 2024). It appears likely then that the amount of brood, pollen, and honey in a hive would significantly affect how quickly vibrations and sounds are attenuated.

In Cyprian honeybees (*A. mellifera cypria*), "acoustic heterogeneity among colonies" was observed in the duration, dominant frequency, and fundamental frequency of anti-predator hissing (Papachristoforou et al., 2008). However, within-colony variation was greater than between-colony variation for each tested characteristic, leading the authors to conclude that there is no colony acoustic signature in the anti-predator hiss. My results do not conclusively show a colony signature either — this would have required directly comparing sound recordings and computing within- and between-colony variation for different metrics, which was outside of the scope of this study. However, I observed a stark lack of consistency between hives in terms of suitability of different features and sensor positions. This strongly suggests fundamental variations in the soundscapes of different hives. Likely, a longer experiment would also

(A)



(B)

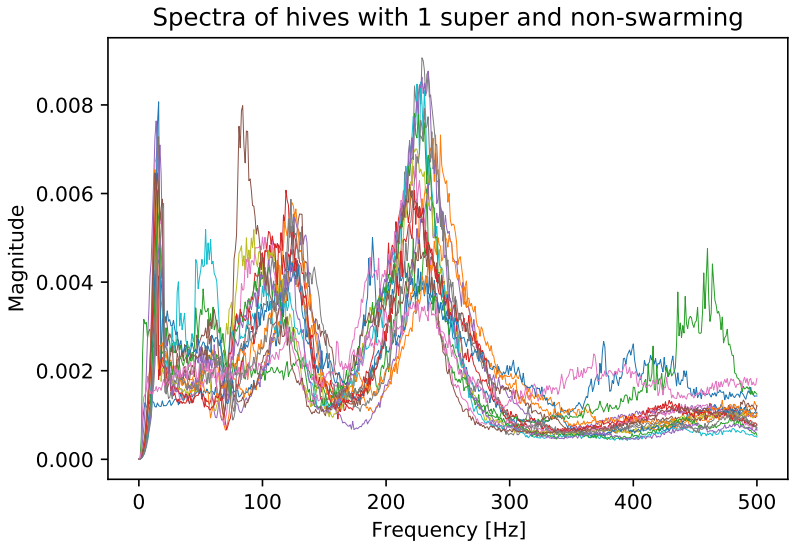


FIGURE 2.5.1: Magnitude spectra of non-swarming hives with 0 or 1 honey supers. Changes to the hive configuration fundamentally alter the frequency spectrum. Unpublished data collected by Huw & Sandra Evans (Canetis SRL) during swarming season of 2010–2011.

demonstrate a lack of consistency *within* hives, where a trained SHM's performance would decrease when tested on later recordings from the same hive.

All of these factors strongly suggest that hive identity is a major variable to consider when developing and testing any algorithm or machine learning model for detecting colony state. Currently, models are often tested on random subsets of recordings, often from a small number of training hives (eg Sharif et al. (2020), Liang (2024),

and Robles-Guerrero et al. (2024)). Other models were tested on unseen hives, but the test sets were limited in the number of hives, timespan, or label diversity (Nolasco et al., 2019; Orlowska et al., 2021; Terenzi et al., 2022; Ruvinga et al., 2023). Though many of these models tested to very high accuracies, the results are highly unlikely to reflect the performance of the model on new hives. The technological advances presented in these studies therefore should be robustly validated before further research builds on them. Hive individuality also presents issues when trying to draw biological conclusions from trained models using feature importance or other model interpretation techniques, as important features for one hive may not translate to others.

Future work on hive monitoring should prioritise robust testing on unseen hives. Ideally, both the training and test set should contain more than one hive if the dataset allows, and the test set should include all classes.

2.5.5 Study limitations

While my dataset was large compared to many other studies on colony state monitoring, the number of hives was still small with 10 individuals. Only two hives were controls, both of which were used for training, which precluded me from testing multi-hive models on hives with different queen state patterns.

It is also possible that different model type or ANN architecture would have led to better performance overall. During model development, I tested the following architectures with handcrafted features: logistic regression, gradient boosted trees (`XGBoost`), and SLPs with 32, 128, and 256 nodes in the hidden layer. With ResNet features I tested the following architectures: SLP with 128 nodes, MLPs with up to 512x256x128x32 nodes, and a multi-instance learning (MIL) model with 128x64x32 nodes. None of these models performed particularly well. However, I found that 256 nodes were sufficient for model fitting on the handcrafted feature set. Rather than continue optimising model architecture, and optimising for each feature set separately, I decided to use a simple ANN with a single hidden layer of 256 nodes for both feature sets to test the factors of interest. Future work could evaluate the impact of network size on model performance.

My data were collected in the same location by the same beekeepers over the same time period. All colonies belonged to the same honeybee race and hives were set up

identically. Controlling for these factors is essential to make colonies as comparable as possible, but it also limits the variation in colony sounds that models can experience during training. For example, transferring a frequency-based model from one dataset to another is only possible if the target hive shares the same configuration as the training hives (cf. section 2.5.4). As the literature on colony monitoring expands and new datasets are made available, the potential for the development of models with greater generalisation power will grow.

In accordance with most of the hive monitoring literature, I treated queen state detection as a closed-set problem with only two distinct queen states: queenright and queen-rearing. In reality, honeybee colonies can experience a wide range of queen states, including: having a virgin queen, being "hopelessly" queenless (without the option to rear a replacement from young brood), having laying workers, or having a mated but unproductive queen. Only some of these queen states require beekeeper intervention — for example, a colony with a virgin queen can be safely left alone in most cases, while an unproductive mated queen should be replaced with a viable queen as soon as possible, and a "hopeless" hive may need to be combined with another strong hive (Thompson, 2016). Despite the full experiment including more than two queen states, I chose to limit my analysis to times where colonies were either queenright or queen-rearing. While this simplifies inference, it further limits the applicability of my results to real-world deployments.

2.5.6 Towards general queen state prediction

For queen state prediction, false positives are more dangerous and damaging than false negatives. It is most important to detect the queen's absence as soon as possible after queen loss. The impact of a false negative is a wasted hive inspection, which causes disruption to the colony — in contrast, a false positive wrongly reassures the beekeeper that the queen is present and well, when in fact the queen has been lost and intervention is necessary. Delays in replacing the queen can result in gaps in the brood cycle, interrupting colony growth and reducing the foraging population down the line. This could impact pollination output, honey production, or overwintering success, depending on the time of year. To avoid false positives and the associated

loss of productivity, a viable model needs to be very good at detecting queen loss, while errors in detecting queen presence can be tolerated.

I have shown that even multi-hive models trained on 9 hives struggle to generalise to new hives. The application of such models is therefore still very limited. Since it is not possible to collect data from all hive states of interest and train a bespoke model for every colony, fine-tuning could offer a solution. In this strategy, a model pre-trained on several hives would be finetuned on a small set of labelled samples from a target hive in its current, healthy state. The model should then be better equipped to detect changes in colony state in this specific hive. However, the effort in customising models to each hive might be prohibitive for larger beekeeping operations. Changes in soundscape over time, linked to seasonal changes in hive activity, would also require periodically updating the hive-specific training set of healthy samples.

An alternative is a stable 'label change detector' as proposed by Bricout et al. (2024). In this strategy, audio samples are compared to a recent reference sample with known label, for example a recording taken just before a hive inspection. A model trained by contrastive learning, which focuses on detecting whether two samples share the same label or not, then detects whether the current sample is similar or not to the reference sample. If the current sample is predicted to be dissimilar, this should trigger an inspection to diagnose the issue. At the next regular hive inspection, the reference sample is updated. This technique has shown a high ability to generalise to new hives without fine-tuning. Since the model itself is never retrained, the added effort would be lower than for the fine-tuning strategy.

2.5.7 Conclusion

In this chapter, I showed that the performance of audio-based queen state prediction models is strongly dependent on the number of hives present in the training set, the sensor position, and the feature set used. In out-of-group tests, the identity of the test hive can further affect model performance, indicating substantial variability in in-hive soundscape between hives.

I demonstrated that deep features, like those derived from CNNs or other image-based models used in many recent studies, are not necessarily better than simpler handcrafted features — in fact, they were consistently inferior to handcrafted features

in my tests. I also emphasised the importance of unbiased model testing. If models are tested on random subsets of recordings from the same hives used to train the model, performance is strongly inflated.

Despite the large and growing number of studies reporting the success of their models, general monitoring of queen state is still out of reach.

Chapter 3

Explainable AI for detecting honeybee queen loss

Contributors:

Stella Marie Felsinger (study design, data analysis, interpretation of results, writing),

Huw Evans (experiment design, data collection),

Sandra Evans (experiment design, data collection),

Beth Mortimer (supervision, interpretation of results),

Geraldine Wright (supervision, interpretation of results).

3.1 Abstract

Predicting honeybee colony condition from acoustic data is an active research area. Despite advances in this field, the underlying soundscape of honeybee hives and how external and internal factors affect it has been studied rarely. The in-hive soundscape is an emergent property of behaviours performed throughout the colony and therefore sensitive to changes in behaviour, as well as being affected by ambient noise.

Here, I show how in-hive soundscapes are affected by a range of factors including colony strength, hive identity, honey stores, and time of day. Queen state substantially impacted soundscape, but the region linked to the queenright (healthy) state differed from colony to colony. I tested different post-hoc methods of processing prediction time series to improve accuracy. The efficacy of these methods also varied strongly between hives. By calculating feature importance values over a selection of best models, I identified features with consistent contributions to predicting queen state across

different hives and links to known honeybee communication signals.

These results demonstrate that hive identity strongly affects model performance due to interacting impacts of a multitude of colony states on the soundscape. Feature importance analysis was nevertheless able to identify common features used across hives to predict queen state. These findings could inform the development of future hive monitoring models with a simplified, biology-informed feature set.

3.2 Introduction

Large parts of agricultural production, notably the Californian almond industry, are nowadays partially or entirely reliant on migrant beekeepers who provide pollination as a service (Goodrich, 2019). This practice can provide a significant additional revenue stream to beekeepers, but also exposes bees to new stressors such as transport (Ahn et al., 2012), pesticide exposure (Krupke et al., 2012), nutritional monotony (Di Pasquale et al., 2013), and facilitated disease and parasite spread between colonies (Dynes et al., 2019). Deleterious interactions between these stressors are well-documented (eg Dolezal et al. (2019) and Pettis et al. (2012)). The high rates of colony mortality experienced by beekeeping operations (NASS, 2024a) are increasingly being understood as a rational consequence of commercial pollination (Ellis et al., 2020; Shanahan, 2022), which itself is a stop-gap measure to maintain the viability of industrial agriculture at unsustainable scales. Simple, scalable, and rapid monitoring and maintenance methods are required for the welfare of honeybee colonies within this system.

As the only reproductive individual in the colony, the queen is of singular importance. Laying hundreds of eggs each day, she counteracts worker mortality to maintain or grow the population (Winston, 1987). A queenless colony, in turn, will quickly decline in strength and eventually die if the queen cannot be replaced by the beekeeper or by the workers themselves through emergency queen rearing. Monitoring queen health is therefore both an ethical and an economic necessity.

Regular hive inspections are central to maintaining hive health. However, inspections also introduce disruption to the colony, causing stress and usually killing some workers during dis- and reassembly of the hive. Inspections can also kill the queen, if she drops outside of the hive or is crushed when replacing a frame. Worker bees can

even react to colony disturbance by killing their queen (Gonçalves, 2021). Limiting the number of inspections, usually to no more than fortnightly, is crucial. Reducing the frequency of inspections even more could further benefit colonies by maintaining a hive environment free from disturbance. Therefore, less invasive methods for detecting queen presence are desirable.

Honeybee colonies are said to produce a characteristic "roaring" sound when queenless. This has been attributed to increased Nasonov scenting, whereby workers expose certain glands on their abdomen and fan their wings to percolate signalling pheromones (Butler, 1954; Fell and Morse, 1984). While the "roaring" sound has not been clearly acoustically defined, most beekeepers claim to be able to recognize a queenless colony by ear. This makes detecting the queenless state a popular target for audio-based hive monitoring studies. While many studies claim to have developed highly accurate queen detection models, they are usually insufficiently tested (as discussed in Chapter 2) and do not interrogate model functioning. Other work in honeybee acoustics is purely phenomenological, describing changes in the in-hive soundscape of individual hives (such as Kanelis et al. (2023)).These studies propose to advance our technological capabilities or provide insights into the behavioural shifts surrounding queen loss. Unfortunately, many of them use datasets with a limited number of hives, meaning that they do not offer any broader understanding of how queen loss affects colonies. If we aim to develop generally applicable audio-based monitoring methods that support honeybee welfare as well as the beekeeping and pollination sector, it is essential to identify common patterns of queen state-linked sound changes that apply to most honeybee hives.

This chapter further presents and investigates model results from Chapter 2. Models were developed on a dataset of 10 hives experiencing queen loss and rearing replacement queens. Here, I explore the audio space occupied by each hive and show how different variables map onto this space. I use SHapley Additive exPlanations (SHAP) values to identify in-hive soundscape changes that occur in most or all hives, identifying candidates for generalisable features. I also discuss between-hive differences in model performance. I suggest and demonstrate strategies for improving the fidelity of model predictions, and discuss differences in their applicability between

hives.

3.3 Experiment methods

3.3.1 Data collection

I used the same dataset as for Chapter 2. The experimental setup and recording schedule are described in section 2.3.1.

3.3.2 Schedule of interventions

Colonies were inspected four times over the course of the experiment (Fig. 3.3.1). During each inspection, Liebefeld grading was carried out to assess the amounts of bees, brood, honey, and pollen in the hive to 1/8th frame precision. Hives were also briefly opened on 11. and 15. August to check colony condition. The queen of each hive (excepting control hives) was caged on 4. August to facilitate her removal on the following day with minimal disruption to the colony. On 5. August, queens were removed from all but two colonies (control hives D1 and D3). All colonies began rearing replacement queens, with closed queen cells being present in all treatment hives on 11. August. Queen cells were removed on 15. August from two colonies (A2 and A4). All other colonies were allowed to requeen themselves. Despite removing queen cells from this hive five days prior, a virgin queen was discovered in A2 on 20. August. On the same day, a remaining queen cell was found in hive A4 and removed.

The experiment continued to run until 10. October and included samples from colonies with laying workers, virgin queens, and newly mated queens. However, to facilitate analysis, I chose to limit the analysis to the pre-treatment and queen-rearing period (27. July – 17. August 2023). This subset of the audio data contained almost 180,000 samples, close to 1/3 of the total dataset.

3.3.3 Inferring queen status

Using the known dates of queen caging and removal, as well as observations of capped queen cells and virgin queens during inspections, I inferred the daily queen status of each hive (see Tab. B.1.1). Capped queen cells were found in all queenless hives on

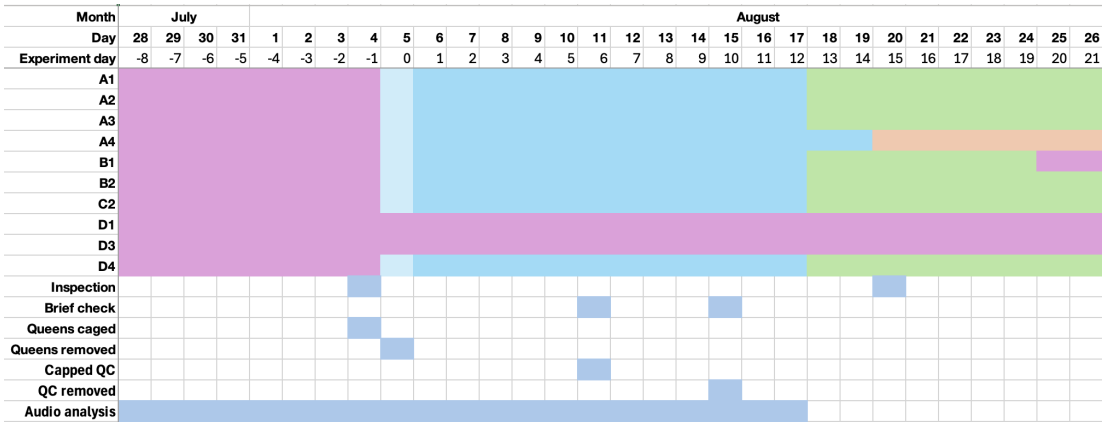


FIGURE 3.3.1: Experiment schedule with queen state by hive (given by A#, B#, C#, D#). Hive inspections using Liebefeld grading were carried out on day -1 (4. August) and day 15 (20. August). Mated queens were removed from 8 of 10 colonies on day 0 (5. August). Capped QC were found in all queenless colonies on day 6 (11. August). On day 10, queen cells were removed from hives A2 and A4, but at least one cell was missed in each hive and colonies continued to rear queens. Pink: mated queen present, light blue: day of queen removal, bright blue: queenrearing, orange: queenless, dark blue: inspections and key events. Data collection continued until 09.09 (inspections) and 30.09. (audio recording), but audio analysis was restricted to the period from day -8 to day 12 to exclude queen states other than queenright and queenrearing from the dataset.

11. August, suggesting that queen rearing began very soon after queen removal. If all colonies began rearing new queens immediately on 5. August from 3-day-old eggs, the earliest that these queens could have emerged is 18. August (day 12). I therefore considered hives to have been rearing queens from 5. – 17. August.

3.3.4 Inspection data

Inspection data including the amounts of honey, pollen, brood, and bees were interpolated to daily frequency with `pandas.Dataframe.interpolate()` using Piecewise Cubic Hermite Interpolating Polynomial (PCHIP). The PCHIP method was chosen because it showed the most intuitive and conservative behaviour out of all methods tested when interpolating between sparse observations. Interpolation was limited to the period between the first and last inspection, since interpolating outside of the bounded time period would be too error-prone. The interpolated time series were plotted to verify the integrity of the interpolation (Figs. 3.3.2, B.1.1a & B.1.1b).

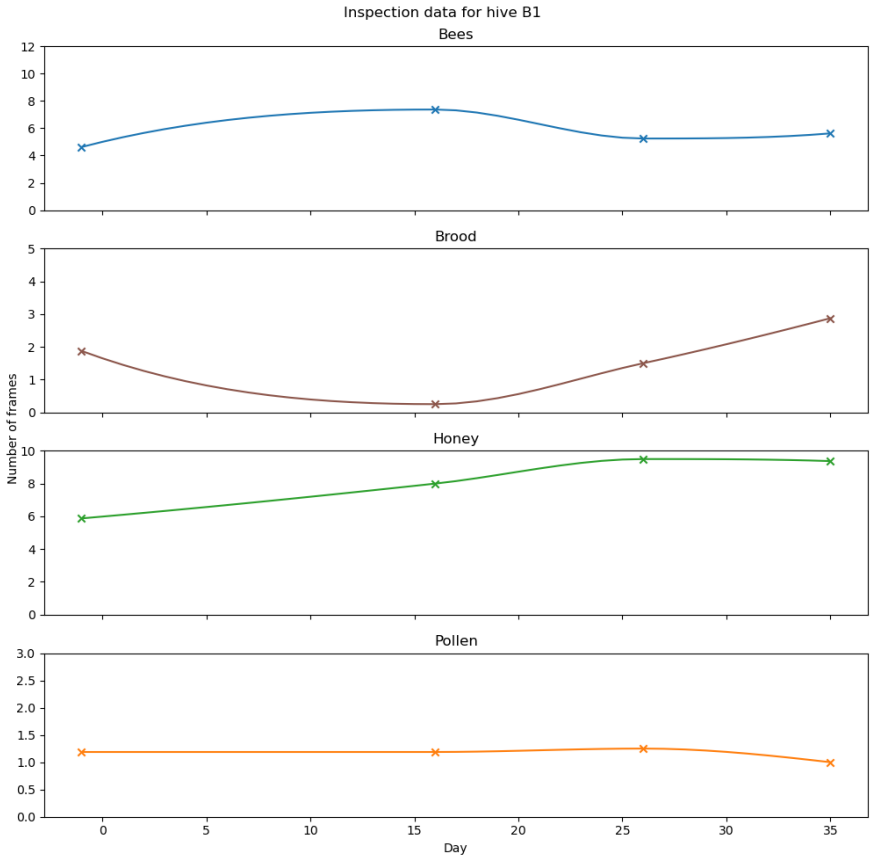


FIGURE 3.3.2: Example of interpolation from fortnightly inspections of census data from a single colony (Liebefeld grading of bees, brood, honey, and pollen) to daily frequency using PCHIP. Crosses indicate Liebefeld inspections and line indicates interpolated values. The graph shows the entire period over which hive census data were collected. Analysis of audio data included the period from day -8 to day 12 (28.07. – 17.08.2023).

This hive (B1) lost its queen on day 0 and began queenrearing. As a result, brood production declined gradually before recommencing when the new queen had mated. Colony strength initially increased but then decreased slightly due to a lack of brood. Honey stores increased and pollen stores remained stable during the period of low brood production, but declined slightly when brood production increased in the latter half of the experiment.

3.4 Analytical methods

3.4.1 Feature extraction

I extracted a handcrafted feature set using the methods described in section 2.3.2. The feature set included power spectral densities between 0–500Hz with ca. 4Hz spacing, as well as the mean, standard deviation, and range of ZCRs and spectral centroids. The resulting feature dataframe contained 179,298 rows and 135 columns.

3.4.2 UMAP embeddings

To visualise the structure and variation within in-hive sound during the experiment, I calculated the Uniform Manifold Approximation & Projection (UMAP) of all audio files to reduce the dimensionality of the data.

The features were first normalised using the `scipy StandardScaler()`.[1] A UMAP reducer (from `umap.UMAP()`) with 3 components and a neighbourhood size of 10 was then fit to the scaled feature data and the data were transformed accordingly to generate a 3-dimensional embedding. Metadata labels, including sensor position within the hive and queen status, were added.

The embedded points were plotted in 2D and 3D using the `seaborn-matplotlib` and `plotly` packages to visualise the distribution of samples in embedding space. I manipulated the boundaries of the embedding space to exclude single outliers in order to better visualise the vast majority of points. To achieve this, I excluded the lower and upper 0.07th percentile along each component axis. In total, 99.58% of points were kept. In order to avoid overplotting and increase visibility of all category levels, the selection of points was then randomly subsampled to include 5000 points from each hive. The final selection of samples for plotting included 50,000 samples evenly distributed across all hives.

[1] Whether or not features should be normalised for UMAP depends on the nature of the data. Where features stand in relationship to each other, it can be advisable *not* to normalise the features since the relationship will be lost (McInnes, 2018). In the case of this dataset, all features are derived from the same audio recording and are therefore related. However, the power spectral density, zero-crossing rate, and spectral centroid features are on very different scales. Therefore, I chose to normalise the data before applying the embedding algorithm.

By mapping marker colour to different variables, I visualised how the distribution of samples in embedding space varies with colony metrics such as queen status, population size, amount of brood, pollen and honey. Similarly, I visualised external metrics such as date, time of day, hive identity, the location of the hive, and sensor position within the hive.

3.4.3 Model training

Using the handcrafted feature set, I trained small ANNs (single-layer perceptron (SLP) with 256 neurons in the hidden layer) to classify the queen state of a hive based on an audio recording. I considered both the full dataset containing recordings from all 5 sensors in a hive, as well as the single-sensor datasets from the 'Outside', 'Centre', and 'Inside' positions (see Tab. 3.4.1 for the number of recordings in each subset).

I followed the methods in sections 2.3.2 to 2.3.5 to train single-hive and multi-hive models on the handcrafted feature set, resulting in a total of 960 SHMs and MHMs[2].

TABLE 3.4.1: Number of recordings in sensor subsets.

Subset	# Samples
All sensors	179,298
Inside	35,884
Centre	35,940
Outside	36,224

3.4.4 Feature importance

I used SHapley Additive exPlanations (SHAP) to estimate the importance of each feature in determining model predictions. This approach also allowed me to compare the most impactful features across models.

SHAP values are rooted in game theory and present a fair allocation of feature importance by averaging each feature's marginal contribution across all possible feature combinations. Exact computation of Shapley values requires exhaustively evaluating all possible subsets of features by including and excluding them, which is computationally impractical for large feature sets. Instead, I used the `shap.PermutationExplainer`

[2]The models are identical to those presented in Ch. 2 and were not retrained for this analysis.

(Python package `shap`, (Lundberg and Lee, 2017)), which approximates Shapley values. The `PermutationExplainer` shuffles feature values across samples to disrupt their relationship with the target variable, then systematically adds and removes features in a random order. This procedure captures the change in model output due to the addition and removal of each feature and produces an estimate of that feature's Shapley value for a single sample. By repeating this process across many samples, we can assess the overall importance of each feature. The SHAP values are contained in an `Explanation` object that can be further manipulated and plotted.

First, I calculated a baseline model output using a set of 1,000 random samples from queenright and queen-rearing hives. The baseline represents the model's expected output when it has no feature-specific information, acting as an approximation of the average output across the entire dataset. Then, I permuted a second set of 10,000 random samples by shuffling feature values and recalculated the model outputs for each permutation. Comparing the changes in output to the baseline allowed me to estimate the sensitivity of the model to changes in individual feature values within each sample, quantifying each feature's contribution to the prediction. The same set of 10,000 samples was used throughout to ensure that the results were comparable across models and hives.

To visualise feature importance, I used 'beeswarm' plots. In these plots, each feature is represented by a distribution ('swarm') of SHAP values across all samples (Fig. 3.4.1). Each point corresponds to a single feature and sample. The point's position on the x-axis represents the SHAP value (indicating the feature's impact on the output for that sample). The colour of the point reflects the original feature value relative to the baseline mean (blue = low, purple = average, red = high feature value). Points are stacked vertically to show the density of SHAP values along the x-axis, allowing me to see how frequently a feature has a given impact. If a feature's swarm had a tail of red points to the right (positive SHAP values), then high feature values increased the output value — conversely, a blue tail to the right meant that low feature values increased the output value. In this way, beeswarm plots were interpreted to understand each feature's impact magnitude as well as impact direction.

In my models, the '0' label was used to denote the queenright state, while the '1' label was used for the queenless or queen-rearing state. Therefore, a low model output

value means that the model considers the hive likely to have a mated queen, whereas a high output value means it is more likely to be queen-rearing.

To get a representative sample of feature importance across hives, I computed SHAP values for an ensemble of models which included the three best-performing MHMs for each test hive (by OOG test AUC).

For each model, I generated two types of beeswarm plot. First, all features were plotted in frequency order, allowing direct comparison of importance between features and between models. Secondly, features were ordered by their mean absolute SHAP value and the swarms for the 20 most important features were plotted. This ordering highlighted features with a consistent and significant influence on model predictions across the dataset, rather than those with occasional high impact.

Across the whole ensemble of models, I calculated the overall mean importance of each feature. Each `Explanation` object contains three arrays: `.values`, `.base_values`, and `.data`. After computing the `Explanation` for each model, I averaged across all models for each array type to obtain three mean arrays. These were formatted as a new, mean `Explanation` object.

Plotting the 20 most important features in the mean `Explanation` allowed me to

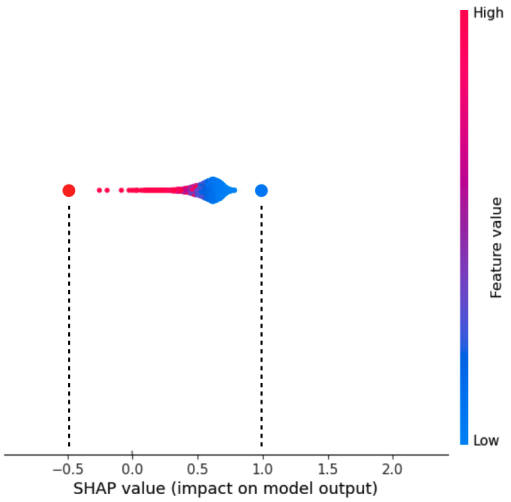


FIGURE 3.4.1: Illustration of 'beeswarm' visualisation of SHAP values. Each point represents one sample. Point colour indicates feature value and x-axis coordinate indicates the feature's contribution to the prediction value. Feature values and prediction values are represented in relation to a baseline mean for each feature and baseline model output. These were computed over a subset of 1,000 samples. By applying the SHAP algorithm to a large subset of 10,000 samples, we obtain a distribution of points ('swarm') showing overall feature impact.

identify features that were broadly used across most hives and models. Features with strong impact directionality (where high feature values consistently increase or decrease the output value, forming a pronounced "tail" in the distribution away from x=0) were considered especially noteworthy. These features may reflect behavioural or physical changes in the hive that occurred during and after queen loss. Therefore, these features were considered the most likely to generalise to new hives.

Below, I compare feature importance results for two representative models as well as mean SHAP values for the ensemble of best-performing models (Section 3.5.3).

3.5 Results

3.5.1 Audio manifold

I used UMAP (see section 3.4.1) to reduce the set of 135 handcrafted features to 3 components in order to visualise the audio feature space. This resulted in a 3-dimensional embedding or 'manifold'. In the view onto components C1 and C2 (projecting along the third component), the main cluster of the manifold has an oval shape with a short process or 'stem' at the low-C2 edge (Fig. 3.5.1a). In the C1/C3 view, the cluster approximated a triangle, with clearly defined high-C3 and high-C1 edges, but a more sparsely-populated low-C1 edge with a curved process (Fig. 3.5.1b). In the C2/C3 view, the distribution resembled a funnel shape with a clear high-C3 edge, but poorly-defined high-C2 and low-C3 edges (Fig. 3.5.1c). A small secondary cluster, along with several outliers, was located in the low-C1/high-C2/low-C3 coordinate space. The manifold had mostly clean boundaries, though some outliers existed.

Figs. 3.5.1–3.5.6 show views of the 3D audio manifold. Points are coloured by different population and experiment metrics, including colony size, time of day, hive identity, and queen presence. For each colour variable, three views of the manifold projecting along each of the component axes are given. Fig. 3.5.2 presents a C1-C2 view of the manifold of each hive individually, showing the localisation of different queen states for each hive.

Queen state

UMAP embeddings did not show evidence of a general signature of queen presence. Queenright samples (from colonies with a mated queen) were distributed throughout the manifold, with a slightly higher concentration in the centre, while queenless samples were more common at the edges of the manifold (Fig. 3.5.1a–3.5.1c). An outlying cluster appeared dominated by queenright samples though the surrounding space contained queenless samples as well.

Queen state by hive

When looking at the audio feature space of individual hives, there is some evidence of hive-specific queen presence signatures. Overall, most hives occupied a similar space within the manifold (Fig. 3.5.2). All hives overlapped in the central region, but varied in density in other areas. This was most notable in the bottom left corner, where hives A3, C2, D3, and D4 were very sparse, and the bottom 'stem' of the manifold, sparse in hives C2 and D3 and dominated in A1 and D1. The top right quadrant was sparsely populated for hives A1, A2, B2, and D2, but contained the bulk of the data for A3 and C2.

All hives (except control hives D1 (Fig. 3.5.2h) and D3 (Fig. 3.5.2i)) were made queenless at the same time. Nevertheless, the samples from the queenright period fell

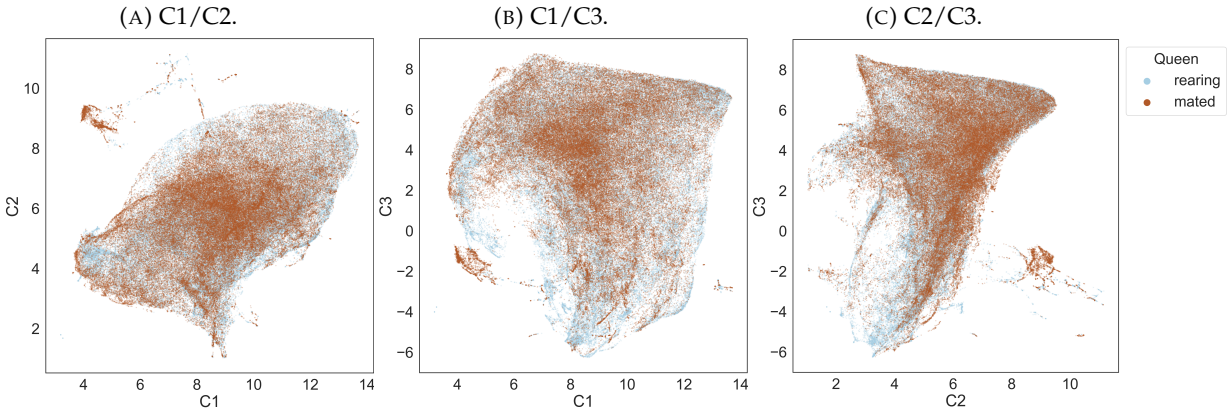


FIGURE 3.5.1: 2D views of 3D UMAP manifold (see section 3.4.1) of in-hive audio data. C1, C2, and C3 refer to UMAP components and indicate viewing direction. Points are coloured by queen state (blue = queen-rearing, orange = queenright). Data from 10 hives over analysis period (27. July – 17. August).

into very different regions of the manifold (compare B1 (Fig. 3.5.2e), C2 (Fig. 3.5.2g), and D4 (Fig. 3.5.2j)). The control hives D1 and D3 also demonstrate the difference in manifold occupation — there was no change in queen state, but the highest density of samples fell into two non-overlapping regions for the two hives. While for each treatment hive, we can identify regions characteristic of the queenless state, the same regions can be characteristic of the queenright state for another hive. This clearly illustrates a major barrier to between-hive generalisability of queen detection models.

Population metrics

Colony size appeared to be slightly correlated with position on C1, with higher C1 coordinates corresponding with smaller colonies (Figs. 3.5.3a–3.5.3c). To a lesser degree, a high coordinate value on C2 also appeared to indicate a small colony size.

The edges of the distribution also tended to contain samples with a small brood area (Figs. 3.5.3d–3.5.3f). Regions with the lowest brood area corresponded with regions that were dominated by queenless samples. This could suggest that these samples came from the later days of the experiment, where treated colonies had been queenless for some time and therefore had a lower brood area.

Colony stores

The size of honey stores was primarily associated with C1, with low stores corresponding to a high coordinate value (Figs. 3.5.4a–3.5.4c). Regions with low honey stores were roughly colocated with those with a small colony size (Fig. 3.5.3a). This seems intuitive given that smaller colonies would have lower foraging strength.

The distribution of pollen stores was less clear. It seems that edges were more dominated by high pollen stores, particular the lower left edge and the upper right edge in the C1/C2 view (Fig. 3.5.4d). These areas somewhat corresponded with regions of low brood, suggesting that when colonies had less brood to feed, the pollen stores grew. The correspondence was also seen in the edge regions in the other views of the manifold, particularly C2/C3 (Figs. 3.5.4e–3.5.4f).

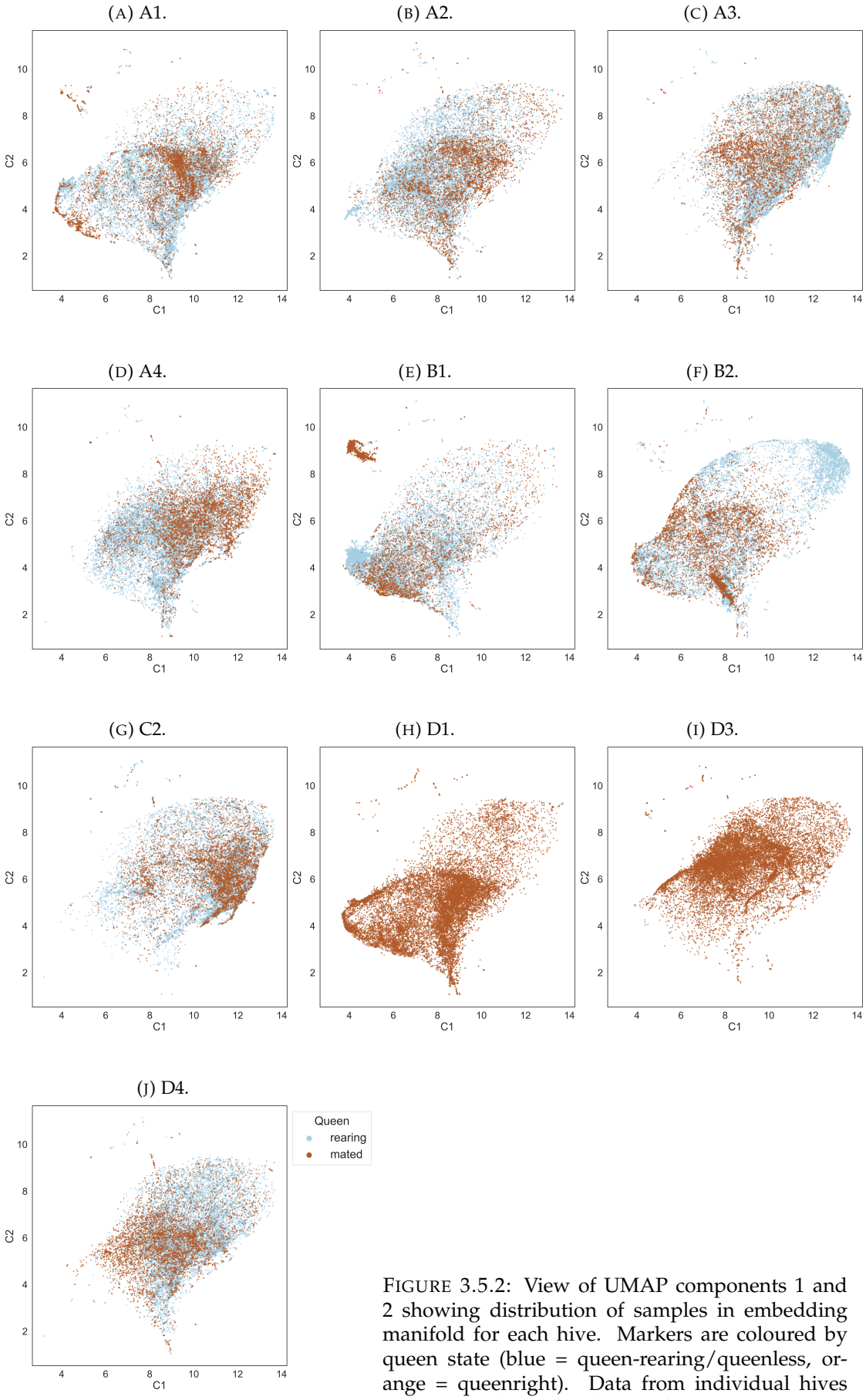


FIGURE 3.5.2: View of UMAP components 1 and 2 showing distribution of samples in embedding manifold for each hive. Markers are coloured by queen state (blue = queen-rearing/queenless, orange = queenright). Data from individual hives over analysis period (27. July – 17. August).

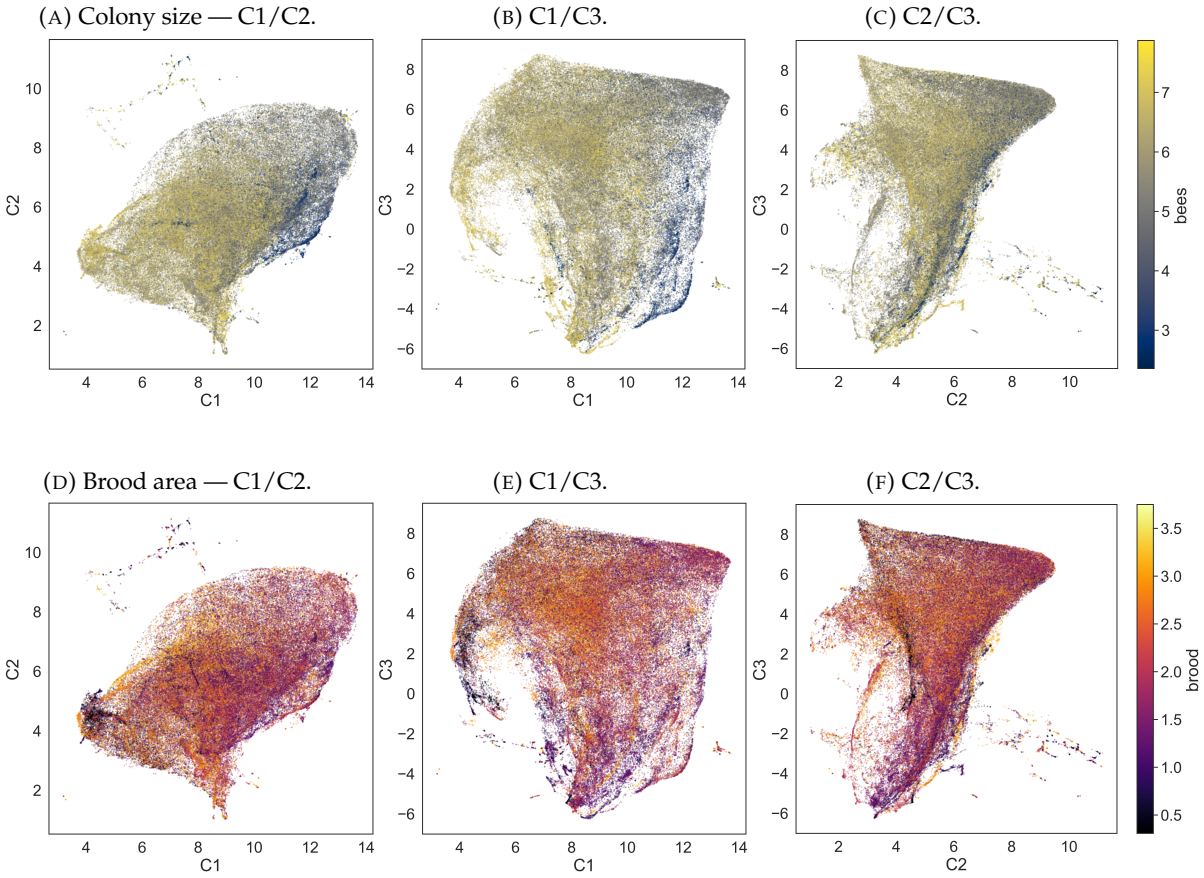


FIGURE 3.5.3: 2D views of 3D UMAP manifold (see section 3.4.1). C1, C2, and C3 refer to UMAP components and indicate viewing direction. Points are coloured by adult population size (A – C; blue = low, yellow = high), or brood area (D – F; purple = low, yellow = high). Population metrics were recorded using Liebefeld scoring. Data from 10 hives over analysis period where census data were available (4. – 17. August).

Sensor & hive placement

Most hives' distributions overlapped in the manifold space, but some regions (especially towards the edges) were hive-specific (Figs. 3.5.5a–3.5.5c). Examples are the stretch of pink (hive C2) to the right middle, the lower left dominated by purple (hive B1), dark blue (hive A1) and brown (hive B2), and the outlying cluster (hive B1) in the C1/C2 view.

Samples from different sensor positions were generally mixed within the manifold (Figs. 3.5.5d–3.5.5f). Recordings from the Inside position appeared slightly more common than others along the top left and bottom left edges. Outside sensor recordings (closest to the hive entrance) were bottom-right of the centre, while Outside 2 recordings (second from the entrance) were more concentrated along the bottom left and

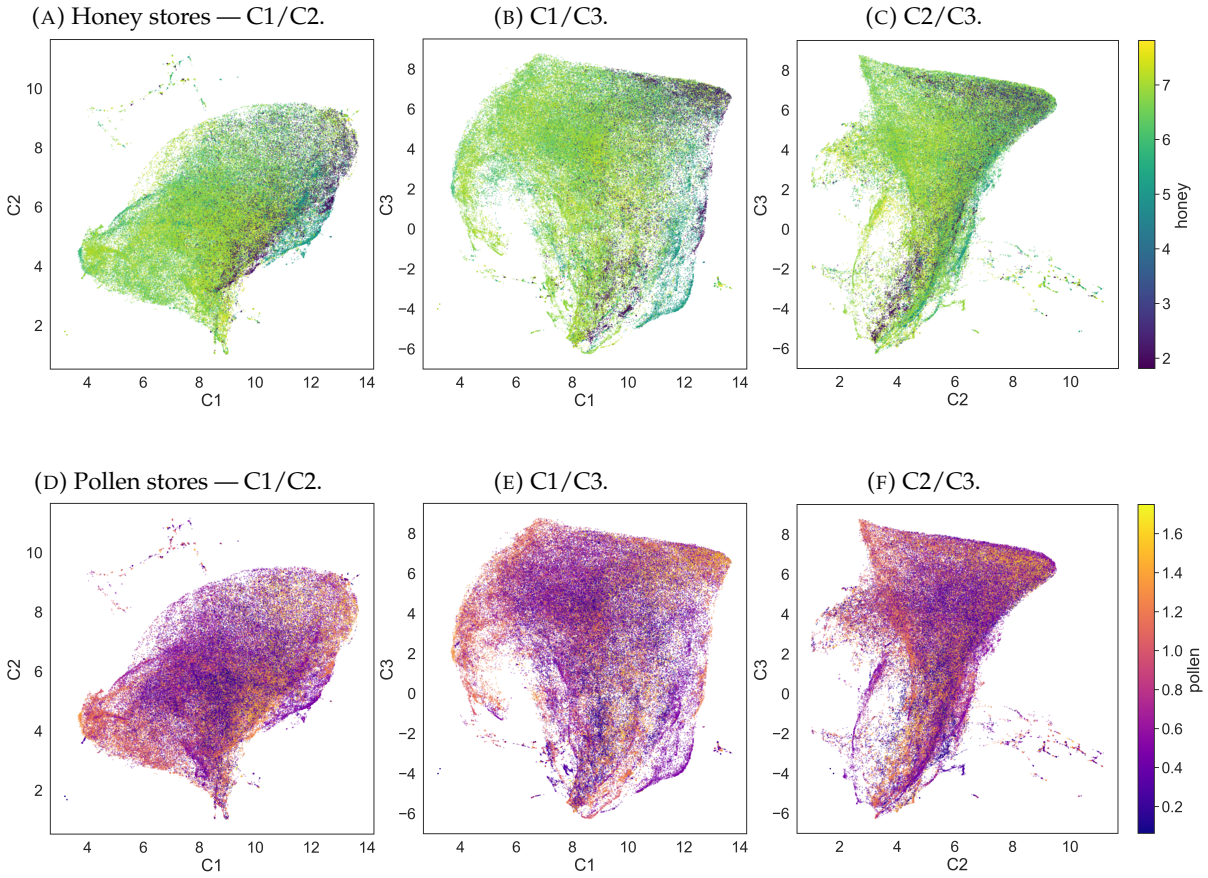


FIGURE 3.5.4: 2D views of 3D UMAP manifold (see section 3.4.1). C1, C2, and C3 refer to UMAP components and indicate viewing direction. Points are coloured by the amount of honey (A – C; blue = low, yellow = high) or pollen (D – F; purple = low, yellow = high) present in the hive. Food storage metrics were recorded using Liebefeld scoring. Data from 10 hives over analysis period where census data were available (4. – 17. August)

right edges. Inside 2 and Centre recordings were distributed throughout.

Similar to individual hive distributions, the pallet distributions overlapped partially with some areas occupied mostly by a single pallet (Figs. 3.5.5g–3.5.5i). One such region aligned with the middle-right area (in the C1/C2 view) that was dominated by hive C2. The A and D pallets were mostly in the centre, while the B pallet had some denser regions in the bottom left and top edge of the manifold.

Time variables

Day within the experiment was not a strong factor (Figs. 3.5.6a–3.5.6c). The outlying cluster mainly consisted of pre-D0 recordings. In the C1/C3 view, recordings from

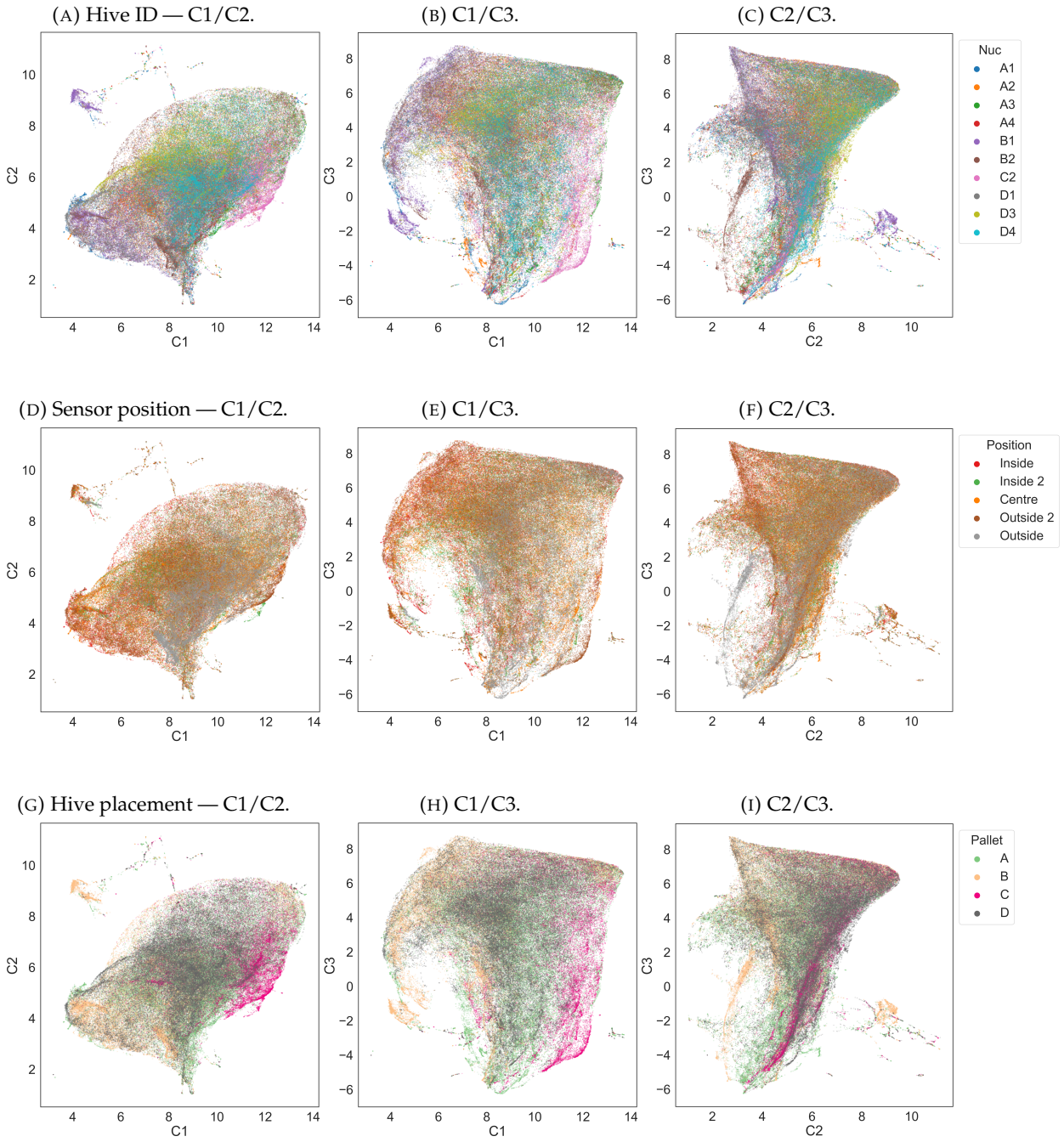


FIGURE 3.5.5: 2D views of 3D UMAP manifold (see section 3.4.1). C1, C2, and C3 refer to UMAP components and indicate viewing direction. Points are coloured by static characteristics — hive ID, sensor position within the hive, and hive position (pallet). Data from 10 hives over analysis period where census data were available (4. – 17. August)

later in the experiment tended to fall outside of the centre of the manifold.

Time of day by contrast had a very strong correlation with location within the manifold (Figs. 3.5.6d–3.5.6f). It appeared strongly correlated with C3, where morning recordings (until 10am) tended to have high coordinate values and afternoon/evening recordings lower coordinate values. The right edge of the manifold in the C1/C3 view (Fig. 3.5.6f) showed a high concentration of early-morning (dark blue) samples.

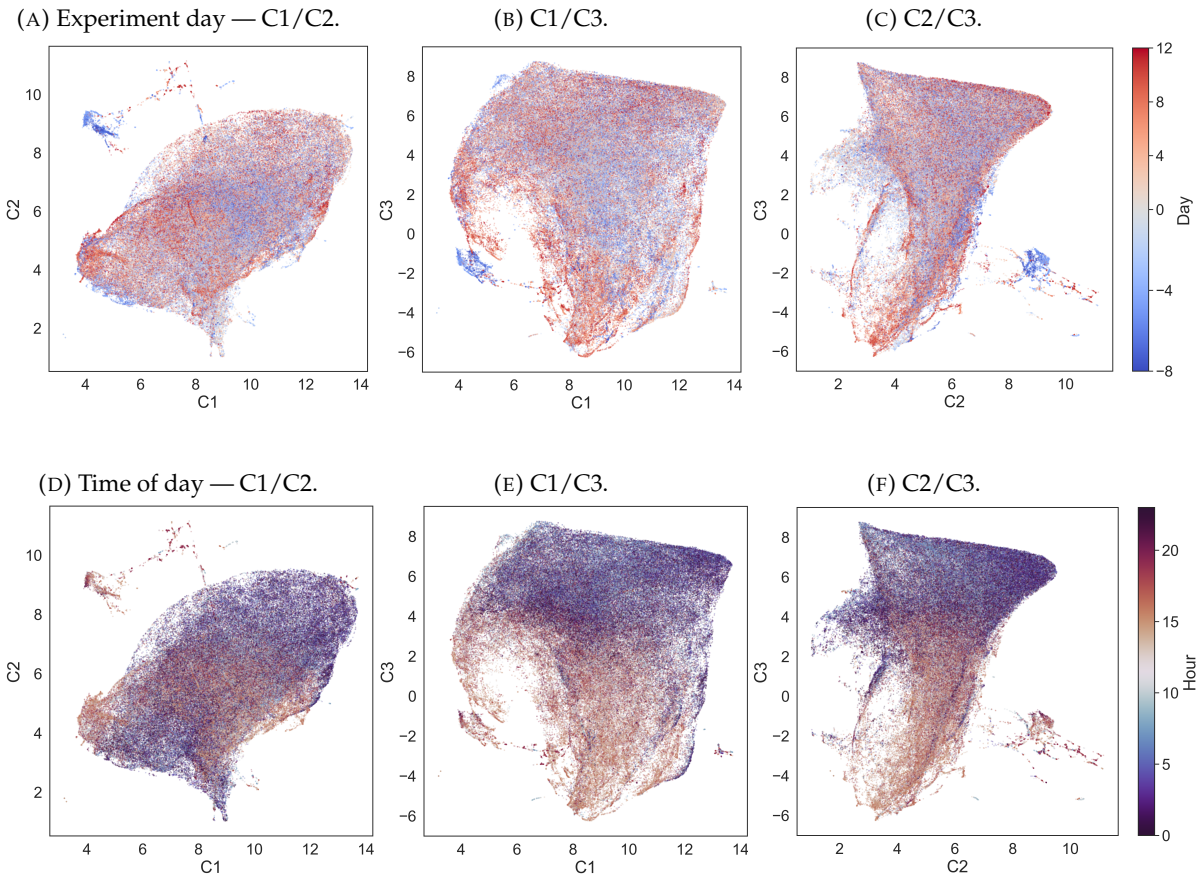


FIGURE 3.5.6: 2D views of 3D UMAP manifold (see section 3.4.1). C1, C2, and C3 refer to UMAP components and indicate viewing direction. Points are coloured by the temporal variables experiment day (A – C; blue = early, white = day 0, red = late) and time of day (D – F; black = midnight, white = noon). Data from 10 hives over analysis period where census data were available (4. – 17. August)

3.5.2 Neural network performance

I used small SLPs with 256 nodes in the hidden layer to predict the queen state label from frequency and waveform features as described in section 2.3.3. With mean OOG test AUCs above 60%, many of the models produced in Chapter 2 were technically usable since their predictions were correct more often than not. However, there were ways to increase prediction accuracy above the per-sample AUC reported.

The simplest strategy was to predict queen state based on several recordings, which was done simply by averaging predictions across several hours or the whole day. This made predictions less vulnerable to spurious high or low model outputs. Prediction time series were refined further by setting confidence thresholds so that only predictions above or below certain quantiles were included (eg predictions between 0–0.15 or 0.85–1), ignoring recordings with ambiguous model outputs. Finally, I explored how model correctness varied with time of day.

Below, I present model outputs for two multi-hive models for hives C2 and A3[3] and discuss how these strategies improve inference. These two hives were chosen since the OOG test performance of their associated MHMs was close to the mean across all hives. Additionally, their respective MHMs were tested on the same control hive D1 and the respective other hive from the training set. This allowed for direct comparison of their performance.

C2

Fig. 3.5.7 shows confusion matrices and prediction time series for a multi-hive model for hive C2. This model with learning rate $\alpha = 0.001$, trained on recordings from the Centre position (repeat 1), was the best-performing model for hive C2. While the model had a very good OOG test AUC of 81%, the false positive rate (FPR) and false negative rate (FNR) were slightly elevated at 21% and 34%. By subsampling predictions to those below and above confidence thresholds at 0.15 and 0.85, discounting any predictions in between, the FPR and FNR were both reduced to 13% and the F1 score was increased from 71% to 88% (cf. Figs. 3.5.7a & b).

[3] In section 3.5.3, I compare feature usage between these same two models.

The prediction time series for test hive C2 (Fig. 3.5.7c) was noisy, but correctly replicated the shift from low to predominantly high predictions. By downsampling the prediction time series, eg by taking the mean across all predictions in a daily window (Fig. 3.5.7d), the noise was reduced and the accuracy was improved.

To assess whether the model had overfit to the training data, I evaluated its performance on three random hives from the training set. Predictions on training hives had a similar or higher level of noise than predictions for test hive C2, suggesting that the model did not overfit (Figs. B.2.1a–B.2.1c). For training hives A2 and A3, model outputs were elevated after day 0, correctly capturing the shift from a queenright state to queen-rearing (Figs. B.2.1a–B.2.1b). For training hive D1, which was a control hive, the model gave predominantly low predictions, correctly predicting the hive to have been queenright throughout (Fig. B.2.1c).

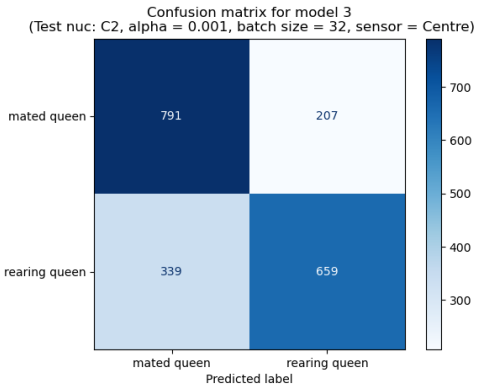
While prediction values were variable, prediction accuracy did not vary strongly with time of day for test hive C2 or training hive A3 (Tab. B.2.1, Fig. 3.5.7e & 3.5.7f). Some minor time dependence was observed for training hives A2 and D1, but accuracy was high overall (Fig. B.2.3b & B.2.3a).

A3

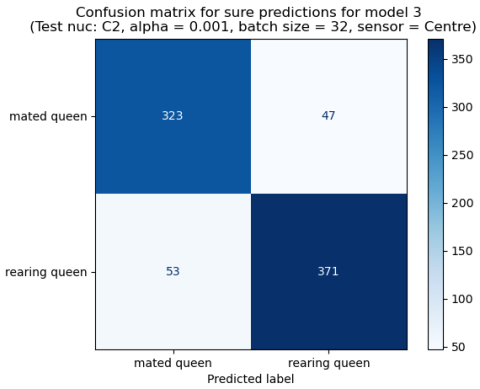
Fig. 3.5.8 shows confusion matrices and prediction time series for one of the three top models for hive A3 (Inside sensor, $\alpha = 0.001$, repeat 3). Despite being the best model for this hive with an OOG AUC of 73% and a very low FPR (7%), it had a low F1 score (45%) and high FNR (69%) (Fig. 3.5.8a). In this case, subsetting predictions to confidence thresholds (accepting predictions between 0–0.15 as negative predictions and 0.85–1 as positive predictions, discarding all other predictions) only marginally improved the F1 score to 48% and lowered the FPR to 4% (Fig. 3.5.8b).

The high FNR of the model is explained when examining the prediction time series (Fig. 3.5.8c). Predictions on the test hive A3 were predominantly low (predicting queenrightness), but exhibited a daily spiking pattern, suggesting that correct predictions were concentrated to a time of day. Examination of test prediction data by time of day showed that between 12:00–16:00, 87% of predictions were correct on average, while outside of this time window, only 57% of predictions were correct (see Tab. B.2.2 and Fig. 3.5.8e in Appendix B). For treatment hive C2 from the training set, there was

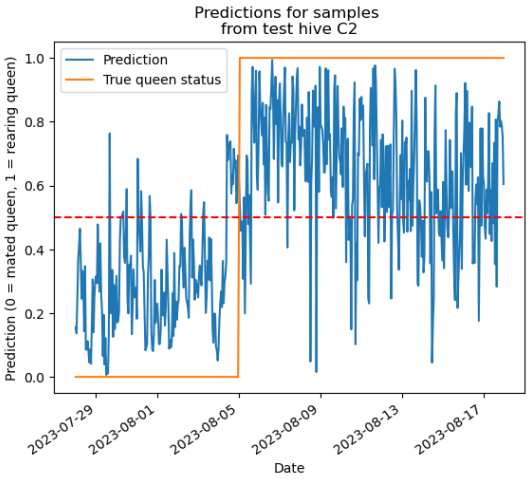
(A) Confusion matrix.



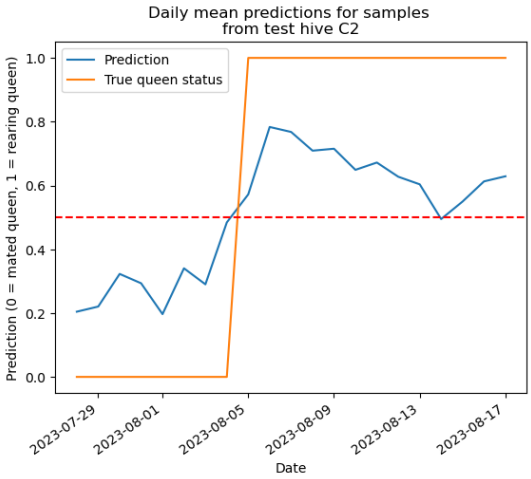
(B) CM — "confident" predictions.



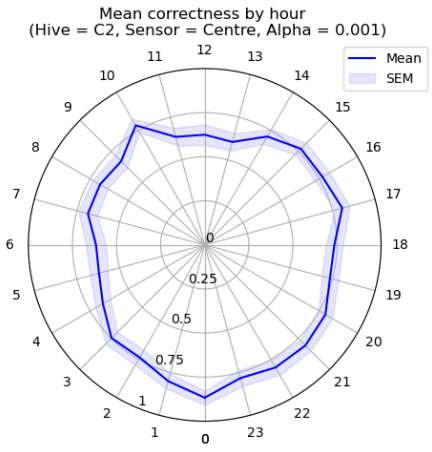
(C) Test predictions — all.



(D) Test — daily.



(E) Test hive C2



(F) Training hive A3

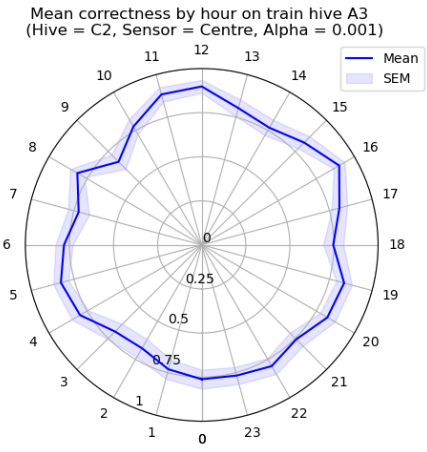


FIGURE 3.5.7: Model outputs of C2 Centre sensor multi-hive model ($\alpha = 0.001$, repeat 1). Confusion matrix of all predictions (A) and confident predictions (B) (y = true label, x = predicted label).
Hive state predictions for test hive C2 over time at sample-level (C) and resampled to daily level (D). Red dashed: 0.5 threshold; orange: true queen state; blue: predicted queen state. 0=queenright, 1=queenless.
Polar plots of mean prediction correctness by time of day for test hive C2 (E) and training hive A3 (F). Blue line: mean correctness (proportion of correct predictions per hour); light blue band: standard error.

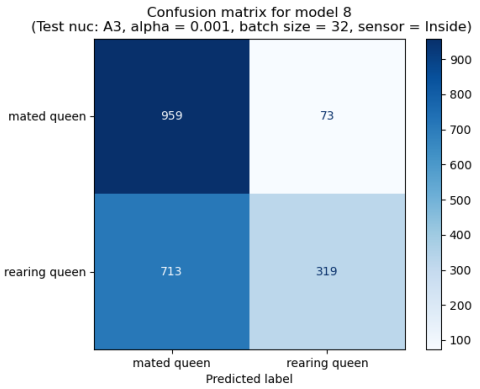
no discrepancy in model performance between the two time windows (85 vs. 86%, cf. Fig. 3.5.8f), but for the control hive D1, prediction accuracy was *lower* from 12:00–16:00 (85 vs. 96%, cf. Fig. B.2.4a).

Downsampling predictions was not as impactful in achieving high accuracy as for the C2 model presented previously. Even post-D0, most predictions remained below 0.5 after resampling to daily mean predictions (Fig. 3.5.8d). However, the predictions presented a clear shift around the date that the queen was removed from the hive.

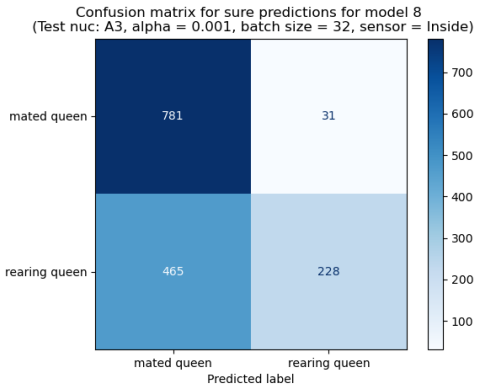
As for C2, predictions on the training set showed no evidence of overfitting, with similar noise levels in prediction time series for test hive A3 and training hives B1, C2, and D1 (Figs. 3.5.8c & a– c).

Though less pronounced than for test hive A3, prediction time series for training hives B1 and C2 also showed some periodicity (Figs. B.2.2a & B.2.2b). The predictions for control hive D1 showed some time heterogeneity — predictions before day 0 were very close to 0, while predictions after day 0 showed a peaking pattern (Fig. B.2.2c). However, the bulk of predictions remained very low, more so than the predictions for test hive A3 (Fig. 3.5.8c).

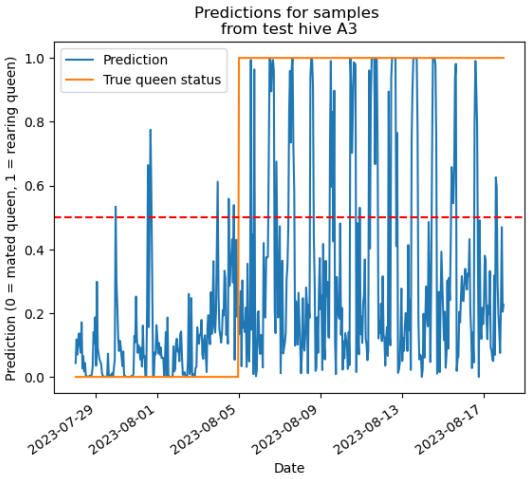
(A) Confusion matrix.



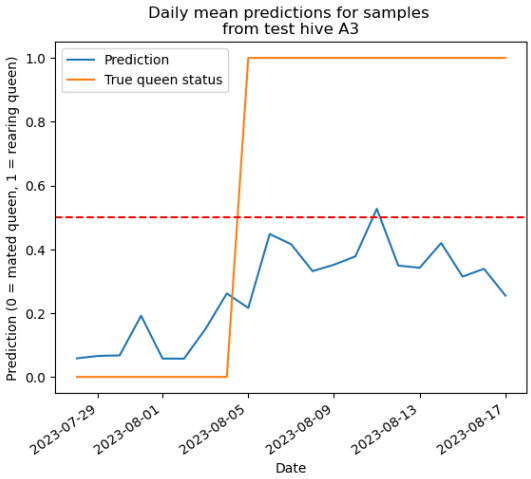
(B) CM — "confident" predictions.



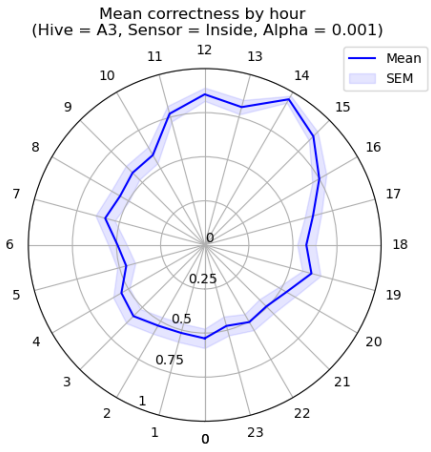
(C) Test predictions — all.



(D) Test — daily.



(E) Test hive A3



(F) Training hive C2

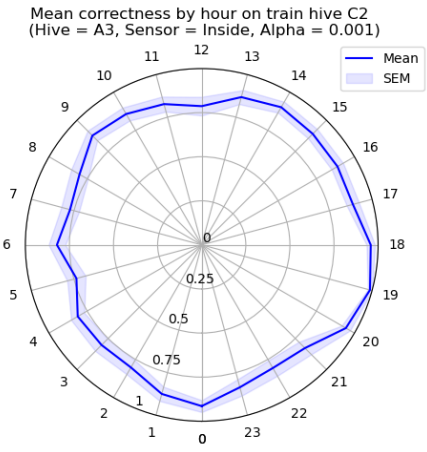


FIGURE 3.5.8: Model outputs of A3 Inside sensor multi-hive model ($\alpha = 0.001$, repeat 3). Confusion matrix of all predictions (A) and confident predictions (B) (y = true label, x = predicted label).
Hive state predictions for test hive A3 over time at sample-level (C) and resampled to daily level (D). Red dashed: 0.5 threshold; orange: true queen state; blue: predicted queen state. 0=queenright, 1=queenless.
Polar plots of mean prediction correctness by time of day for test hive A3 (E) and training hive C2 (F) Blue line: mean correctness (proportion of correct predictions per hour); light blue band: standard error.

3.5.3 Feature importance

I used SHAP values (Python package `shap` v0.42.1, (Lundberg and Lee, 2017)) to understand how the best-performing models used handcrafted frequency and waveform features to predict a hive's queen state.

For an ensemble of the best 24 MHMs (3 best models per hive), I calculated the mean SHAP values for each feature across a consistent set of 10,000 samples (Figs. 3.5.9 & B.3.1). Each point represents how a given feature has contributed to the model output ie the prediction value. A positive contribution increases the output, pushing the model prediction for the given sample towards predicting queen-rearing. Conversely, a negative contribution decreases the output value and makes a queenright prediction more likely.

For most features, the distribution of SHAP values had a high concentration of blue points (representing low feature values relative to the baseline) close to $x = 0$ and a red tail to one side. This suggests some non-linearity in the response of most features: decreasing a feature value below the mean did not tend to have a strong effect on the prediction value, while increasing it often did. For example, in Fig. 3.5.9, low feature values in 'PSD 101.6Hz' did not impact the prediction, whereas high values in this feature biased the model output towards predicting queenrightness. Features with more balanced impact, where both increases and decreases to the feature value affected the prediction in opposing directions, were much rarer. For example, in Fig. 3.5.9, high and low values in the features 'Centroid Mean' and 'ZCR Mean' each contribute to the model output value, but in opposite directions.

Plotting feature impacts in frequency order revealed spectra of impact direction and magnitude (Fig. B.3.1). Impact direction tended to be similar between neighbouring frequency features. This is intuitive since most sounds produced in the hive will show a slight natural pitch distribution, leading to correlation between neighbouring frequencies (cf. section 2.4.1). Therefore, any in-hive sounds like fanning, flight noise from the entrance, or vibrational signals could vary depending on, for example, the body size of the bee or the comb contents (cf. section 2.5.4).

For many features, the positive tails were narrow, indicating that the impact is only observed in a subset of samples (see, for example, 'PSD 43.0Hz' in Fig. 3.5.9).

This could be an indication of interaction effects above second order, where changes to a feature value can have varying effect sizes depending on other features. It could also be an effect of feature correlation (see section 2.4.1). If relevant signals span a frequency range wider than my frequency resolution ($\Delta \approx 3.9Hz$), this could cause importance to be arbitrarily split across neighbouring frequencies, resulting in a smaller contribution allocated to each individual feature. This interpretation is also supported by the observation that neighbouring frequency features tend to share similar SHAP value distributions (see Fig. B.3.1).

Mean feature importance in MHMs

Fig. 3.5.9 summarises the most impactful features across the ensemble of best MHMs. By averaging across models, features with spurious effects in single models are removed since their SHAP values will approach 0 through averaging. Only features with consistent impact on model output will remain. While this method removed features that are highly impactful for some models, the remaining features should be

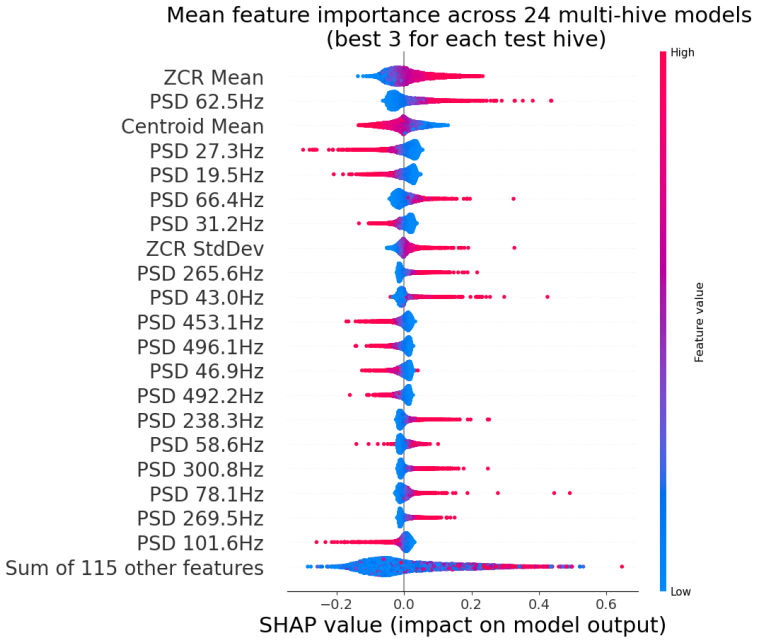


FIGURE 3.5.9: SHAP value distributions of handcrafted features for 24 best multi-hive models. 20 most important features ranked by mean absolute SHAP value. SHAP values represent a feature's contribution to the model output. Features with a positive SHAP value contribute to detecting queen-rearing, while those with a negative value contribute to detecting the mated queen's presence.

broadly applicable across hives, and hopefully represent information about fundamental differences between sounds from queenright and queen-rearing colonies.

Non-PSD features were well-represented. The most important feature was the zero-crossing rate mean, representing percussiveness, closely followed by the spectral centroid mean, representing timbre, ranked 3rd. The ZCR standard deviation was ranked 8th. ZCR mean and centroid mean had balanced impact: A high ZCR mean increased the model output, while a low value decreased the output (ie both red and blue points were concentrated away from 0). For the centroid mean, low feature values increased the output while high values decreased it. A high ZCR standard deviation sometimes increased the output value, but low standard deviations tended to have no effect (ie blue points are close to 0).

Among frequency features, the prevalence of low frequencies (up to ca. 100Hz) was striking. Out of 17 frequencies among the best 20 features, 10 fell below 102Hz. At 19.5Hz, 27.3Hz, 31.2Hz, 46.9Hz, and 101.6Hz, a high Power Spectral Density decreased the output (ie biased prediction towards queenlessness). At 43.0Hz, 62.5Hz, 66.4Hz, and 78.1Hz, a high PSD instead increased the output (ie biased the prediction towards queen-rearing). The PSD at 58.6Hz does not show a clear impact direction — high feature values can both increase or decrease the prediction value, though increases appear more prevalent.

At higher frequencies, impactful frequencies were more sparse. High PSD values at 238.3Hz, 265.6Hz, 269.5Hz, and 300.8Hz (frequencies which have been linked to flight noise and waggle dancing sounds (Esch, 1961b; Vance et al., 2014)) all increased the output value. At 453.1Hz, 492.2Hz, and 496.1Hz, high PSD decreased the output.

Feature importance in individual models

Individual models trained for different test hives did not use the same features, but there was some overlap between their feature lists (Fig. 3.5.10). 8 of 20 features in the list of the 20 most important features for hive C2 also appeared in the list for A3. Of these, 6 matched in impact direction (High PSD at 19.5Hz, 27.3Hz, 226.6Hz, 398.4Hz, 402.3Hz, and 492.2Hz all decreased the model output). The two exceptions, PSD at 43.0Hz and mean spectral centroid, had mixed impact in the model for A3.

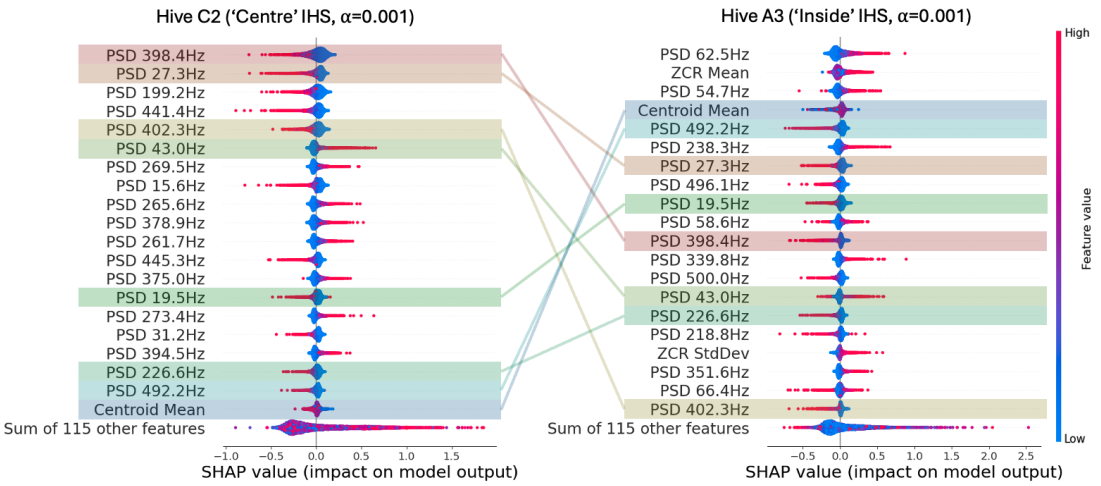


FIGURE 3.5.10: Impacts of handcrafted features on prediction value for best individual MHMs for C2 and A3. 20 most important features sorted by mean absolute SHAP value. 8/20 features were shared between the two models (indicated by matching colour bars).

Each individual model's feature list can also be mapped onto the list of the 20 most important features overall (Fig. 3.5.11). 8 features from the mean 'top 20' list (middle plot in Fig. 3.5.11) were also used by the MHM for hive C2 (left plot). All matching features had the same impact direction across the two lists. High PSD at 19.5Hz, 27.3Hz, 31.2Hz, 492.2Hz, and centroid mean decreased the model output, while high PSD at 43.0Hz, 265.6Hz, and 269.5Hz (perhaps representing waggle dancing sounds) increased the output.

Overlap between the mean top 20 list and the most important features for the hive A3 model was even higher with 11 features in common. However, several were inconsistent in impact direction either within or between the two lists. The ZCR mean and standard deviation both increased the model output when high. The centroid mean overall decreased the output when high, but for the A3 model, its impact on the output value was very inconsistent.

High PSD at 19.5Hz, 27.3Hz, and 496.1Hz decreased the output value, while at 62.5Hz and 238.3Hz they increased it. The other overlapping frequency features (43.0Hz, 58.6Hz, and 66.4Hz) had inconsistent impacts in the model for hive A3.

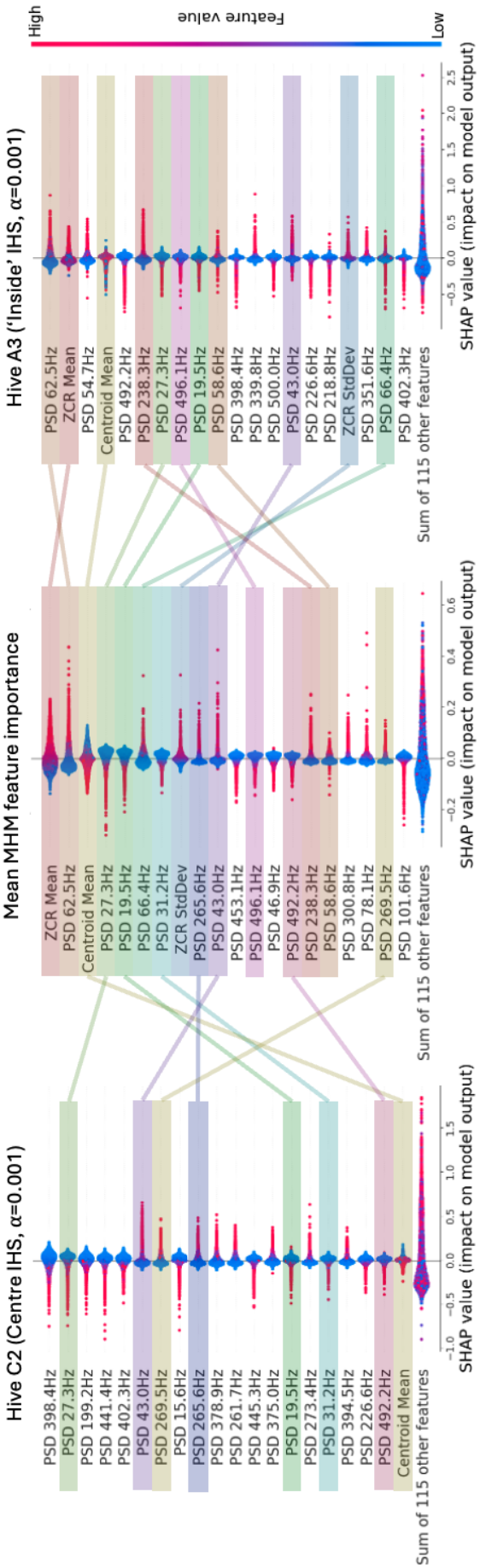


FIGURE 3.5.11: Impacts of handcrafted features on prediction value for best individual MHMs for C2 and A3 compared to mean MHM feature impacts. 20 most important features sorted by mean absolute SHAP value. 11/20 features from the mean top 20 were present in the A3 list (right), and 8/20 features were present in the C2 list (left) (indicated by matching colour bars).

3.6 Discussion

3.6.1 Summary of results

Audio manifold

1.a All investigated metrics (colony size, brood area, honey and pollen storage area, hive ID, sensor position, hive position, experiment day, and time of day) showed some degree of regionalisation in feature space.

1.b The strongest patterns linking metric value and location within the audio manifold were found for time of day, hive ID, hive position, colony size, and honey stores.

1.c Queen state also showed a regionalisation pattern overall, with queenright samples broadly localised to the center and queen-rearing samples towards the edges of the manifold.

1.d Hive-by-hive, the queenright state was highly localised, but the areas of highest density differed strongly between hives, showing that hive ID was a strong determinant of location within the manifold.

Model performance

2.a Performance of multi-hive models can be improved through simple strategies such as downsampling predictions, setting conservative detection thresholds, or only using samples from a specific time window.

2.b Success of these strategies differed between hives.

Feature importance

3.a SHAP analysis can be applied to interrogate model functioning and identify features with consistent impact on model predictions.

3.b Over a subset of 24 high-performing models, the most impactful features included the mean and standard deviation of the ZCR, the mean spectral centroid,

10 low-frequency PSD features, and several frequencies from the flight noise and waggle dancing ranges.

3.c Overlap between the 20 most important features for individual models, as well as with the 20 most important features by mean absolute SHAP value, was considerably higher than expected by random chance (8/20 and 11/20). Averaging SHAP values across an ensemble of models and a defined subset of samples appears to be a meaningful method for testing feature generalisability.

3.6.2 Audio manifold

After extracting audio features, I performed dimensionality reduction with UMAP. I then plotted different hive metrics onto the resulting 3-dimensional manifold to identify regions of the manifold linked to particular hive states or covariates such as hive ID and time of day. Overall, I identified regionalisation patterns in most, if not all metrics tested.

The patterns overlap for several metrics — for example, low honey stores and low colony size both mapped to the same region (C1+ edge of manifold, Figs. 3.5.3a & 3.5.4a), as did low brood areas and high pollen stores (curved process at C1- and parts of C3- tip, Figs. 3.5.3e & 3.5.4e). The latter regions fell into the areas characteristic for queenlessness (cf. Fig. 3.5.1b). These correlations are interesting and biologically reasonable: Small colonies have lower foraging capacity and downregulate nectar foraging (Fewell, Ydenberg, and Winston, 1991; Beekman et al., 2004); queenless colonies produce less brood and colonies without brood consume less pollen (Hrassnigg and Crailsheim, 1998).

However, even stronger patterns of regionalisation emerged for other factors: hive ID was a strong determinant of a sample's location in the audio manifold, as was hive location (a supercategory of hive ID, since hives were placed on pallets in groups of up to 4). Time of day showed the strongest pattern within the manifold, being very strongly linked to components C2 and C3.

Hive identity was also a major determinant of manifold localisation of different queen states. Different hives' audio recordings localised to different regions of feature space even if they shared a queen state, or conversely, occupied the same region of

feature space in different queen states. In all hives, the queenright state mostly was localised to the center of the manifold and the queen-rearing state to the edges. However, the area of highest density of queenright samples always varied, perhaps suggesting that in queenright hives, a variety of other hive conditions affect and dominate colony sounds. Also, the queenright samples were densely and narrowly distributed, while queen-rearing samples were more dispersed throughout the manifold. Based on these results, it may be possible to identify a characteristic signature of queenrightness for each hive, but taken together, there did not appear to be a region of audio feature space indicative of queenrightness or queen-rearing.

Sources of bias

This hive-dependence in the distribution of audio samples within the manifold also colours the interpretation of other hive metrics. The area that appears characteristic of low brood area (curved process at C1-, Fig. 3.5.3e) is in fact simply dominated by hive B1. It is likely that this area is not, in fact, a general signature of low brood area, but is only characteristic of low brood area for this individual hive. Further, the change in localisation of samples within the manifold may not even be linked to brood area directly, but could be an expression of experiment time since the two variables are correlated and cannot be unlinked. A larger dataset, where experiment time and queen state (as well as the other hive metrics) are unlinked would be necessary to investigate connections between audio feature space, different hive metrics, and their co-localisation in greater depth.

A further challenge to this analysis was the interpolation step. Hive inspections cannot be carried out too often at the risk of disruption to the colony. The daily data I used to assign point colour based on each hive metric were therefore based on fortnightly inspections, interpolated to daily frequency using the `PCHIP` method. This method was intuitively the best fit at replicating realistic patterns of hive development — however, its use was not experimentally validated. Colony dynamics models exist (eg Torres, Ricoy, and Roybal (2015)), but their application was outside of the scope of this project and a simpler method was needed. The simplicity of smoothly interpolating between observations can also lead to errors. Brood area can change

quickly from one day to the next, if a large group of capped brood emerge simultaneously. Bad weather can prevent foraging, leading honey and pollen stores to decrease more rapidly. Not knowing how accurately the interpolated time series of hive metrics reflect the true conditions in the hive, the UMAPs of hive metrics prone to sudden changes should be interpreted with caution.

3.6.3 Improving model inference

In section 3.5.2, I discussed the performance and prediction timeseries of two multi-hive models with different prediction profiles. Both models made usable predictions on their respective test hive. Performance was improved by applying simple post-hoc filtering methods, but the best method differed between the models and test hives.

Narrow thresholding

The performance of the C2 model was improved by thresholding predictions. Accepting only predictions within 0.15 of 0 or 1 improved the FNR, FPR, and F1 score. However, for A3, the same strategy was unsuccessful, since this model's predictions were concentrated below 0.5. Therefore, most positive detections were ignored when applying this simple thresholding technique. Even though this strategy can only be applied when models exploit the full prediction space between 0–1[4], it is powerful when it can be used.

Resampling

The model for C2 benefitted greatly from resampling predictions to daily frequency, which removed noise and consistently elevated predictions above the queenless detection threshold at 0.5. By contrast, the A3 model did not benefit in the same way from this method — here, downsampling removed the predictions above the detection threshold since most correct high predictions were only slightly elevated. This presents an argument for a baseline-guided approach, rather than a thresholded approach, to making hive state predictions. By checking whether model outputs are consistently elevated above a baseline, eg the mean output over the past two weeks,

[4]Some models rarely output values close to 1, see section 3.5.2 — A3.

model performance could be improved for every individual hive, even if the output does not exceed the standard threshold of 0.5. Here, I applied a very simple downsampling procedure without overlapping windows, but a moving average would work very similarly and would be more suitable in a deployment scenario, where regular updates are desirable.

Periodicity

Predictions for A3 were strongly periodic, with a high proportion of correct predictions during the early afternoon and a low proportion during the rest of the day. This periodicity was not replicated for other treatment hives, but can be observed in control hive D1, where accuracy was decreased during the 12:00–16:00 window by 9 percentage points. No periodicity was observed in any of the prediction time series of the C2 model.

It is not clear why the A3 MHM might have utilised time-dependent information to make predictions, while the presented model from C2 showed evenly high performance around the clock. While this seems counterintuitive, it could reflect an ability to make more nuanced predictions during this time window, while at other times the model might have been biased towards low predictions. It could also indicate that behavioural shifts occurring after queen loss are most notable at a particular time of day in some hives, like A3. This could be the case either if the behaviour is only performed within a certain time window, or if its sounds are overwhelmed by other sources of noise outside of this time window. In other hives, like C2, the behavioural shifts seem to be detectable throughout the day and night.

For most of their lives, adult worker bees do not experience rhythmic daily cycles of activity and rest (Moore, 2001). Only foragers and food storers were found to follow diurnal cycles (Klein et al., 2008). Nevertheless, groups of bees together express day-night cycles in their locomotor activity (Grodzicki, Piechowicz, and Caputa, 2020), and diurnal variation in in-hive noise has been previously established, including my own results (Figs. 3.5.6d–3.5.6f). Upon queen loss, I expected to see changes mainly in brood rearing behaviour, since there is less brood and therefore less brood care required. Brood care activities, however, are performed around the clock by young workers which do not follow a diurnal cycle. It therefore seems more likely that daily

variation in SNR would be caused by other, obscuring sounds. The drop in in-hive population caused by foraging during the day could impact noise levels and SNR in populous hives, like A3 (Fig. B.1.1b). Smaller hives, like C2 (Fig. B.1.1a) might have less daily variation in SNR since the foraging population is smaller. The factors influencing model performance throughout the day, including variation between hives and models, have not yet been investigated in depth and this should be a next step.

Sources of bias

Developing a queen state detection model on this dataset was slightly complicated by the non-matched experiment design, where the number of control hives was lower than that of the treatment hives. This limited my model testing procedure so that MHMs were only tested on a single treatment hive, since both control hives were used for training. Holding out a control hive for testing would have further imbalanced the training set and increased the chance of environmental effects dominating model output. The limited test set of a single treatment hive only allowed me to verify that models were able to detect a shift in queen state on day 0, but not that they could also detect the *absence* of the shift in control hives that were continuously queenright, or a shift to queenlessness on another day.

There were potential sources of location- and time-linked bias. Control hives were located on the same pallet (D). This could have led to similar levels in environmental sound contamination in recordings from both control hives, which could make them distinguishable from treatment hives.

Such patterns in environmental sounds could also provide the model with clues to time-within-experiment. Queen loss followed the same schedule in all treatment colonies. While this is typically a prerequisite control for any experiment, it can also limit the inference we can draw from results of predictive ML models. This is because a model could potentially learn to associate external information, like an increase in rainy or windy weather, with the shift in queen state, and use this information to make largely correct predictions. Such models would not perform well on any colonies outside of the experiment, since they had not actually learnt to extract information about colony state from hive sounds.

The hive A3 MHM showed some time heterogeneity in its predictions, both throughout the day and the experiment. Predictions on the test hive A3 as well as on the control hive D1 (used for training) were correctly low prior to the queen removal date, but showed a regular peaking pattern post-D0. The sudden start of the peaking coinciding with the removal of queens in other hives suggests that the model may, to some degree, be using environmental noise or other external information to make predictions.

No comparable peaking pattern or time heterogeneity was observed for prediction time series of the MHM for hive C2. The prediction shift coincided with the true queen state shift, and for control hive D1, no shift was predicted.This indicates that the model did not rely on environmental sounds to make its predictions — if it did, then prediction values would have increased after day 0. While prediction values increased slightly on day 0, most remained far below the standard detection threshold at 0.5.

Balancing future experiments by having the same number of control and treatment hives would allow for more rigorous testing and prevent models from using such workarounds. Another measure would be to break the association of time with queen state by staggering queen removal dates between hives, and potentially reintroducing the queen after a queenless period. The greater number of state changes and their asynchronicity in time would reduce the model's ability to derive hive state predictions from external information.

3.6.4 SHAP value averaging over model ensembles

I used SHAP analysis to identify which audio features consistently contribute to classifying recordings as queenright (decreasing model output) or queen-rearing (increasing model output).

Multi-hive models cannot specialise to a single-hive, since they were trained across a number of them and then evaluated on an unseen test hive. That test hive's characteristic sounds may differ in critical ways to what the model has been trained on (as demonstrated in Fig. 3.5.5a – 3.5.5c). The out-of-group test performance (AUC) consequently was lower than the within-group test performance (where models were tested on a random subset of data from training hives). Nevertheless, OOG test performance remained good enough to make functional predictions, especially when averaging

model outputs over several hours and making decisions based on the agglomerated prediction.

To understand which features had consistent impacts on model predictions and might offer some ability to generalise, I calculated SHAP values for an ensemble of models with high OOG test performance. The three best models were selected from each test hive. I considered the mean feature importance distributions derived from this ensemble of multi-hive models most reliable, since they included two levels of averaging — first, within the model, which had to take into account information from 9 different hives in order to successfully predict on an unseen hive; second, across 24 models, which had been optimised for 8 different hives. The features that, on average, had a clear impact direction and ranked highly were considered most likely to be transferable features that could help classify queen state on recordings from new hives.

Feature importance values showed some overlap between multi-hive models for different hives, but also many unique features (shown for C2 and A3 in 3.5.10). This suggests that models trained on different hives utilised distinct sets of features to make predictions. It was also consistent with the hive-dependent differences in manifold regionalisation of queenright samples (see sections 3.5.1 and 2.5.4). This again demonstrates that in-hive soundscapes differed strongly between hives.

Overlap in feature usage by individual models and the ensemble was similar or higher than the overlap between individual models (Fig. 3.5.11). Strikingly, 15 of 20 features from the averaged list were among the 20 best features for at least one of the two individual models. Overlap between the averaged list and A3 was higher than between the averaged list and C2. This is slightly counterintuitive, since the A3 model showed worse performance on the test set and indications of environmental noise contamination in the training set.

Frequency features

My SHAP analysis revealed a suite of frequency features that were consistently more or less prevalent in queenless hives compared to queenright hives. By linking the frequency ranges found to be important for distinguishing between queen states to behaviours known to be associated with that frequency range, we can gain insights

into which behaviours might be up- or downregulated during early queenlessness and queen-rearing.

A remarkable outcome of the feature importance analysis is the high prevalence of very low frequencies among the most impactful features. When considering mean SHAP values across the best MHMs from each hive (Fig. 3.5.9), 9 of the 20 most important features are frequencies below 100Hz.

The role of low-frequency thoracic vibrations in the hive as individual markers of activity level has been demonstrated by Hrncir, Maia-Silva, and Farina (2019) (see section 1.3.2). Both the amplitude and frequency of this thoracic vibration increased with level of activity, from ca. 30Hz for inactive hive bees to ca. 45Hz for returning foragers. These thoracic vibrations could function as reliable activity indicators for nestmates. While is is unclear if the activity vibrations of all workers together would produce an audible sound signature, 31.2Hz and 43.0Hz were both among the 10 most important features across all 24 best models. High PSD at 43.0Hz and low PSD at 31.2Hz were associated with queenlessness. This could suggest a general increase in activity, with fewer vibrations of inactive hive bees and more vibrations of active foragers recorded.

The PSD at 19.5Hz was among the 20 most important features for both individual MHMs as well as when averaging across all 24 best-performing models. This frequency falls into the range associated with the shaking signal (though it is not known if the shaking signal has a significant acoustic component, see section 1.3.2; Łopuch and Tofilski (2019)). A decrease in PSD at this frequency is linked to queenlessness, suggesting that fewer activity-inducing shaking signals are performed after queen loss.

The frequency range typically associated with flight noise (200–250Hz, Vance et al. (2014) and Papachristoforou et al. (2008)) was represented by a single PSD feature. 238.4Hz appeared among the 20 most important features in several models and is ranked 17th on average across multi-hive models, but no other flight frequency appeared in the mean top 20.

Two frequencies from the waggle dance frequency range of 250–300Hz (Esch, 1961b; Esch, 1961a; Michelsen, Kirchner, and Lindauer, 1986) were represented among the 20 most impactful features by mean absolute SHAP value in multi-hive models (265.6,

269.5, and 300.8Hz). Out of these, 269.6Hz was also an important feature for the C2 multi-hive model, but was not used by the A3 MHM. Increases in flight noise and waggle dancing could be explained by an increase in the forager population resulting from the release from QMP control (Pankiw et al., 1998). The relevance of this frequency range to queen loss detection was also highlighted by Kanelis et al. (2023).

Above the waggle dance range, 453.1Hz, 492.2Hz, and 496.1Hz were important. These fall into the range of queen piping, specifically tooting (see section 1.3.2). Tooting is typically performed by the first virgin queen to emerge. The timespan of my dataset was chosen to only include pre-emergence recordings, so I did not expect to find any queen piping. Nevertheless, it is possible that queen replacement proceeded faster than I anticipated, leading to the inclusion of recordings with queen piping, a clear signal of queen rearing.

During queen replacement, we would expect to see tooting at 300–500Hz followed by quacking at 180–370Hz (fundamental frequencies, Michelsen et al. (1986)). Since the quacking frequency range overlaps with the flight and waggle dance ranges, we cannot distinguish these behaviours by frequency alone. Instead, a more in-depth spectral analysis of the harmonics and time-domain patterns would be required to determine if these impactful PSD features are linked to queen piping or to other behaviours.

Other features

Two broader features often had very high SHAP values. In particular, the mean ZCR was ranked 1st when averaging SHAP values across all best MHMs (and ranked 2nd for the A3 MHM), while the mean spectral centroid was in the top 5 for many models. Both means had consistent impacts on the prediction: a high ZCR increased the output, while a high mean centroid generally decreased it.

The ZCR is extracted from the waveform by calculating the rate at which the waveform crosses the zero-amplitude line. It functions as an approximation of the dominant frequency in a signal and can be used to describe the timbre of a sound sample. Since high-frequency waveforms oscillate between positive and negative air pressure more rapidly than low frequencies, recordings with more high frequencies would tend to have a high ZCR. This feature has in the past been used for the discrimination of

percussive sounds, as well as voice detection and musical instrument classification (Gouyon, Pachet, and Delerue, 2000; Alías, Socoró, and Sevillano, 2016).

The spectral centroid is also an indication of which frequencies dominate a recording, since it measures at which frequency most of the energy in a sample is concentrated. The association of a lower spectral centroid with queen-rearing aligns with the high importance of low-frequency features previously discussed.

My feature importance analysis found that a low spectral centroid was associated with queen rearing (increase in model output) whereas a low ZCR was associated with queenlessness (decrease in model output). Both the spectral centroid and the ZCR are ways of estimating the dominant frequency, but they affected model outputs in opposite directions. It is unclear why these ostensibly similar features had opposing effects, but they appear to be distinct enough for models to use them differently.

Limitations and further work

The in-hive environment is very noisy due to the large number of individuals and sound sources. In addition, noises from the outside environment may be audible inside of the hive, especially if sensors are placed close to the entrance. This could lead to contamination of in-hive recordings, allowing hive state detection models to use additional information when making predictions. Environmental noise contamination of the in-hive environment should be investigated in future work to better understand how prevalent it is and, if necessary, how it could be filtered out.

Bees naturally cover foreign objects in the hive with propolis or wax (Winston, 1987). Due to how the BeeHero IHS are placed over the top bar of a frame, the sensor becomes integrated into the comb. It is possible that this affects the resonant properties of the sensor itself and potentially the recorded vibrational spectra as a result. In this experiment, the IHS were placed into the hive one week before they started recording. Whether this was enough time for the sensors to be stably integrated into the comb is not clear. If not, then early recordings would differ from later ones due to any effects of being partially covered in wax. Whether such an effect exists and how long it takes for the effect to stabilise should be investigated. Barring experimental evidence, future experiments should aim to place recording equipment into the hive well in advance

of the scheduled recording start date and verify by observation that it has been stably integrated.

Honeybee lifecycles are strongly seasonal, and hive activities and colony composition, as well as principal dangers, vary throughout the year. Controlled experiments manipulating honeybee hive state on the other hand are limited in time. It is highly likely that sound signatures of a particular hive state are not transferrable to other seasons due to the covariation of many other factors determining hive state. The effect of weather, especially wind and rain, on in-hive sounds also has been demonstrated in the past (Tlačbaba et al., 2014; Farina, 2023). While the models and features presented here have demonstrated some ability to generalise, applying them to hives in a different location or at a different time of year would almost certainly present a substantial challenge. Models with greater ability to generalise can only be developed with a much larger and well-annotated dataset spanning multiple seasons and climate zones.

My methods for increasing model accuracy have shown that some models are more accurate during certain time windows. Including time of day as a feature could be a promising way to improve model performance. I also observed that some models rarely output very high values, but that predictions increased at the appropriate time nevertheless. This suggests that baseline-guided models, which take into account the recordings taken across eg the previous 14 days and detect deviations from this baseline, could be a promising new approach.

A possible limitation to my feature importance results is the choice of explainer. I used `shap`'s permutation explainer, which may not be suitable for correlated features. It attempts to fairly assign impact to all important features so that the sum of individual contributions is equal to the model output (Additive feature attribution, Lundberg and Lee (2017)). However, if several features are correlated, they would effectively share impact, reducing the SHAP value of each individual feature. Beeswarm plots of feature importance across the full feature spectrum (Fig. B.3.1) show some evidence of feature correlation — neighbouring features often affected model outputs in the same direction. However, in groups of features with similar importance there were usually one or two features with the biggest impact. It appears that despite the possible limitations, the permutation explainer was able to identify the most important features

this way. Nevertheless, future work could consider increasing the interval between PSD features to reduce the strength of correlations between neighbouring features.

Finally, I focused my analysis on the frequency range 0–500Hz, since most fundamental frequencies of honeybee vibrational signals fall into this band. Other studies found that the frequency band of 0–550Hz gave the highest classification accuracy compared to 450–1050Hz and 950–1550Hz, but using the full range of 0–1550Hz produced the best results (Biswas, 2020). Future experiments could try to extend the frequency range by recording at higher sampling rates, if transmission and data storage constraints allow.

3.6.5 Conclusion

In this chapter, I presented an in-depth investigation of the feature space, predictive performance, and feature usage of some of the models developed in Chapter 2.

I found that the set of handcrafted features formed a continuous manifold in UMAP embedding space. Several hive metrics, most notably time of day, hive ID, colony size, and honey stores, showed patterns of regionalisation, where particular states mostly occurred in one region of the manifold. I suggested possible reasons for colocation between different metrics, but also discussed the likely outsized influence of individual hives on these patterns. Queen state showed strong regionalisation overall, but patterns for individual hives deviated strongly from the average, highlighting substantial differences in the queenright soundscape between hives.

I tested three methods for improving prediction accuracy on the test set post-hoc by downsampling predictions or filtering them based on the time of day or confidence thresholds. Downsampling was successful for both models examined in denoising the prediction time series. Subsetting predictions to conservative confidence thresholds (0–0.15 and 0.85–1) strongly improved prediction accuracy for one model (C2), but not for the other due (A3) to its bias towards low model outputs on the test hive and therefore few predictions above the confidence threshold. Filtering predictions by time of day proved more suitable for this model (A3), which showed stark contrasts in prediction accuracy depending on the time of day. Prediction accuracy of the other model (C2) was mostly independent of time. When examining prediction time series

on training data, I also found some evidence of environmental sound contamination in one model (A3) but not the other.

I applied SHAP analysis to identify features with high impact on model predictions. Individual models for hives C2 and A3 used different suites of features, though there was some overlap. Both feature sets also overlapped significantly with the 20 features with highest mean absolute importance among the 24 best models. The most impactful features included the ZCR, the spectral centroid, 10 PSD features below 102Hz, as well as frequencies in the flight noise and waggle dancing ranges. I discussed likely behavioural changes in the hive following queen loss and placed my feature importance results in the context of honeybee vibrational communication.

Chapter 4

Acoustic monitoring of colony nutritional state

Contributors:
Stella Marie Felsinger (study design, experiment design & setup, data collection, data analysis, interpretation of results, writing),
Léa Beaupère (experiment setup, data collection, image analysis),
Kieran Walter (data collection, image analysis),
Raquel Texeira de Sousa (experiment setup),
Rui Gonçalves (BroodCam development),
Beth Mortimer (supervision, interpretation of results),
Geraldine Wright (supervision, study design, interpretation of results).

4.1 Abstract

Pollen nutrition affects a honeybee colony's productivity and long-term survival through direct and rapid effects on brood production and interactions with disease stress. Though monitoring nutritional state is a central beekeeping task, acoustic hive monitoring of nutritional state has never been attempted previously. Further, it is unknown how extended periods of nutritional stress impact brood rearing when pollen becomes available again.

Here, I collected a set of in-hive audio and microclimate data from 27 honeybee colonies experiencing pollen starvation and infection with a fungal pathogen. I observed that colonies starved for up to 3 weeks recommenced brood rearing once pollen

feeding was reinstated; however, brood rearing capacity after 2 weeks of feeding decreased with longer starvation periods. Chalkbrood infection was not found to significantly interact with starvation in affecting brood production. Handcrafted audio and microclimate features and time of day were not sufficient to predict nutritional state using neural networks, but showed some success for colony strength prediction in unseen hives.

These results reveal the relative difficulty of different prediction tasks, which likely relate to the underlying biology and behaviours associated with the predicted colony states, as well as how the prediction task is formulated. My results suggest that honeybee colonies are remarkably resilient to starvation by employing rescue mechanisms like brood cannibalism and forager reversal. Nevertheless, extended periods of starvation critically reduce brood rearing, with fatal consequences late in the season.

4.2 Introduction

In commercial apiculture systems providing pollination services, honeybee colonies are routinely kept in monoculture settings where pollen sources are limited to one or a few different plant species. This can impact colony productivity and health.

Pollen dearth affects brood production primarily through the nurse bees' ability to feed young brood. Nurses, 6–13 day old workers, are responsible for feeding the entire colony (Winston, 1987; Crailsheim, 1991). Nurses process hard-to-digest pollen into protein- and lipid-rich secretions of the hypopharyngeal gland, which are then fed to other hive members, most importantly young brood and the queen (Wright, Nicolson, and Shafir, 2018).

Under conditions of pollen starvation, brood development is impaired. Pollen-limited worker brood develop into poor foragers and waggle dancers and pollen-starved drone brood have lowered reproductive quality (Scofield and Mattila, 2015; Szentgyörgyi, Czekońska, and Tofilski, 2017). Prolonged pollen dearth can even lead to brood cannibalism, where nurses consume eggs and young larvae to recover the contained nutrients (Schmickl and Crailsheim, 2001; Schmickl and Crailsheim, 2002), leading to overall reduced brood production.

Low quality and diversity of pollen diets have also been linked to reduced immunity and higher pathogen loads (Alaux et al., 2010; Di Pasquale et al., 2013; Cavigli et al., 2016; Dolezal et al., 2019), as well as lower resilience to pesticide exposure (Pettis et al., 2013; Archer et al., 2014). Some diseases, like chalkbrood disease (*Ascosphaera apis*), do not kill the colony directly, but can significantly reduce brood production (Aronstein and Murray, 2010).

Drops in the brood production and foraging strength of the colony have knock-on effects on pollination strength and ultimately colony survival, impacting crop yields and beekeepers' operational costs (Klatt et al., 2014). It is therefore essential to avoid prolonged periods of pollen dearth.

Monitoring nutritional state currently relies on beekeeper inspections, as no automated methods exist for this task. Given typical inspection rates of 1-2 weeks, pollen dearth could go unnoticed for several weeks, potentially causing irreparable damage to the colony. Pollen starvation is known to profoundly affect colony behaviour and condition, including anatomical changes (Szentgyörgyi, Czekońska, and Tofilski, 2017), increased and precocious foraging (Schulz, Huang, and Robinson, 1998; Schulz et al., 2002; Mayack, Natsopoulou, and McMahon, 2015) and brood cannibalism (Rösch, 1930; Woyke, 1977). These shifts may alter the colony soundscape sufficiently to enable passive acoustic monitoring of colony nutritional state.

Pollen supplements or substitutes have been shown to increase brood production (Avni, Dag, and Shafir, 2015), but their use in recovering from a period of pollen dearth has not previously been explored. Understanding the ability of honeybee colonies to survive prolonged periods of pollen dearth and restart brood production has clear implications for colony recovery following deployments in monoculture settings.

In this chapter, I investigate the impact of a period of pollen starvation, followed by a period of pollen feeding, on brood production and chalkbrood infection in 27 honeybee colonies. I test the use of in-hive sound and microclimate data in predicting brood production and colony strength and look at effects of training set size and diversity, test set choice, and prediction target on model performance.

4.2.1 Hypotheses

Experimental

H1 Numbers of capped brood will decrease during a period of pollen starvation and will increase again once pollen feeding is restarted.

H2 There exists a point after which a colony is too depleted of nutrients to recover, even with optimal feeding (multifloral pollen diet).

H3 The critical pollen starvation period after which a colony will not recover lies between 7 to 21 days.[1]

H4 Chalkbrood severity will be lower in control colonies than in colonies experiencing pollen starvation.

Modelling

H5 Handcrafted audio features, temperature, humidity, and time of day are sufficient to allow prediction of colony condition with ANNs.

H6 Training set hive diversity increases cross-hive generalisation ability.

H7 Within-hive generalisation in time is easier than cross-hive generalisation.

H8 Colony strength is an easier prediction target than nutritional state.

[1] The tested durations were chosen based on a preliminary experiment which saw strong effects after 2 weeks of pollen starvation, including brood cannibalism and queen rejection.

4.3 Experiment methods

I manipulated the nutritional state of 32 honeybee mini-colonies in a controlled environment while recording their in-hive sounds and performing regular hive inspections. The experiment was carried out in a semi-field environment without natural food sources. Colony nutritional intake was controlled by providing patties of either pollen or beekeeping candy in an internal feeder.

The experiment included 4 groups, each containing 6–7 hives: 1 control group, who received pollen patties throughout, and 3 treatment groups, who experienced varying lengths of pollen starvation from 1–3 weeks. I monitored colony condition before, during, and after starvation to assess impacts of starvation length on capped brood production.

The experiment began with 36 pollen-fed colonies, but only 32 survived to D0 (1. September 2023) to be included in the treatment phase of the experiment. One colony from the control group (nuc 25) absconded in the first week after treatment began. A further 4 colonies declined to below 1 beeseam in the 4 weeks following D0 and were excluded from analyses of hive audio and census data. Another 2 colonies did not have sufficient audio data available, likely due to low batteries in the IHSs. The final dataset for audio analysis contained 25 colonies while the census data included 27 colonies.

4.3.1 Enclosure

Colonies were kept in a purpose-built polytunnel environment (Fig. 4.3.1). The mesh covering of the polytunnel allowed natural light, wind, and rain in, but prevented foragers from leaving the experiment space. The ground was soil and grass, which was kept mowed to prevent the growth of nectar- or pollen-producing plants.

Large buckets of water were placed in all corners of the room. Each bucket contained bricks and stones at water level to provide places for water collectors to land safely. Ad-lib hanging syrup feeders were placed along the length of the room on each side, each containing between 250–400ml of sucrose syrup. Feeders were cleaned and refilled once per week. The sucrose concentration was slightly adjusted from week to week depending on ambient temperature and existing honey stores in hives, but

always ranged between 20–30% sucrose. The provision of water and syrup feeders allowed colonies to express natural foraging behaviours.

4.3.2 Hive configuration

I used APIDEA mating hives to house my mini-colonies consisting of a mated queen and ca. 2000 young worker bees. Each hive consisted of a brood box with 5 frames, a top feeder with a hole cut into it, an acrylic cover, and a lid (Fig. 4.3.2). Hive boxes, lids, and entrance doors were spray-painted different colours to help foragers return to the correct hive.

4.3.3 Colony setup

Of the 36 colonies used in the experiment, 19 were used in a previous experiment in June–July 2023 and 17 colonies were newly established for this experiment. On 20. July, I collected nurse bees from full-size hives in the university apiary (John Krebs Field Station, Wytham, Oxfordshire) by shaking frames with open brood into a mesh



FIGURE 4.3.1: Experiment setup in polytunnel.

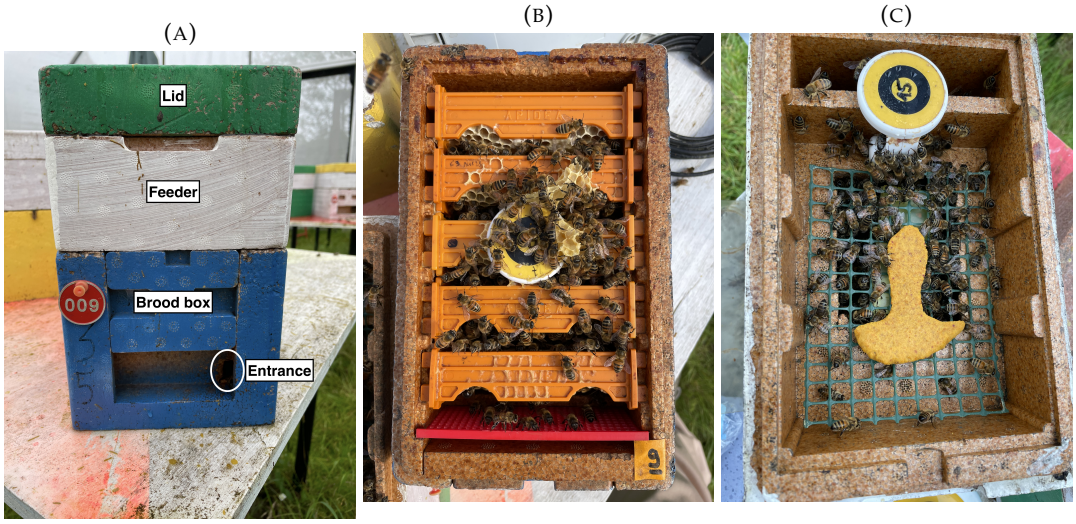


FIGURE 4.3.2: Configuration of experimental hives using APIDEA mating hives (A). Each hive consisted of a brood box with five frames (B), a top feeder with a hole cut in the bottom (C), an acrylic sheet (not pictured), and a lid. IHSs were placed on the middle frame (B) and on the dividing wall of the top feeder (C). Diet patties were suspended over the hole in the top feeder on a large mesh for optimal accessibility.

bucket. The nurse bees were provided with 70% invert syrup and placed in a cool, dark environment. This mixed any hive-specific scents, effectively removing each bee's hive-identity markers (orphanising).

The following day, orphanised nurse bees were distributed into the prepared hives. Each hive was given a mated queen (obtained from a breeder) and placed for 24hrs into a dark, air-conditioned shed to acclimate and form a colony. Colonies were provided with ad-lib 70% invert syrup in an internal feeder during acclimation.

Using the same orphanised nurse bees, pre-existing colonies were topped up to have about the same population size in all colonies. To improve acceptance of the added nurse bees, the hive, colony, and new nurses were sprayed with lemongrass-scented sugar syrup. After a 24hr acclimation period in the shed, newly-established colonies were moved into the polytunnel. Colonies were arranged in a 'Mercedes'-pattern with three hives to each table and hive entrances facing away from each other (Fig. 4.3.1).

4.3.4 Group allocation

The 32 colonies alive on D0 were assigned to treatment groups based on colony strength, capped brood production, queen presence and egg-laying, hive entrance orientation, and location within the polytunnel (see Tab. C.1.1 & 4.3.1).

4.3.5 Experiment schedule

I carried out my experiment between 20. July and 18. October 2023 (Fig. 4.3.3). Between 20. – 25. July, I topped up existing hives, set up new colonies, and installed sensors in all hives. IHSs were installed in three positions: on the centre frame ('frame'), in the top feeder ('top'), and on tables between hives ('ambient'). I began data collection with weekly full hive inspections from 28. July (see section 4.3.8). During this phase (feeding baseline), all colonies were fed a multifloral pollen diet for optimal nutrition (see section 4.3.6).

The starvation period began on 1. September (D0) for treatment groups FG1, FG2, and FG3. The colonies in these groups received candy patties (see section 4.3.6) for 1, 2, or 3 weeks. Following the starvation period, I recommenced pollen feeding. The beginning of the recovery period (T0) was staggered between groups. For FG1, T0 was on D7, for FG2 it fell on D14, and for FG3 it was on D21 (Tab. 4.3.2).

I performed weekly full hive inspections to record colony strength, brood production, and diet consumption, and replaced diet patties when necessary. From D0 on, I performed reduced "quick" inspections weekly, where I assessed colony strength and diet consumption (see section 4.3.8). Full inspections, including taking photos of all frames for precise brood counts, were performed on milestone days D0 (start of starvation period), T0 (start of recovery period), T14, T28, and T42 where possible (rows 'BRASS'[2] in Fig. 4.3.3).

4.3.6 Diet preparation

Colonies were fed with pollen patties (Feeding treatment/control) and candy patties (starvation treatment). Regardless of type, patties weighed 100g. Pollen patties consisted of 75–80% pollen and 20–25% 70% invert syrup (Tab. 4.3.3). Patties were mixed

[2] BRood ASSessment

TABLE 4.3.1: Allocation of hives into experimental groups. 17 hives were set up for this experiment (marked in 'New colony' column), 19 were taken over from a previous experiment. 4 hives died before D0 (unassigned in 'Group' column). 5 others died within 4 weeks of D0 and were excluded from analyses of census data and hive audio (unmarked in 'In analysis' column). No colonies survived more than 6 weeks post-D0 due to cold weather.

Hive	New colony	Group	In analysis	Alive in week
				-2 -1 0 1 2 3 4 5 6 7
1	X	1		
2	X	1	X	
3	X	C	X	
4	X			
5	X	1	X	
6		2		
7	X			
8		3	X	
9		C	X	
10		C	X	
11		1	X	
12		C	X	
13		3	X	
14		2	X	
15		2	X	
16		3	X	
17		3	X	
18		C	X	
19		1	X	
20		2	X	
21		2		
22	X	C	X	
23	X	2	X	
24		1	X	
25		C		
26		3		
27	X	3	X	
28	X	3	X	
29	X	2	X	
30		1	X	
31	X			
32		3	X	
33	X	1	X	
34	X	C	X	
35	X			
36	X	2	X	

TABLE 4.3.2: Mapping between notations of experiment time. D = days since start of starvation period. T = days since start of recovery period. D0 fell on the same date for all hives, but T0 differed between groups.

Day – Group	Control	FG1	FG2	FG3
D0	01.09. (D0)	01.09. (D0)	01.09. (D0)	01.09. (D0)
T0	01.09. (D0)	08.09. (D7)	15.09. (D14)	22.09. (D21)
T14	15.09. (D14)	22.09. (D21)	29.09. (D28)	06.10. (D35)
T28	29.09. (D28)	06.10. (D35)	13.10. (D42)	—

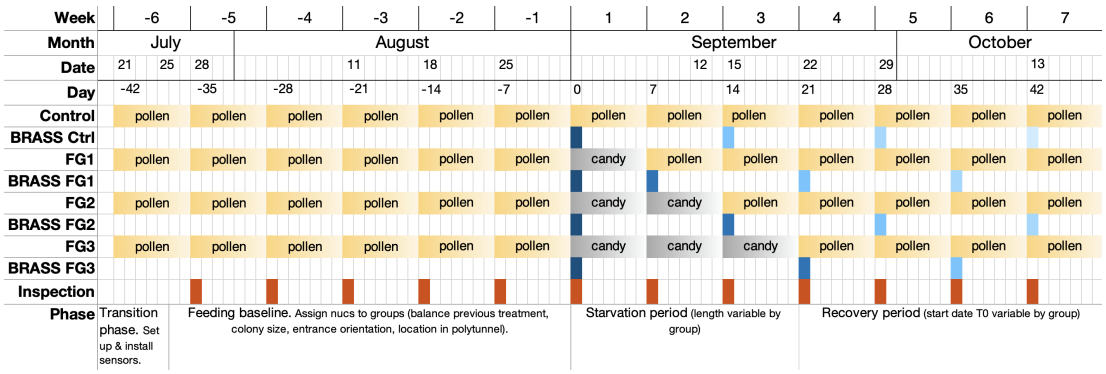


FIGURE 4.3.3: Experiment timeline with sampling and feeding schedules of 4 groups. The transition phase and feeding baseline (D-43 – D0) during which all colonies were fed pollen served to bring newly established and pre-existing colonies to an even and good nutritional condition. The starvation period during which colonies received candy patties differed in length from 0–3 weeks. Finally, colonies were fed pollen again during the recovery phase. The start date of the recovery phase was staggered. Each group's brood production was measured by photographing each frame on D0 (first day of starvation period, cobalt blue (darkest)), T0 (first day of recovery period, azure blue), T14 (14. day of recovery period, caribbean blue), and T28 (28. day of recovery period, sky blue (lightest)). For the control group, D0 was identical to T0. All colonies were inspected weekly to assess strength (beeseams) and diet consumption, as well as brood production prior to D0. From D0 on, brood production was assessed every two weeks (rows 'BRASS') while strength and consumption were assessed weekly (see section 4.3.8). During the starvation period, colonies were not inspected to reduce stressors.

in bulk in a professional-grade stand mixer[3]. The pollen mix consisted of pollen from a range of sources based on availability, including:

- monofloral French fresh frozen pollen from different plant groups (poppy (*Papaver sp.*), tree heather (*Erica sp.*), rockrose (*Cistus sp.*), cherry (*Prunus cerasus*), hawthorn (*Crataegus sp.*), and chestnut (*Castanea sp.*))
- Italian multifloral pollen, freeze dried and frozen
- Hungarian multifloral pollen, freeze dried and frozen

Each batch of pollen patties contained 250–500g of French poppy pollen, 250g of another French pollen, and 650–820g of multifloral pollen. The freeze dried pollen was ground using a spice grinder or coffee grinder on espresso setting, while fresh frozen pollen did not need to be ground. 15% of the pollen weight was added in invert syrup and everything was mixed in a stand mixer until homogeneous. More invert syrup was added if necessary to achieve a pliable and stable consistency. Each patty was measured out to approx. 100g and rolled out between sheets of parchment paper to 1cm thick. Patties were numbered, their exact weights recorded, and frozen at -20°C for future use (if not used the following day).

Candy patties were designed to be appetitive to the workers, while offering no nutrition beyond carbohydrates. The aim behind providing candy patties was to decouple patty provisioning from nutritional state. I wanted to provide a food source so that all colonies would eat throughout the week (likely producing associated noises), but only some would gain nutrition from the food consumed.

After colonies were found to fully consume patties made from pure beekeeper candy very quickly, I added maltodextrin as filler. A small amount of water was added to aid mixing. The maltodextrin content ranged from 20–25% and the added water content was 1%. Candy patties were mixed individually by hand as the candy was too dense to be mixed in bulk with our stand mixer. Each patty was weighed out to 100g, rolled out, weighed again, and stored at -20°C until use.

[3] KitchenAid Bowl-Lift 6.6L

TABLE 4.3.3: Example pollen and candy patty recipes. DI water = deionised water

Recipe	Ingredient	amount (g)	proportion of total recipe (%)
Pollen patty	Italian multifloral pollen	784.1	49.03
	French poppy pollen	247.5	15.48
	French *Cistus* pollen	244.8	15.31
	Pollen total	1276.4	79.82
	Invert syrup (70%)	322.77	20.18
Candy patty	Beekeeper candy	74	74
	Maltodextrin	25	25
	DI water	1	1

4.3.7 Chalkbrood

I discovered in early August that several colonies were infected with chalkbrood, a fungal brood disease which infects and kills larvae. Chalkbrood is spread by spores, which can be introduced to the hive by returning foragers, contaminated wax or hive components, or dirty beekeeping tools (Aronstein and Murray, 2010).

The infection was attributed to the persistently rainy conditions and unseasonally low temperatures. While initially I had planned to start the starvation period on 8. August, I decided to monitor infection levels and wait for warmer and drier weather to reduce infection levels. Going forward, all equipment and gloves were washed with chlorinated water after each inspection to avoid spreading spores between hives. Tables and hives were cleaned externally each week to reduce buildup of infectious material.

Since weather conditions continued to be changeable, I finally decided to start the starvation period on 1. September. To assess the disease's impact on brood production, I continued monitoring chalkbrood infection levels by counting the number of chalkbrood "mummies" in cells, in the bottom of the hive, and on the table at the hive entrance.

4.3.8 Hive inspections

I performed two types of hive inspections on a fortnightly rolling basis. A quick inspection was performed weekly on each hive. This involved:

- photographing the hive entrance and recording the number of chalkbrood mummies
- removing the diet patty, replacing the hive lid and weighing the hive
- weighing the diet patty to estimate consumption (and adding a new patty if less than 60g remained or the patty showed signs of spoiling)
- removing the top feeder, photographing the colony from above, and counting the number of beeseams to estimate colony strength

These quick inspections were kept as short as possible to avoid any unnecessary disturbance to the colony.

Every two weeks (and every week prior to D0), a brood assessment was carried out in addition to the quick inspection. This involved:

- finding the queen, if possible, and recording her mark colour and condition
- examining each frame to estimate the area of each brood stage and honey
- sampling 5 pupae and 5 nurse bees for tissue analysis
- From D0: photographing each frame to later accurately count the number of cells with each brood stage

4.3.9 Brood assessment

To assess the amount of each brood stage in the hive, I used both visual estimates and image-based counts. After verifying that the queen was not present on a frame, it was cleared of bees by a few shakes and taps over the hive. Any remaining bees were removed with a soft brush.

Visual estimates

For visual estimation, the frame was mentally divided into four quarters (two halves per frame side). I estimated the amount of each brood stage to the nearest quarter frame (1 half frame-side).

This method required some subjective rounding up or down. Where several frames had small amounts of one brood stage, which taken together would add up to 1 or

more frame quarters, these were recorded together. In turn, a quarter was discounted from a cell type which had previously been rounded up and was therefore slightly 'over-recorded'. Despite the subjectivity involved, I used my best judgement to ensure that the overall estimates of each brood stage and honey reflected the allocation of space in the hive as faithfully as possible.

The resulting recorded values expressed the number of frame sides with each cell type (honey, eggs, larvae, and capped brood) out of a total of 10 available frame sides (5 frames) in the hive. Beebread (stored pollen) was recorded when it was found, but this was very rare. I also recorded the number of larvae with chalkbrood infection inside cells, as well as the number of chalkbrood mummies in the bottom of the hive. While chalkbrood mummies in cells were counted exactly, mummies in the bottom of the hive were sometimes estimated rather than counted when there were a high number of them, since precise counting would have taken too long and negatively affected the colony.

Image-based counts

In addition to visual estimates, I photographed each frame side in order to accurately count the number of cells with each brood stage. Frame photos were taken with a custom-built solution principally developed by Rui Gonçalves. The solution involved a plexiglas "collar" with two integrated Arducam 64MP Hawkeye cameras connected to two Raspberry Pi microcontrollers (Pi 3 model B and Pi 4 model B), which operated a touchscreen and directed image capture (Fig. 4.3.4).

To take frame images, the collar was placed onto the open brood box and a cleared frame was placed into the frame holder in the middle between both cameras. The plexiglas collar helped to funnel bees from the frame back into the hive, preventing loss of hive bees when clearing the frame. Photos were captured, automatically labelled with the hive ID, frame number and side, and date, and saved to a microSD card.

Images were analysed in ImageJ v1.54f (Schneider, Rasband, and Eliceiri, 2012). Each cell was annotated based on its contents using the multi-point tool (Fig. 4.3.5). I

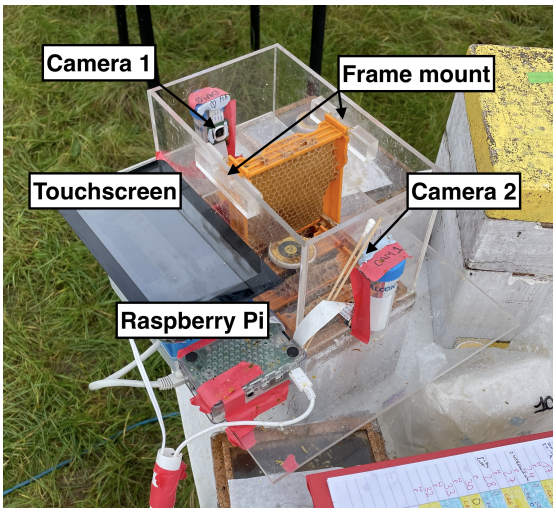


FIGURE 4.3.4: A custom frame imaging solution ("BroodCam") was used to take standardised photographs of frames. The system was chiefly developed by Rui Gonçalves.

distinguished eggs, larvae, capped brood, as well as problematic brood cells (chalkbrood mummies, dead larvae or pupae, cannibalised pupae, cells with multiple eggs, and damaged capped cells).

Due to the proximity of the cameras to the frame, not all parts of the frame were equally in focus. Additionally, the autofocus often focused on the cell rims, rather than the darker bottom of the cells. This meant that small larvae and eggs were sometimes hard to see. Similarly, while cells in the centre of the frame were clearly visible, on the periphery the cell bottoms were partially or fully obscured by the cell walls. Therefore it was much harder to detect eggs and small larvae in peripheral cells.

(A) Original image (B) Annotated image

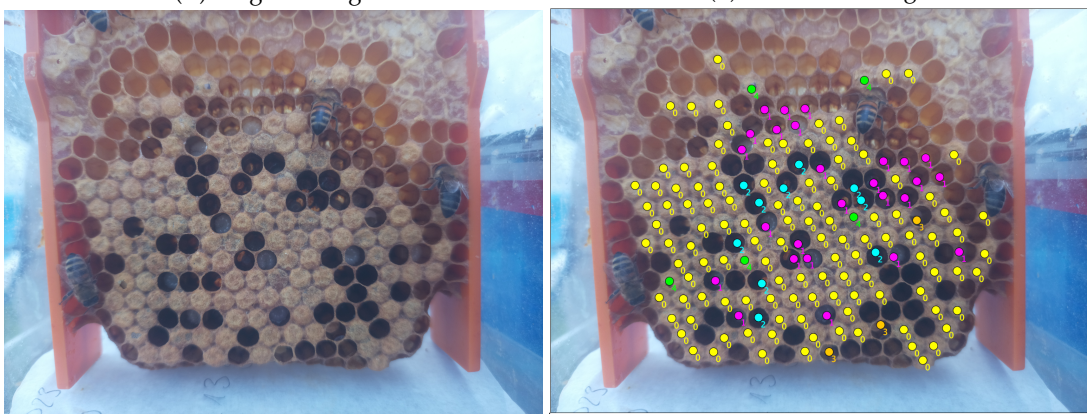


FIGURE 4.3.5: Frame photo taken with "BroodCam" (A) with cell content annotations (B). Cell types annotated: 0, yellow: capped brood; 1, magenta: larva; 2, blue: egg; 3, orange: chalkbrood mummy; 4, green: damaged capped brood cell.

4.3.10 Reliability of brood counts from images

To see if the issues with frame imaging affected brood counts, I assessed the reliability and comparability of my two brood metrics against each other using linear regression.

Correlation strength between the two metrics differed by brood stage (Fig. 4.3.6). For eggs, the correlation was moderate (R^2=60.8%). For larvae and capped brood, the frame and count metrics were strongly correlated (R^2 = 74.7% and 79.5%). The slope representing the number of cells equivalent to one frame side (unit of frame estimate) also differed strongly between brood stages. For eggs, one frame side was equivalent to 36.4 cells. For larvae, a frame side represented on average 87.2 cells, whereas for capped brood, one frame side was equivalent to 238.6 cells.

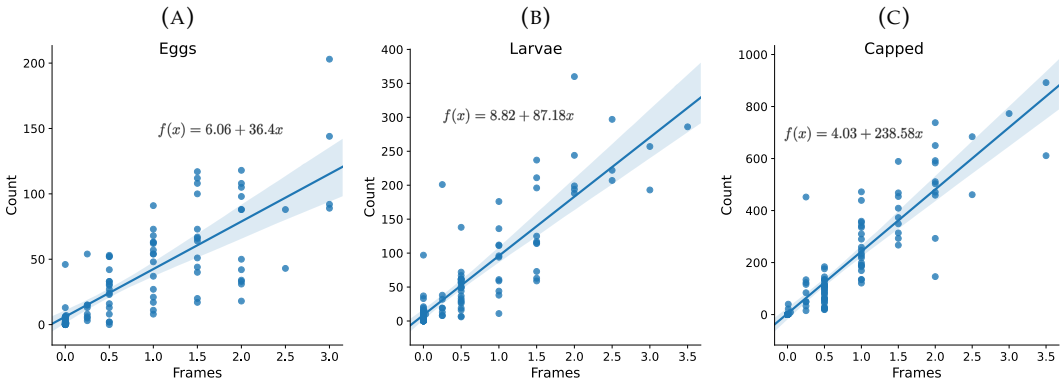


FIGURE 4.3.6: Regression of brood counts to frame estimates. For larvae and capped brood, both metrics are well-correlated (R^2 = 74.7% and 79.5%). For eggs, the two metrics are less correlated (R^2 = 60.8%). N = 101 observations (100 for capped brood) from 32 colonies[4].

4.4 Analytical methods

All data cleaning, analyses, and visualisations were performed in Python v.3.10.9 (Van Rossum and Drake, 2009). The main libraries used were `pandas v2.2.3` (The pandas development team, 2020), `numpy v1.26.4` (Harris et al., 2020), `librosa v0.10.2.post1` (McFee et al., 2024), `scipy v1.15.1` (Virtanen et al., 2020), and `sklearn v1.6.1` (Pedregosa et al., 2011) for data processing and feature extraction; `keras v3.6.0` (Chollet et al., 2015), `pytorch v2.3.1` (Paszke et al., 2019), and `tensorflow v2.17.0` (Abadi et al., 2015) for model training and testing; and `matplotlib v3.10.0` (Hunter, 2007) and `seaborn v0.13.2` (Waskom, 2021) for visualisation.

4.4.1 Feature extraction

Audio features

IHSs recorded a total of 52,015 audio samples from three sensor positions, including 20,125 recordings from the 'frame' position. I extracted a set of handcrafted audio features using the same approach as for Chapter 3. Due to the different audio recording settings (4s sampling at 8kHz sampling rate every 30 mins), a greater frequency range could be extracted from the recordings (up to 4kHz). I calculated the PSD of each sample using segment_length = 4096[5]. This yielded a native frequency resolution of ca. 2Hz. The frequency spectrum was resampled to 10Hz (by rounding frequency values to -1dp and calculating the mean PSD of all frequencies in the 10Hz bin) to reduce correlation between neighbouring frequency bins. Additionally, I truncated the spectrum to 0–500Hz, since higher frequencies were observed to be largely invariant. In addition to PSD features, I also extracted spectral centroids and ZCRs with window length 8000 and hop length 4000. I then calculated the mean and standard deviation of ZCR and spectral centroid within each sample.

[5] ca. 10% of sample length

In-hive microclimate

In-hive temperature and humidity were sampled every 10 minutes, collecting 1,014,726 records. Temperature and humidity samples were pre-filtered to exclude any obviously false records by removing any samples above or below set thresholds. Temperature readings below 0°C and humidity readings below 0 and above 100% RH were discarded. This resulted in the removal of 2386 temperature samples (0.24% of recorded values) but no humidity samples.

$$T = \begin{cases} T, & \text{if } 0 < T \\ \text{NA}, & \text{otherwise} \end{cases} \tag{4.1}$$

$$H = \begin{cases} H, & \text{if } 0 \leq H \leq 100 \\ \text{NA}, & \text{otherwise} \end{cases} \tag{4.2}$$

After this step, I applied a dynamic filter to exclude any spurious values which did not match the general trend of the temperature or humidity curves. This was done by calculating an expectation for each sample value based on the previous and following samples and then comparing the recorded value to the expectation. If the recorded value fell outside of a defined window around the expectation, it was rejected and replaced with the expected value.

$$T_{e,i} = \frac{T_{o,i-1} + T_{o,i+1}}{2} \tag{4.3}$$

$$\begin{cases} T_o, & \text{if } 0.7 \leq \dfrac{T_o}{T_e} \leq 1.3 \\ T_e, & \text{otherwise} \end{cases} \tag{4.4}$$

$$H_{e,i} = \frac{H_{o,i-1} + H_{o,i+1}}{2} \tag{4.5}$$

$$\begin{cases} H_o, & \text{if } 0.8 \leq \dfrac{H_o}{H_e} \leq 1.2 \\ H_e, & \text{otherwise} \end{cases} \tag{4.6}$$

where $T_{e,i}$ is the expected temperature for sample i, T_o is the observed temperature, and H_e and H_o the expected and observed humidity samples. The acceptance intervals were calibrated based on a pilot dataset which included unrealistic spikes

in temperature and humidity curves. The intervals $[0.7; 1.3]$ and $[0.8; 1.2]$ were effective at removing these spikes while keeping as much of the original data as possible. This step replaced 854 temperature samples and 957 humidity samples with predicted values (0.09% and 0.08% of recorded values). The final dataset contained 1,012,340 samples[6].

To every audio sample, I matched the nearest IHS record of temperature and humidity. The matching threshold was set to 60min.

Finally, I extracted the hour from each audio sample's timestamp. In Chapter 3, time of day was found to align with a main component of the UMAP embedding space, suggesting that in-hive sound varies on a daily cycle. Therefore I chose to incorporate time information as an additional feature for this analysis.

4.4.2 Label engineering

Nutritional state & Brood production

Nurse bees directly regulate brood production in response to the nutritional intake and stores of the colony. To reach the capped stage and complete metamorphosis, a larva must receive adequate nutrition from nurse bees producing brood food. When nutritional resources are low, young larvae (1–3 days old) are fed less frequently or even cannibalised to support rearing of older larvae to the capped stage. The loss of those larvae leads to a reduction in the number of capped brood later on. Eventually, nurses' internal reserves are depleted and no brood are able to survive until the capped stage. Therefore, brood production can be used as a proxy of colony nutritional state.

Seasonality

Brood production varies seasonally and declines with the onset of autumn. I constructed a label representing the reduction in brood due to poor nutrition while accounting for seasonality.

I calculated the average number of cells with each brood type across all control hives for each time step (fortnightly, see Tab. 4.3.2) to get a baseline of brood production in pollen-fed colonies throughout the experiment. The averaged brood counts

[6] including the samples replaced by predicted values

were interpolated using the PCHIP method to generate an approximate expected brood count for each day of the experiment. This baseline showed a seasonal decline in brood production due to cold temperatures and autumnal weather at my experiment site.

After interpolating brood counts for all treatment colonies, also using PCHIP, I subtracted the daily baseline from each colony's brood counts. This gave me the daily difference in brood production between each colony and the average pollen-fed control colony. I called this metric the "brood deviation".

Weighting

The age of a capped cell cannot be easily determined visually. Eggs hatch into larvae after three days and are capped after a further 5 days. Days 9–20 of development are spent in the capped stage, during which young bees pupate and finally emerge on day 21 (see Tab. 4.4.1). Nutrition is received only on days 4–8. The successful production of bee brood therefore depends on the nutritional state of the colony's nurse bees during this short time window of the developmental cycle.

The amount of capped brood at time t_0 represents the nutritional state of the colony during the preceding weeks when those now capped brood were being fed. However, the exact feeding windows of the capped brood, and therefore the time window over which colony nutritional state affected the capped brood, depends on the age distribution of the capped brood. This is linked to the queen's egg-laying rate. Neither the egg-laying rate nor the age distribution can be determined visually. I therefore assumed that the egg-laying rate was constant throughout the experiment[7].

Any larvae observed on day t_0 are between 4–8 days old. 4-day old larvae have been receiving food from nurses only since today (t_0), whereas 8-day old larvae have been fed for the past 5 days. The larvae count on day t_0 therefore represents the nutritional state of the colony over the period $[t_{-4}; t_0]$. Older larvae are more representative of the nutritional state than young larvae, since in total they required more feeding to survive until t_0 to be counted.

TABLE 4.4.1: Honeybee brood cycle.

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Stage	Egg			Larva					Capped												

[7]The expected seasonal decline in brood production was accommodated by normalising brood counts to the mean of the control group.

The long time window during which a population of capped cells on day t_0 received feeding is composed of the 5-day time window during which a given larva was fed, coupled with the range of days $[t_{-13}; t_{-1}]$ on which a given cell may have been capped. Any capped cells observed on day t_0 are 9–21 days old. Freshly capped, 9-day old brood were fed from $[t_{-5}; t_{-1}]$, while pupae about to emerge were fed for days $[t_{-17}; t_{-13}]$. Therefore, overall, the capped cells observed on day t_0 were fed from $[t_{-17}; t_{-1}]$.

Feeding windows overlap fully for capped brood cohorts of ages 13–17 days, but only partially for older and younger cohorts. Days with only partial overlap ($[t_{-17}; t_{-14}]$ and $[t_{-4}; t_{-1}]$) are less representative of nutritional state than days with full overlap, since fewer cohorts that were fed on those days are observed on t_0.

Among larvae, older individuals represent a greater investment by nurses and thus their numbers are more indicative of good nutritional state. Based on these assumptions, we can further weight each day's larvae and capped brood counts to account for their different contributions to the nutritional state metric.

Weighting was initially done using a linear function (Tab. C.3.1). Each further day of age indicates that there was sufficient nutrition available for nurses to 1) not cannibalise the larva and 2) feed her. However, larval growth does not follow a linear slope — instead, it more closely resembles an exponential function (Tab. 4.4.2, Fig. 4.4.1, Stabe (1929)). Using the growth curve by Stabe (1929), I derived a second set of weights by normalising daily larval weights with respect to the weight on day 5 (120hrs) (Tab. C.3.1). The difference between the brood deviation curves using linear weighting and growth-derived weighting is minimal (Fig. C.3.1). The greatest difference can be seen between 22.08.–01.09.2023 where the brood deviation increase in the control group and the decrease in other groups are slowed. While growth-derived weighting may be more biologically sound, the difference is small enough as to make no difference to the results. Therefore, it was judged unnecessary to repeat analyses with growth-derived weighting and subsequent graphs and model results were produced using linear weighting.

Formula

Using the known edge dates of honeybee development and assuming a constant egg laying rate (relative to control) we can refactor the relationship to derive a daily index

TABLE 4.4.2: Mean larval weight by day of development. Results by Straus (1911) and Nelson & Sturtevant (1924) are reported in Stabe (1929). The ranges given in Tab. IV of Stabe (1929) are assumed to be standard deviations. Day/Hour 0 weights were taken as soon as possible after hatching, and larvae are typically capped on day 5.

Day	Hour	Mean larval weight (mg)		
		Straus (1911)	Nelson & St. (1924)	Stabe (1929)
0	0	0.06		0.112
0	6			0.17
0	12			0.29
0	18			0.36
1	24	0.3	0.65	0.52
1	30			1.01
1	36			1.29
1	42			1.63
2	48	3.4	4.745	3.03 ± 0.36
2	54			5.87 ± 0.19
2	60			11.12 ± 0.48
2	66			18.3 ± 0.33
3	72	33.3	24.262	20.56 ± 0.40
3	78			25.92 ± 0.55
3	84			47.79 ± 1.88
3	90			66.76 ± 1.25
4	96	100.1	93.378	80.19 ± 1.98
4	102			115.62 ± 1.79
4	108			138.9 ± 0.79
4	114			159.06 ± 0.36
5	120	134.5	144.85	152.38 ± 0.42
5	126			145.57 ± 0.62
5	132			143.03 ± 0.65
5	138			141.83 ± 0.50
6	144	153.2	149.966	143.94 ± 0.51

of nurse bee nutritional state from future brood counts:

$$NSt(t_0) = \sum_{i=1}^{17} a_i B_C(t_i) + \sum_{i=0}^{4} b_i B_L(t_i) \tag{4.7}$$

where $NSt(t)$ is the nutritional state on day t, $B_C(t_i)$ is the observed or interpolated count of capped brood on day t_i, and B_L the count of larval brood. a_i and b_i are daily weighting factors for capped brood and larvae respectively. Egg counts were not considered when deriving this metric.

Implementation

To implement the weighted brood deviation metric, the interpolated brood deviation curves for capped brood and larvae were shifted back in time by the required number



FIGURE 4.4.1: Mean larval weights by day of development in Straus (1911), Nelson & Sturtevant (1924), and Stabe (1929). Results from all three sources are reported in Stabe (1929). Shaded area around Stabe (1929) data represents standard deviation.

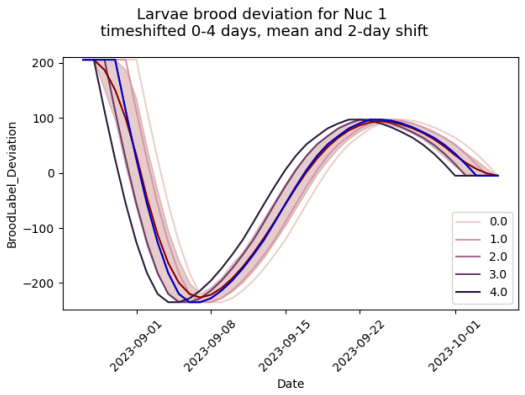
of days (for capped brood, 1–17 days, and for larval brood, 0–4 days). The first and last values of each curve were back-filled and forward-filled to avoid missing values. I then calculated the daily weighted mean of all curves within a brood type, using the weighting factors in Tab. C.3.1 (Figs. 4.4.2a & b). The daily mean larval and capped brood deviation were summed to give an overall estimate of the feeding provided on this day by nurses to the brood, serving as a general proxy for the colony's nutritional state.

The daily weighted mean brood deviation's unit was number of brood cells. I chose not to standardise this label between $[-1;1]$ so that it remained directly interpretable as "the number of brood produced less than or more than the average control colony".

Chalkbrood

Most colonies experienced some level of infection with chalkbrood. I measured the severity of the infection by the presence of 'mummies', the characteristic dead larvae, in comb cells, on the floor of the hive, and in front of the hive entrance. Each chalkbrood metric was interpolated to daily frequency. The daily total number of mummies

(A) Hive 1 — larvae



(B) Hive 1 — capped brood

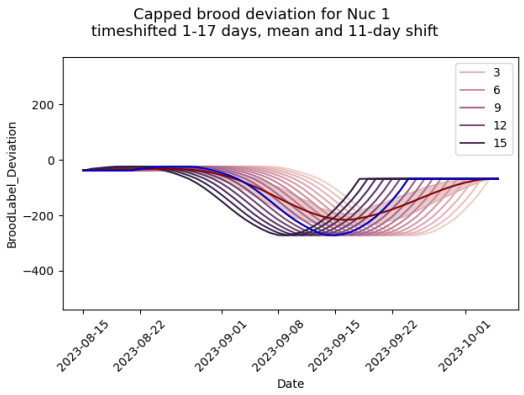


FIGURE 4.4.2: Timeshifted larva and capped brood counts for hive 1. Brood curves are coloured by the amount of shift (0–4 days for larvae, 1–17 days for capped brood). Blue: median, red: weighted mean across all timeshifted curves. The uncertainty envelope around the mean indicates the confidence interval.

was calculated and used to estimate the severity of the disease in the hive. The mean total of the control group was subtracted from each individual hive's total counts. I called this metric the total chalkbrood deviation.

4.4.3 Compiled dataset

The experiment recorded data on 32 colonies, but not all survived until the end of the data collection period. To avoid including samples from very small or dead colonies, I placed a cutoff date of D0 + 4 weeks (28.09.2023). 27 colonies survived to this point and were included in analyses of brood production and chalkbrood data. Unfortunately, two colonies did not have sufficient frame audio data available due to sensor malfunctions, leaving 25 colonies with adequate audio sampling. The final audio dataset contained data from 25 colonies over 8 weeks (04.08. – 28.09.2023) and included 12,587 audio samples matched with temperature and humidity records and interpolated hive census data.

4.4.4 Dataset splitting

Single-hive models

Single-hive models (SHMs) were trained using data from individual hives. I used a week-based split to obtain data subsets separated in time.

The data were partitioned by week and the order of weeks was randomised. The first week in the list was used as the validation week and the second week was used as the OOG test set. All other weeks were used for training. The assignment was then rotated until each week had been used once as the OOG test set (Fig. 4.4.3).

A pre-defined set of days, one per week, was used as a WG test set. Data from days within the training set matching the set of WG test days were removed from the training set and only used for WG testing. WG test days in the validation or OOG test sets were ignored for WG testing.

Multi-hive models

Multi-hive models (MHMs) were trained, validated, and tested on data from different hives with rotating set assignments (Fig. 4.4.4). To create 6 hive-stratified folds, colonies were divided into their experimental groups (control, feeding group (FG) 1, FG2, FG3) and their order was randomised within each group. For fold `i`, the `i`th colony from each group was assigned to the test set. The `i`-1th colony was assigned to the validation set, and all remaining hives were assigned to the training set. This was repeated until 6 folds were created. Because group 1 had 7 hives while all other groups had 6 hives, one hive from group 1 was always in the training set (hive 19 for colony strength label, hive 2 for brood production label).

Each set contained the same number of hives from each experiment group (excepting the training set which had 5 FG1 hives and 4 of each other group). Therefore the distribution of labels was approximately the same across all folds, and across all sets within each fold.



FIGURE 4.4.3: Example of train-validation-test split by week for a single-hive model.

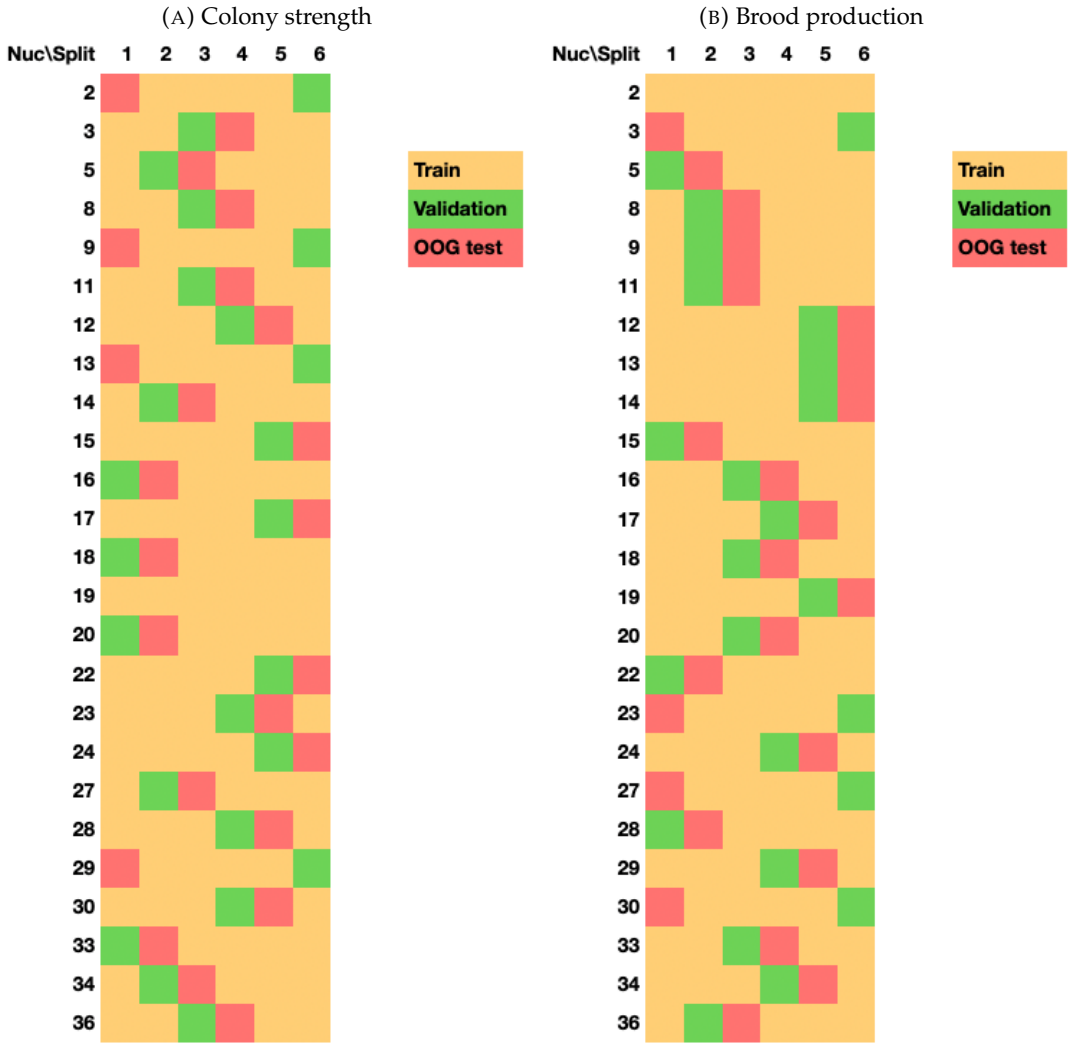


FIGURE 4.4.4: Train-validation-test split scheme stratified by hive and experiment group. Allocation of hives to test and validation sets differed between target labels (A: Colony strength, B: Brood production).

4.4.5 Model training

I trained single-hive models (SHMs) and multi-hive models (MHMs) to predict both the weighted mean brood production deviation (as a proxy for nutritional state) and the colony strength (number of beeseams). I included colony strength prediction after initial problems with finding a set of suitable hyperparameters for brood count prediction, because I assumed this to be an easier task. With colony strength as the target label, I was able to achieve overfitting to the training set. This demonstrated that the chosen neural network architecture was appropriate for the number of samples and features.

I also tested a model using Fourier Neural Operators (FNOs) and an FNO-inspired

MLP using spectrograms as inputs, but neither method was able to learn on the training data. I believe that these models were too complex for the size of the dataset. In the following, I only present results from MLPs using the handcrafted feature set described in section 4.4.1.

All models were trained in `Python` using `keras` with the `TensorFlow` backend. I used the Adam optimiser (Kingma and Ba, 2015) with learning rates 1e-2 – 1e-5 and batch size 32 (Tab. 4.4.3). SHMs were smaller than MHMs, with 1 hidden layer of 16 – 64 nodes compared to 3 hidden layers with 256 nodes. I reduced the network size for SHMs to avoid overfitting on the much smaller training sets.

4.4.6 Convergence monitoring

I used a validation set to monitor generalisation loss during training.

For SHMs, the validation set contained data from a single week. For MHMs, the validation set contained data from four hives, one from each experiment group, that were not included in the test or training set.

I used checkpointing to avoid overfitting. Validation loss was evaluated at each epoch and the checkpointed model was updated if the loss improved compared to the current checkpoint. The final model was also saved. During model evaluation, OOG and WG test loss (Mean Standard Error (MSE)) and R^2 were evaluated with the final model as well as the checkpointed model.

TABLE 4.4.3: Hyperparameters for nutritional state detection models.

Hyperparameter	Value	
	SHM	MHM
# hidden layers	1	3
hidden layer size	16 – 64	256
batch size	32	
learning rate	1e-2 – 1e-5	
optimiser	Adam	

4.5 Experiment results

4.5.1 Brood production

Brood counts were recorded at fortnightly intervals by photographing all frames and counting all visible eggs, larvae, and capped brood cells in ImageJ[8]. Counts were interpolated to daily frequency. I calculated the mean daily count of each stage in the control colonies to establish baselines of egg, larvae, and capped brood production which accounted for seasonality. I then subtracted this control mean from each hive's brood production time series to obtain a brood deviation metric showing how much more or less brood was produced compared to the expected seasonal baseline.

I found strong differences between treatment groups, as well as between all treatment groups and the control (Fig. 4.5.1).

Egg production gradually decreased in the treatment groups throughout the experiment period (top panel in Fig. 4.5.1). In control colonies, egg production increased following D0, peaking on T14 (mid-September), then dropped. At T28 (late September and early October depending on treatment group), egg counts were similar in all groups. In group 2, egg production increased from T14 to T28, whereas egg counts in the other groups decreased.

Larva production dropped in all treatment groups to T0 and T14 (early September in Fig. 4.5.2a) while increasing in the control group (middle panel in Fig. 4.5.1). In each group, the lowest larvae count was recorded on T0, the end of their starvation period (Group 1: D7, group 2: D14, group 3: D21). All groups increased their larva production again over two weeks following the end of the starvation period. While the total increase was different for each group (Group 1: +100, group 2: +75, group 3: +30), all recovered back to the control baseline, ie to a seasonally appropriate level of brood production (middle panel in Fig. 4.5.2b).

The number of capped brood decreased in all groups from D0 (bottom panel in Fig. 4.5.1). In the control group and group 1, capped brood counts stabilised at high levels (ca. 400 and 250 cells respectively on T0), whereas in groups 2 and 3, mean counts dropped to below 50 cells. Capped cell counts in groups 1 and 2 initially increased after feeding recommenced on T0 (D7 for group 1 and D14 for group 2), but

[8]Photos of frames were only taken during the treatment period from 01.09.2023 (D0) – 08.10.2023.

dropped again in synchrony with the control baseline (Fig. 4.5.2a). In group 3, capped brood counts recovered to a mean of ca. 80 cells by T28.

4.5.2 Chalkbrood

Chalkbrood severity was measured by counting the number of dead larvae with the disease ('chalkbrood mummies') in front of the hive, in the bottom of the hive, and in comb cells. All metrics were interpolated and summed to reflect the daily infection level.

The three chalkbrood metrics show different patterns in time (Fig. 4.5.3). Mummy counts at the entrance (top panel in Fig. 4.5.3) spiked in the first quarter of the experiment in all groups to different degrees (though all groups were pollen-fed during this time). Counts in treatment groups 1 and 2 remained higher than in the other groups until mid-September. After this, group 1 experienced a last spike before mummy counts decreased in all groups.

Mummy counts in the bottom of the hive (second panel in Fig. 4.5.3) spiked in group 3 in late August (prior to D0) and then remained very low throughout September with a slight upward trend in October. In the other groups, mummy counts increased from mid-September. The increase and count variance was especially marked in group 2.

In-cell mummy counts (third panel in Fig. 4.5.3) were variable throughout with spikes in all groups in mid-August and all but group 3 in late August. Counts steadily increased in all groups except group 3 throughout September and October.

There did not appear to be a link between experiment group and chalkbrood severity. The only exception is group FG3, which had very low mummy counts in all categories from mid-September onwards. Their brood production had been highly impacted by the starvation treatment, so there was less susceptible brood present. The total count of chalkbrood mummies in cells, in the bottom of the hive, and in front of the hive was not correlated with any brood production metrics (number of larvae, number of capped brood, and weighted brood deviation) or with colony strength (Fig. C.1.1).

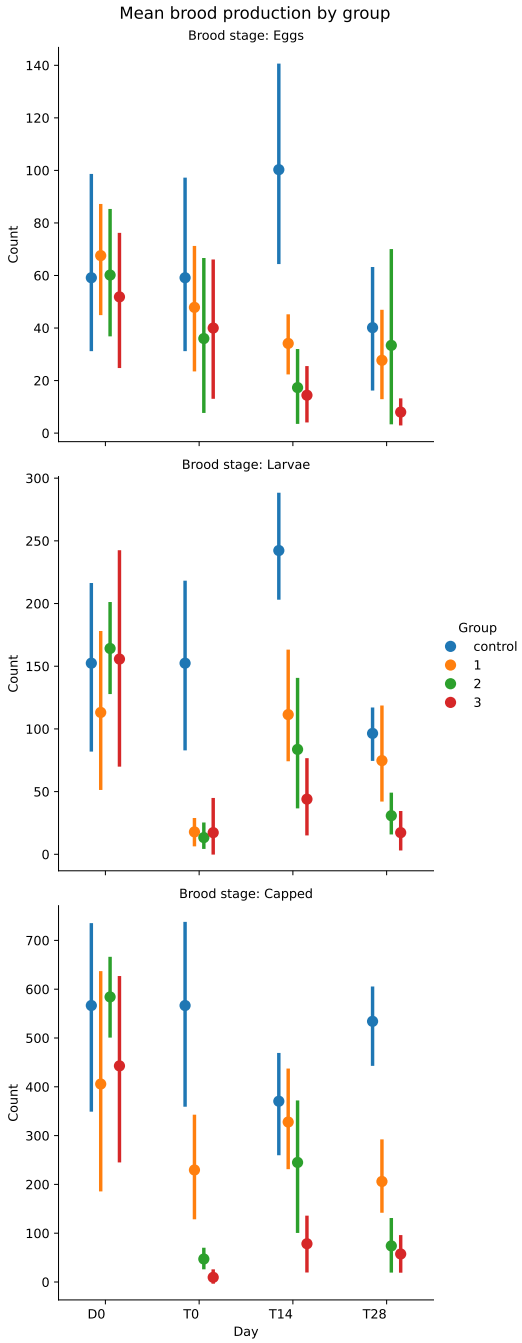


FIGURE 4.5.1: Mean brood counts by experiment group at each milestone. Error bars represent 95% confidence interval. N = 27 colonies. D0: first day of starvation period; T0: first day of recovery period; T14: 14 days into recovery period; T28: 28 days into recovery period. T0 was staggered between groups: For the control group, T0 = D0, for group 1, T0 = D7, for group 2, T0 = D14, and for group 3, T0 = D21.

(A) Interpolated brood counts.

(B) Difference to control mean.

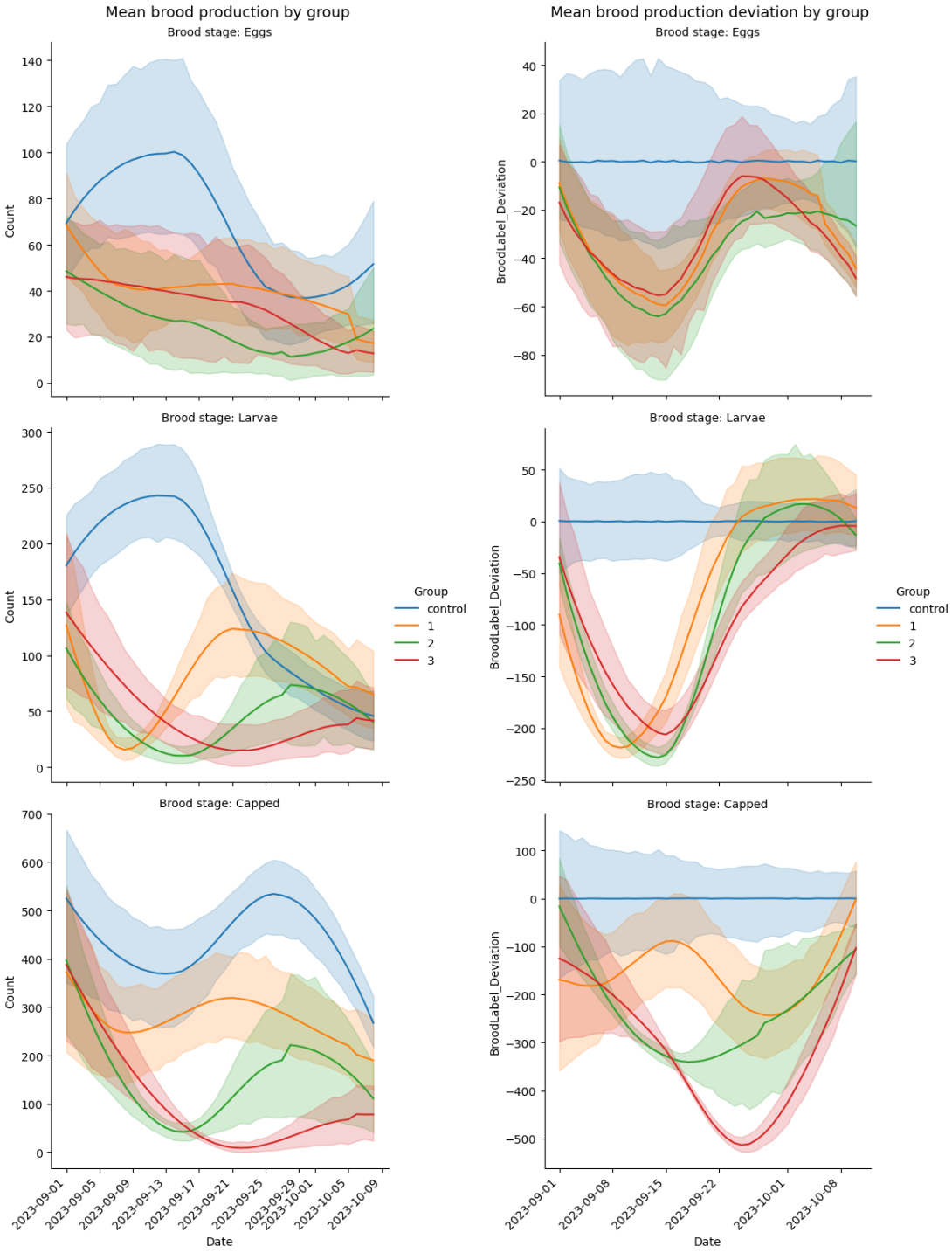


FIGURE 4.5.2: Interpolated brood counts in each brood stage (eggs, larvae, and capped brood) averaged by experiment group (A) and relative to control group mean (B). Brood production declined with length of starvation and recovered in all groups to seasonally appropriate levels. Uncertainty envelopes represent 95% confidence interval. N = 27 colonies.

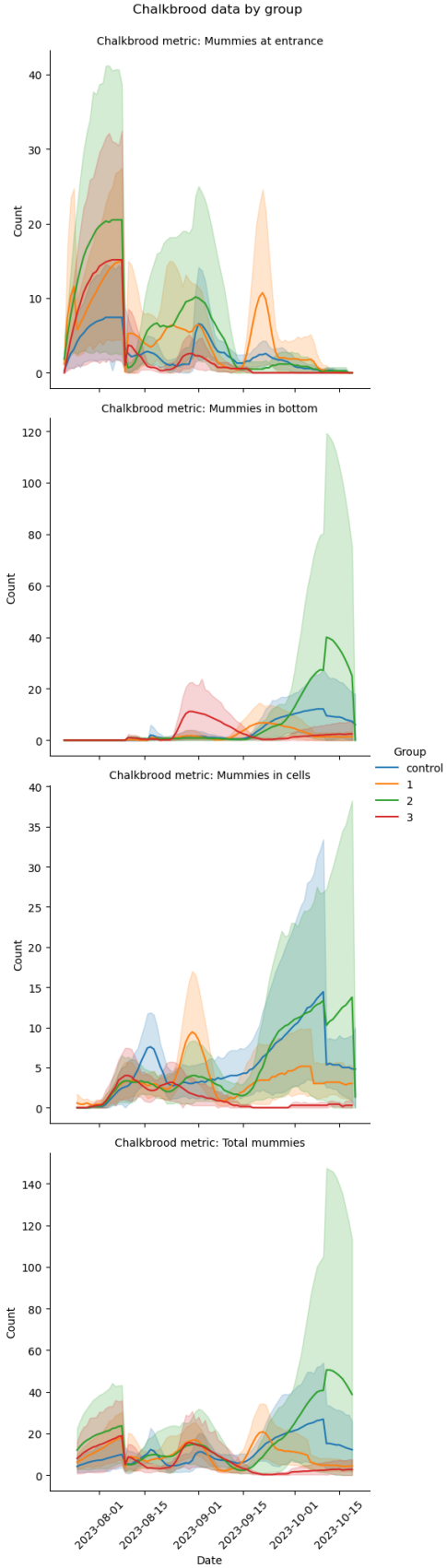


FIGURE 4.5.3: Mean chalkbrood severity by experiment group. Chalkbrood mummies were counted at the hive entrance, in the bottom of the hive, and in cells (top to 2nd from bottom). Total mummy counts are shown in the bottom graph. Length of pollen starvation did not affect chalkbrood severity. Uncertainty envelopes represent 95% confidence interval.

4.6 Modelling results

4.6.1 Soundscape differences by group

Just as spectrograms show changes in magnitude/PSD by frequency over time within a sample, we can use a similar approach to look at soundscape shifts across longer timescales by aggregating consecutive samples and plotting their average Power Spectral Density (PSD) over time. Then, instead of plotting the PSD itself, we can plot the contrast in Power Spectral Density between two groups. This shows the change in soundscape contrast over time — how different the soundscapes of two groups were initially, and how that difference evolved.

Differences between the control group and each treatment group were mainly concentrated in the range of 70–190Hz (Fig. 4.6.1, red dashed line = D0). The range was slightly wider for FG3 compared to the other treatment groups. Strikingly, there were differences in soundscape between groups even before D0. We observed some changes from pre- to post-D0 during midday (10:00–16:00, Fig. 4.6.1), but not at other times of day (Figs. C.2.1 – C.2.3).

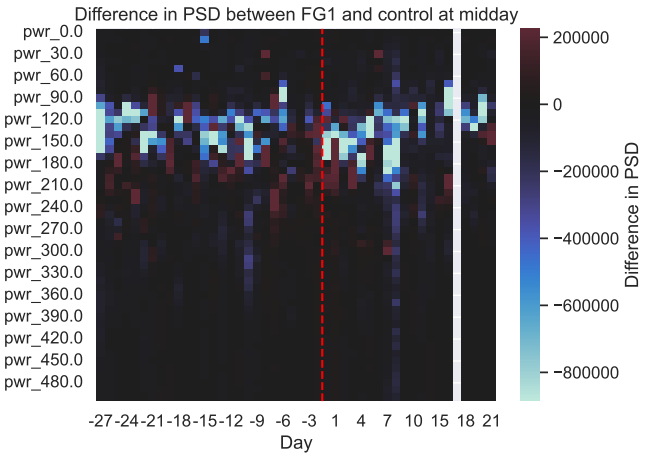
Across all three treatment groups, PSD in the frequency range of 120–180Hz decreased following D0. In particular, groups FG1 and FG2 had very similar patterns of PSD difference relative to the control. This was true even after D7, where FG1 received pollen feeding again while FG2 was still being given the non-nutritive candy replacement. The similarity of the post-D7 pattern of FG1 and FG2 suggests that these groups' soundscapes were not strongly influenced by their respective feeding states.

In contrast, colonies in FG3 were often louder between 90–200Hz than control colonies both pre-D0 and at the end of the experiment (Fig. 4.6.1c). In the first week post-D0, FG3 colonies tended to be quieter than control colonies, but not as much as the other treatment groups.

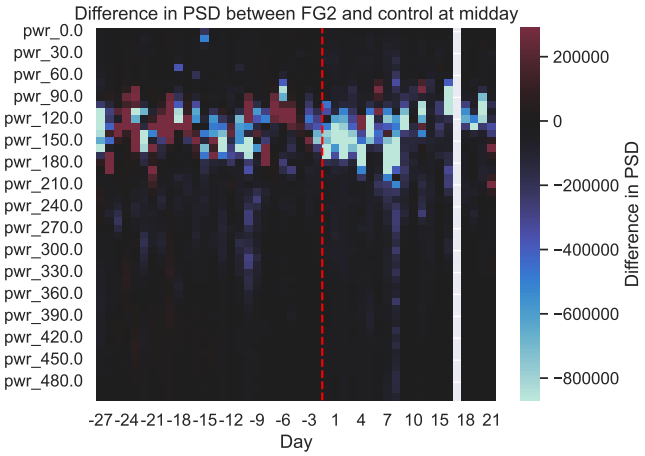
4.6.2 Model results

I trained ANNs to predict two target labels from a handcrafted feature set including the Power Spectral Density (PSD) from 0–500Hz, spectral centroid and ZCR mean and range, in-hive temperature and humidity, and time of day. Models included single-hive models (SHMs), which were trained and tested on unseen data from one single

(A) FG1



(B) FG2



(C) FG3

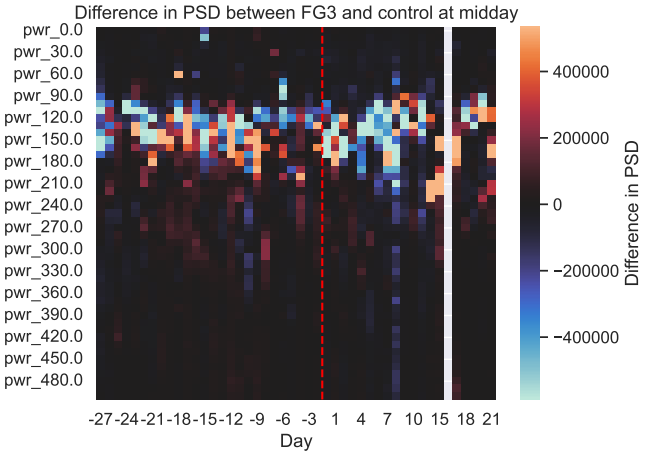


FIGURE 4.6.1: Soundscape contrast between each treatment group and control. Heatmap colour represents difference in PSD at each frequency and on each day at midday (10:00 – 15:59). High (red) difference values indicate that treatment group has higher PSD, while low (blue) values indicate that treatment group PSD is lower. Red dashed line indicates D0. D17 has missing data in control group, leading to missing difference data. N = 25 colonies.

hive, and multi-hive models (MHMs), which were trained on 21 hives and tested on 4 unseen hives. Models were trained to predict colony strength ("Beeseams") and colony nutritional state ("weighted brood deviation"). Both labels were regression labels (continuous counts), so the metrics used to assess model fit were MSE and R^2.

I retained a subset of all trained models to include only those model checkpoints (Fig. C.3.2) and learning rates (Fig. C.3.3) with the overall best performance. The final set of models used for plotting included 108 MHMs with learning rates 1e-3, 1e-4, and 1e-5, and 5,850 SHMs with learning rates 1e-2 and 1e-3.

Effect of model type and target label

Overall, my models struggled to make accurate predictions for either target label. SHMs had lower R^2 scores and higher MSE than MHMs both OOG and WG (Fig. 4.6.2). SHM OOG R^2 scores very rarely exceeded 0, showing no ability to generalise in time within a hive.

A small number of MHMs had positive R^2 scores when predicting the nutritional state label and MSEs were lower than for SHMs (Fig. 4.6.2, bottom panels). When predicting colony strength (top panels), MHMs performed much better than SHMs, with a median R^2 score exceeding 0.2 (though median MSE was similar between SHMs and MHMs).

Effect of treatment group and test hive group in MHMs

Multi-hive model performance varied between OOG test hive groups and between experiment groups. R^2 scores were very low for the control group for both target labels (Fig. 4.6.3). Models predicting the colony strength label had some OOG results above 0, with the best performance for groups FG2 and FG3 (Fig. 4.6.3a).

When predicting the nutritional state label, MHMs had low OOG R^2 scores for all groups. For FG2, WG performance was slightly higher than for the others, but the median WG R^2 score was 0 (Fig. 4.6.3b).

Model performance also varied between groups of test hives. Each group contained one hive from each experiment group. Therefore, the distribution of nutritional state labels was similar between test hive groups, but the distribution of colony strength labels may have been different.

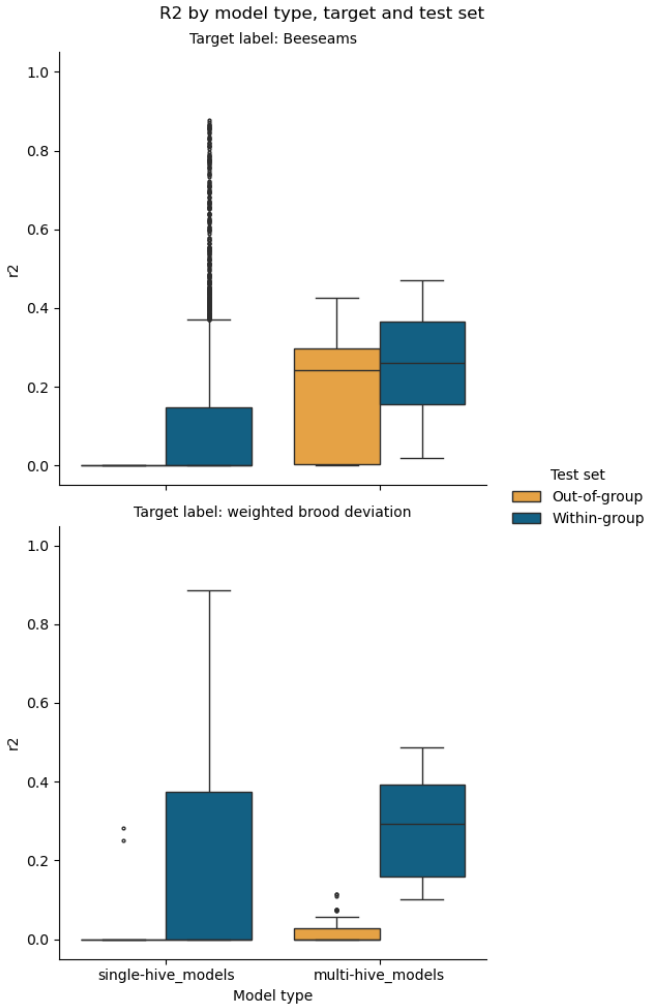


FIGURE 4.6.2: R^2 score of neural networks trained to predict colony strength (top panels) or weighted brood deviation (bottom panels) from IHS data including audio, temperature, humidity, and time of day. Results from 108 MHMs and 5,850 SHMs. MHMs were tested on 6 different fully unseen test sets with one hive from each treatment group, while SHMs were tested on one week of held-out data from the training hive. The held-out week was rotated until every week with available data had been tested on.
Central line represents median, box represents upper and lower quartiles, and whiskers represent 95% confidence interval. Points show outliers.

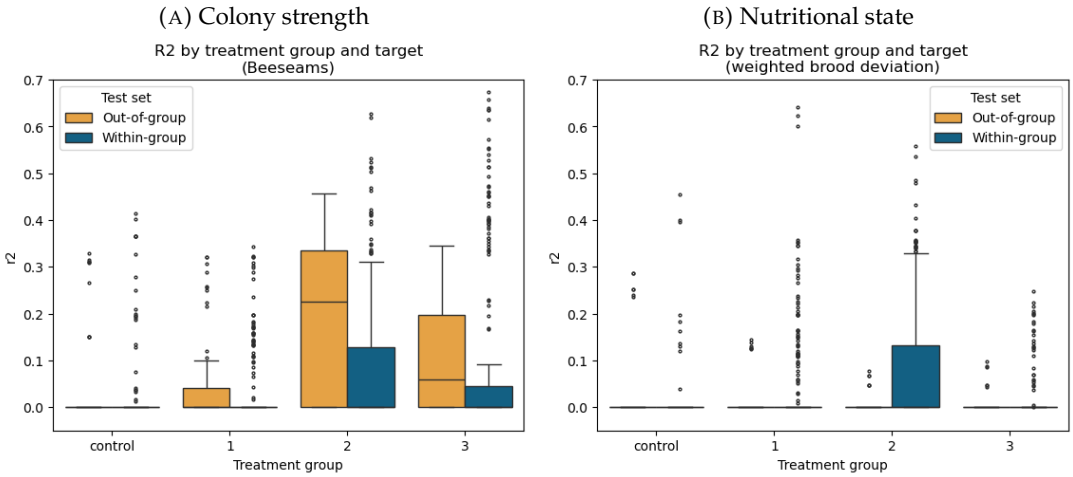


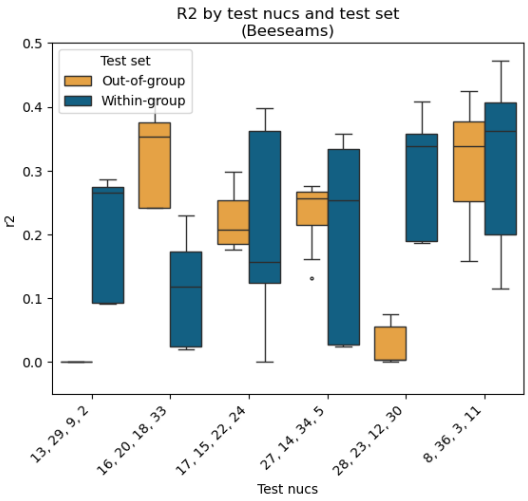
FIGURE 4.6.3: MHM prediction performance by experiment group for colony strength label ("Beeseams", A) and nutritional state label ("weighted brood deviation", B). Results were subset to include only models trained with α = 1e-3–1e-5, as these learning rates produced the best models overall. Model testing and graph specifications are identical to Fig. 4.6.2 above.

Prediction accuracy for the colony strength label was much more variable by test hive group than for the nutritional state label (Fig. 4.6.4). When predicting colony strength, median OOG R^2 scores ranged from 0 (test hives 2, 9, 13, 29) to 0.35 (test hives 16, 18, 20, 33 and 3, 8, 11, 36), with upper quartiles between 0.0–0.45 and lower quartiles between 0.0–0.25 (Fig. 4.6.4a). WG performance also varied strongly between test groups and was inconsistent with OOG performance. For example, of the two groups with the best OOG performance, one had the highest (3, 8, 11, 36) and one the lowest (16, 18, 20, 33) median WG R^2 score.

When predicting nutritional state, WG performance was more consistent between test hive groups compared to predicting colony strength (Fig. 4.6.4b). Median R^2 scores were between 0.2 and 0.35, with upper quartiles ranging from 0.35–0.5 and lower quartiles ranging from 0.05–0.15. Only two groups had positive OOG R^2 medians (5, 15, 22, 28 and 17, 24, 29, 34).

Differences in model performance between experiment groups and test groups can be further resolved into differences between test hives. Hives with a non-zero median OOG R^2 score were 8, 11, 15, 16, 17, 18, 20, 23, 24, & 36 for the colony strength label and 2, 11, 13, & 14 for the nutritional state label (Fig. C.3.4).

(A) Colony strength



(B) Nutritional state

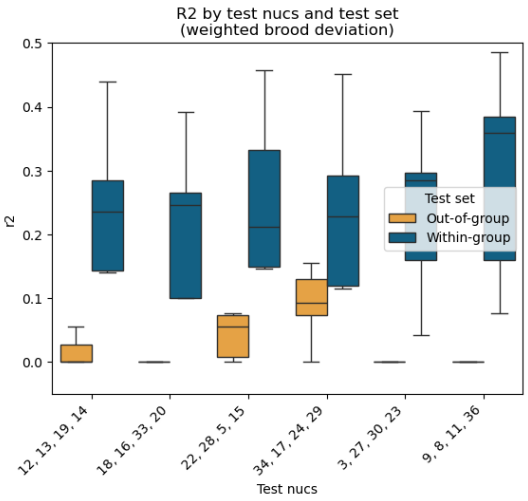


FIGURE 4.6.4: MHM prediction performance by unseen test hive group for colony strength label ("Beeseams", A) and nutritional state label ("weighted brood deviation", B). Model training and selection and graph specifications are identical to Fig. 4.6.3.

4.7 Discussion

4.7.1 Summary of results

Experimental

H1 Capped brood counts decreased during pollen dearth and increased with pollen feeding.

H2 I did not find clear evidence of a critical "point of no return", but observed decreased rearing capacity in all starved groups.

H3 While all groups recommenced brood rearing, peak rearing capacity decreased with the length of the pollen starvation period.

H4 There was no link between chalkbrood infection levels and length of starvation.

Modelling

H5 Handcrafted audio features, temperature, humidity, and time of day were not sufficient to predict nutritional state, but showed some utility for colony strength prediction.

H6 Cross-hive generalisation was only possible with multi-hive training sets.

H7 Model performance was lower for within-hive generalisation across time (forecasting, backcasting, and interpolating) in single-hive training sets than for cross-hive generalisation in multi-hive training sets.

H8 Model performance was higher when predicting colony strength than nutritional state (via brood production).

4.7.2 Brood production

The starvation treatment had a clear impact on brood production in all groups, consistent with H1. Production of larvae and capped brood dropped in all treatment groups (FG1 was starved for one week, FG2 for two weeks, FG3 for three) and increased again when pollen feeding recommenced. Larvae counts dropped close to 0 in all three treatment groups. Capped brood counts on the other hand dropped only slightly to ca. 250

for FG1 while for FG2 and FG3, capped brood counts decreased to nearly 0. This can be explained with the difference in "gestation" time at each brood stage (Tab. 4.4.1). Larvae become capped after 5 days, so any larvae counted on D0 were either dead or capped by D7 when the next brood count took place for FG1. The other treatment groups were only assessed on D14 and D21 respectively, to reduce disturbance during the sensitive starvation period. Capped brood on the other hand take 13 days from capping to emergence. Therefore, the majority of the capped brood cells from D0 still would have been present on D7 in FG1. However, due to the lack of larvae between D0–D7, many fewer new capped cells would have been added, leading to an overall decrease in capped brood as adult bees emerge.

FG1 experienced 7 days of pollen starvation and responded with a near-total loss of larvae by T0 (D7 in FG1). The number of eggs also decreased from D0 to T0 (D7) in FG1, but it did not drop as low as the larvae count. This suggests that the queen continued laying eggs, while the nurse bees were regulating the brood population, likely by brood cannibalism — consuming eggs and young larvae to recover nutrients for feeding older brood (Woyke, 1977; Weiss, 1984). Brood cannibalism strongly reduces honeybee brood survival rates during food insecurity, with only 50–75% of larvae surviving until capping during autumn (Woyke, 1977). Even short periods of pollen dearth (1–5 days) can decrease the amount of feeding received by young larvae (1–3 days old), while older larvae are kept well-fed until they are capped (Schmickl and Crailsheim, 2002). Since my experiment took place in late summer – early autumn, it is possible that workers were already primed for higher rates of cannibalism to conserve energy and prioritise rearing older brood.

On T14 (14 days after the period of pollen starvation ended), the number of larvae and capped brood increased back to the D0 baseline for FG1. This was not the case for FG2 and FG3, where mean brood counts were lower than on D0. This can be in part explained by the difference in dates: for FG1, T14 was on D21, whereas for FG3, T14 was two weeks later on D35, when conditions were gradually getting colder (Tab. 4.3.2).

I observed a drastic effect of pollen dearth reducing capped brood counts to nearly 0 on T0 in FG2 and FG3, the colonies that were deprived of pollen for 2 and 3 weeks. This would have led to a gap in emergence of young adult bees, the "nurses" which

perform brood care from ages 6–13 in a typical, healthy colony (Winston, 1987). Under normal conditions, hive bees then age into foraging from as early as 18 days old (Winston, 1987). However, honey and pollen starvation can lead to precocious foraging, whereby hive bees transition to foraging up to 14 days early, further reducing the number of nurses in the hive (Robinson et al., 1992; Schulz, Huang, and Robinson, 1998; Amdam and Omholt, 2003). Therefore, a substantial proportion of workers present in FG3 colonies on D0 already could have become foragers by the end of the starvation period. The combined effect would have been a strong reduction in the number of nurse bees in the hive, leading to a shortage of nurses available to care for any new brood that the queen is laying.

A small proportion of foragers can revert to nursing when all nurse bees are removed from the hive (Rösch, 1930; Page, 1992). However, reversal probability is dependent on the amount of time since the worker had made her initial transition to foraging, as well as her initial vitellogenin levels (Page, 1992; Amdam and Omholt, 2003). The quality of brood food provided by these reverted foragers also is doubtful, since foragers are nutrient-depleted compared to hive bees (Toth et al., 2005). Page (1992) observed reverted foragers interacting with brood, but did not test the nutritive quality of the brood food provided.

Overall, it seems likely that colonies in FG3 (and to a slightly lesser extent in FG2) were so depleted in workers capable of nursing that even 4 weeks of continuous pollen supplementation were unable to rescue the colonies. Colonies in FG1 on the other hand re-established capped brood production to the level of control colonies. However, the effect of the season cannot be underestimated, since even continuously fed control colonies radically decreased their brood production and declined with the onset of autumn. Repeating this experiment during early summer would help untangle seasonal effects from feeding-related effects.

My results are consistent with H1, showing that capped brood counts decrease during pollen starvation and increase again once pollen is present. I found no evidence of a clear "point of no return", and instead observed a gradual decrease in peak brood production capacity corresponding to the length of pollen starvation (H2, H3). Even the group with the longest starvation period was able to produce capped brood, and the least-starved group did not maintain brood production at the levels of the

continuously-fed control group.

4.7.3 Chalkbrood

Overall, I found no difference in infection level between treatment groups. Unexpectedly, chalkbrood severity was not lower in the control group compared to starved colonies (H4). Some isolated peaks only occurred in a single group, but these increases did not appear to reflect general patterns, since they were neither persistent in time nor consistent between groups with similar lengths of starvation. The only effect I observed was the strong reduction of chalkbrood in FG3 during and after the starvation period. Only very few larvae were present in these colonies after D0, leading to a negligible susceptible population.

The gap in brood rearing also did not alleviate chalkbrood severity later on, as the increase in mummies in FG1 and FG2 shows. This is consistent with chalkbrood spores persisting on in-hive surfaces and workers (Aronstein and Murray, 2010).

My observations contrast with recent work which observed differences in micronutrient content in larvae from infected and healthy colonies, suggesting that poor nutrition could increase chalkbrood susceptibility (Pavlovic et al., 2024). Pollen diet quality is also known to affect the resilience of honeybee workers to other diseases (Alaux et al., 2010; Di Pasquale et al., 2013).

4.7.4 Modelling

I trained ANNs to predict nutritional state or colony strength from handcrafted audio and in-hive microclimate features. I tested my SHMs' ability to generalise in time (backcast, forecast, or interpolate) within one hive, and my MHMs' ability to generalise to unseen hives.

Some SHMs showed good performance WG, with some models achieving an R^2 score of over 0.8. Nevertheless, these models were unable to generalise in time even within the same hive.

MHMs, trained on data from 17 hives, were slightly more able to generalise to unseen hives, though goodness-of-fit (R^2) was low overall. This aligned with my expectation that more varied training sets lead to improved performance on OOG data

(H6). It also showed that within-hive generalisation across time is not inherently easier than cross-hive generalisation (H7).

However, the training sets used for the two model types differed not only in the number of hives represented (1 vs. 17), but also in the total number of samples. On average, SHMs training sets contained only 150 examples for nutritional state prediction, and 300 examples for colony strength prediction. In contrast, MHMs were trained on ca. 4250 and 8551 examples for the respective target labels.

The checkpointed model (with the lowest validation loss) usually performed slightly worse at OOG generalisation, independent of model type. This was surprising since model optimisation on an unseen validation set is designed to preserve generalisability. In this analysis however, longer training often benefitted the model. Validation loss should reflect test loss if the data used in each set are equivalent — when losses are different, it suggests that the sets are too disparate. The apparent disconnect between validation and test loss points to fundamental differences in the soundscapes of different hives.

Generally speaking, more training data leads to better model performance (Liu et al., 2017; Althnian et al., 2021). This could easily explain the failure of most models (especially SHMs) to generalise. However, the poor performance of models trained on small datasets is usually due to overfitting, which I did not observe - otherwise, WG prediction accuracy would be higher.

Comparing my models' performance to baselines from literature, even models trained on small datasets (in the hundreds of recordings) were able to achieve high accuracy on random WG test sets (Robles-Guerrero et al., 2023; Maralit et al., 2023). But crucially, each individual recording in a random test set is adjacent to samples that are in the training set. In contrast, my WG test set contained full days of test samples.

Random test sets allow models to rely on temporal autocorrelation — samples close in time will sound similar to each other. Using separate WG test days interspersed between training days reduced the strength of temporal autocorrelation and required models to identify somewhat transferable patterns, even if only from one day to the next. This made my WG prediction tasks more difficult than the typical random test sets used in the literature, potentially explaining the decrease in model performance.

Colony strength and nutritional state prediction

I expected colony strength to have a clear effect on in-hive sound — more bees should be expected to make more noise. Preliminary work on a pilot dataset found a slight correlation between the mean PSD at 400-500Hz and colony strength (Fig. 4.7.1). I did not test whether the same correlation existed in this dataset. It is possible that the correlation between in-hive sound and colony strength is weak because each additional bee does not contribute much noise overall. It has previously been observed that background noise on a frame does not scale linearly with the number of bees on it, which has been attributed to mass damping (Smith and Chen, 2017).

While my MHMs were not excellent at predicting colony strength, they were able to explain up to 40% of change in this label using my handcrafted feature set. The average goodness-of-fit was 25%. In contrast, the same features, model architecture, and hyperparameters only achieved a goodness-of-fit of up to 10% for the nutritional state label, with a mean R^2 of 0.0.

Colony strength and nutritional state prediction from honeybee audio data have not yet been attempted in the literature. A link between colony strength and temperature in the centre of the hive has previously been demonstrated (Cook et al., 2022). This could have contributed to the improved model performance of colony strength prediction models. Lacking a true baseline for both prediction tasks, it is hard to assess whether the overall low rates of generalisation are due to an inherent difficulty in predicting these labels. Nevertheless, we can consider colony strength prediction an overall easier prediction target than brood production/nutritional state (H8).

Overall, my feature set appears to carry some signal about colony condition, at least colony strength. However, the features I chose do not seem to be sufficient to achieve high prediction accuracy on these tasks (H5). A different type of architecture, as well as a larger dataset, could improve model performance.

4.7.5 Limitations and further work

Experimental conditions

The 2023 field season was unseasonably wet and cold. This affected my experiment by increasing humidity in the polytunnel. The wooden tables I used had not been treated

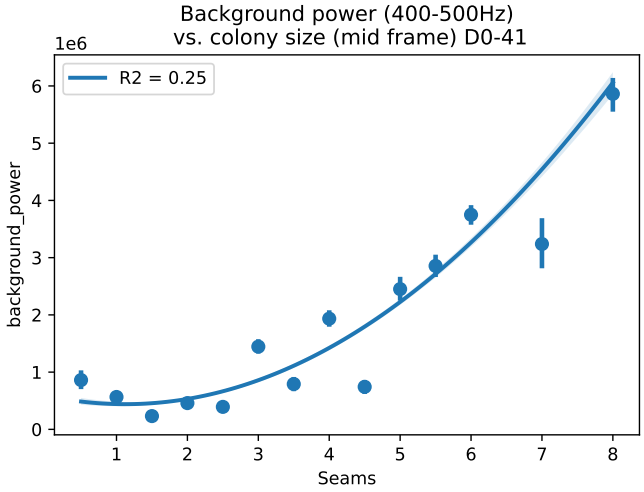


FIGURE 4.7.1: Correlation between PSD at 400–500Hz and colony strength ("Seams"). Data from 2021 pilot experiment (unpublished).

with water-resistant paint and were susceptible to mold under these conditions. I gradually repainted tables with bee-friendly, low-VOC (Volatile Organic Compound) paint[9]. While the new paint finish prevented mold on the tables, it did not decrease in-hive humidity. Colonies experienced significant chalkbrood loads as a result. This had not been an issue during pilot experiments which had been conducted in a sealed glasshouse and in an outdoor apiary on metal mesh tables.

I expected chalkbrood rates to decrease once the weather improved and delayed the experiment start by over three weeks waiting for sunnier weather. This pushed the recovery period into September and October, when honeybee colonies naturally begin to decrease brood production. Unfortunately, this meant that my three treatment groups did not go through their recovery period under identical conditions, limiting the conclusions we can draw from the experiment. Instead of staggering the end of the starvation period (T0) between experiment groups, I could have staggered the start. This would have meant that all colonies undergo their recovery period with pollen feeding simultaneously, increasing comparability between groups.

If the experiment had been able to start earlier, if I had used metal tables with better drainage, and if I had staggered the onset of starvation, rather than the onset of recovery, we might have seen lower rates of chalkbrood infection and avoided the seasonal downturn in brood production affecting my results.

[9]Protek Royal Exterior Finish

Dataset

The size of the audio dataset was likely a major factor limiting prediction accuracy. Despite having the in-hive sensor set to record every 30min, the total number of recordings obtained fell short of what I had expected. Sampling frequency dropped especially following D0, severely limiting the number of recordings during this sensitive experiment period where data were most valuable.

It is unclear what led to this loss of sampling. It is possible that sensors and gateways were damaged by the persistent wet conditions. In-hive sensors (IHSs) themselves were not protected and were exposed to atypically high humidity in the hives. These conditions could have led to malfunctions and loss of samples. Gateways were sealed in plastic bags with paper towels to absorb any condensation, so moisture damage to the gateway electronics themselves is unlikely. Gateways were battery-powered and a drop in power could have also led to lost samples. I replaced batteries in late September when I noticed the drop in sampling, but at this point the most sensitive experiment period had already passed.

In previous experiments, I was able to maintain high sampling density by recharging gateways with solar panels. In the polytunnel, this was not possible given the number of gateways used. In future experiments, solar power should be used over battery power where possible to remove this source of error.

Hive type

I conducted my experiment with nucleus colonies, which are 10–20x smaller than a typical honeybee colony and contain only a queen with ca. 2000 workers. The size and construction of my hives also differed from standard hives, being made from styrofoam rather than wood. I used nucleus colonies to reduce the number of workers required for the experiment. Smaller colonies also should be more sensitive to changes in nutrition, since the total nutrient buffer contained in nurse workers' internal reserves is smaller than in a more populous colony.

The low colony strength and small physical size of hives may have increased ambient sound contamination of the in-hive audio recordings. A concurrent ambient sound dataset was recorded, but not utilised in this analysis due to time constraints.

Future work on this dataset should evaluate the severity of ambient noise contamination and, if necessary, use the ambient sound recordings to decontaminate the in-hive audio samples.

Model choice and hyperparameterisation

I encountered some difficulties in achieving a good fit to the training data for many of my nutritional state prediction models. To validate the underlying architecture given the number of samples and features, I additionally predicted colony strength, where I was able to achieve overfitting. I then used the same basic ANN architecture and hyperparameters for both prediction tasks. However, a different set of hyperparameters or type of model might have been more suitable for predicting colony nutritional state.

Both target labels were largely interpolated between inspection dates. Colony strength was assessed weekly and brood production, the nutritional state proxy, was assessed weekly, then bi-weekly from D0 on. This naturally introduced uncertainty into the label, which can confuse supervised machine learning models. Weakly-supervised models on the other hand can deal with label uncertainty more readily. An example of this paradigm is MIL, where labels are applied to groups of samples rather than individual samples ("instances"). The group of samples ("bag") is considered to contain a mix of informative and uninformative samples. It is the model's task to learn to ignore uninformative instances and use only informative instances for the bag-level prediction. Given the weak nature of my labels, future analyses of this dataset should test whether multi-instance learning can improve prediction accuracy.

Recording brood production

Brood production was measured in two ways: by counting individual brood cells from photographs of frames, and by visually estimating the overall amount of each brood type on a frame. Brood photos were only taken every 2 weeks, since this was time-intensive and invasive for the colony. Prior to D0, I used visual estimation to record the total amount of each stage per hive to a precision of one half frame side.

Visual estimates and image-based counts for larvae and capped brood were well-correlated ($R_l^2 = 74.7\%$, $R_c^2 = 79.5\%$). The brood production estimates from before D0

could in future be used to extend the brood labels into this period of the experiment. This would significantly increase the number of training examples available to train my models and could lead to better performance.

The number of eggs and larvae that was equivalent to one frame side were much lower than for capped brood. This could suggest that visual brood assessments overestimated the amount of eggs, and to a lesser degree, larvae. Fewer eggs and larvae were 'required' to be seen to record a half frame side compared to capped cells. As described in section 4.3.9, brood amounts were sometimes combined across multiple frames when there were not enough of that brood stage present on one frame to record. This was done to reduce inspection duration, since these are strong stressors for bees and cause cooling of the colony (especially for small colonies like ours). Eggs and small larvae are also rarer than capped cells, since bee brood only spend 3 days as eggs and 5 days as larvae, compared to 13 days as capped brood. For example, multiple frames often contained small amounts of eggs, but less than a half frame side. In these cases, I recorded a half frame side across two frames. This method might have overestimated the true number of eggs and larvae present.

On the other hand, small brood were very difficult to identify on images. Eggs and young larvae are inconspicuous and easily confused with reflections of the wax in the cell bottom, especially when the camera focused at the cell cap level. Identification becomes even more difficult at the peripheries of the frame, where only a part or no part of the egg or larva might be visible. The inconsistency in the number of eggs, larvae, or capped brood equivalent to a frame side could reflect undercounting of these smaller brood stages. Whereas eggs and larvae can be found in all areas of the frame during inspections, later counting from images would typically only detect those in the centre of the frame and likely left a significant proportion unrecorded.

Despite the inconsistencies between brood stages regarding how many cells one frame side represented, the high correlation coefficients for larvae and capped brood show that for these brood stages, my visual assessment method was consistent between hives and over time. Whether or not the exact brood counts are accurate, they can be used to evaluate hives' brood production in relation to each other and throughout the experiment.

4.7.6 Conclusion

Pollen dearth strongly impacts a honeybee colony's ability to rear brood and maintain its size. "Rescue feeding" with supplemental pollen can help restart brood production after a period of pollen starvation and avoid colony decline. However, it is not fully known how long a colony can be without pollen and still recover. Further, there are currently no methods for remote monitoring of nutritional state.

In this chapter, I evaluated the ability of honeybee colonies to recommence brood rearing after a period of full pollen starvation when fed a multifloral pollen diet. I found that all colonies, even those starved for 3 weeks, restarted brood production, but the peak brood rearing capacity decreased gradually with increasing length of starvation. Counter to expectation, the length of the starvation period did not affect chalkbrood levels.

Based on known development times of honeybee brood, I derived an indicator metric of colony nutritional state. My metric calculates a weighted average of interpolated larvae and capped brood counts over several days and weeks.

Audio and in-hive microclimate data captured during the experiment, as well as time of day, were used to train ANNs to predict colony strength and my nutritional state metric. I found that this feature set contained information about colony strength, but not about nutritional state. Model performance increased with training set size and hive diversity. Within-hive forecasting, backcasting, and interpolation of colony strength or nutritional state was not possible, likely due to insufficient training set size. In contrast, I achieved cross-hive generalisation of colony strength prediction, but model performance was highly dependent on the test hive.

My results indicate that nutritional state prediction based on audio and microclimate features is fundamentally more difficult than colony strength prediction. Nutritional state prediction was not possible in my dataset with the chosen model architecture. I identified several confounding factors in my experiment, chiefly the seasonal downturn in brood rearing coinciding with the last weeks of data collection, which complicated the interpretation of the experimental results. However, there is scope for improving the nutritional state label and doubling the dataset size. These improvements in combination with a model type more suited to "weak" labels, such as

multi-instance learning, could in future improve model performance for cross-hive nutritional state prediction.

Chapter 5

General Discussion

5.1 Summary of key findings

Despite recent advances in hive monitoring, fundamental questions remain unanswered. These include:

- What factors influence the performance of colony state prediction models?
- What constitutes best practice in hive monitoring, including experiment design, data collection, and model testing?

By filling these knowledge gaps, we can focus our efforts to generate better datasets and make progress towards truly generalisable models with stable performance across contexts. To this end, I analysed two novel datasets, collected by me and Canetis SRL, of in-hive sound and microclimate data from honeybee colonies experiencing queen loss, nutritional stress, and chalkbrood disease.

In Chapter 2, I tested how dataset characteristics and feature extraction affected the performance of acoustic queen loss detection models. I found that prediction accuracy was sensitive to dataset size and diversity, sensor position, and feature set. Handcrafted audio features (**pds!**, spectral centroid, and ZCR) proved more suitable than ResNet encodings of spectrograms, likely because they were less noisy. Compared to models trained on a single hive (SHMs), models trained on data from multiple hives (MHMs) performed worse when tested on unseen data from training hives (ca. 85% vs 95% WG test AUC), but had a higher capacity to generalise to unseen hives (ca. 70% vs 60% OOG test AUC). Contrary to expectation, sensor position did not significantly affect model performance. However, when accounting for hive identity, sensor

position strongly impacted performance, with as much as 30 percentage points difference between positions. Hive identity further interacted with feature set and sensor position, affecting model performance unpredictably.

In Chapter 3, I investigated the underlying causes of hive-dependent differences in model performance. I found that in-hive soundscapes were most strongly shaped by time of day, hive identity, hive position, colony size, honey stores, and queen state. However, the effect of queen state on the soundscape differed strongly by hive, making it difficult to define a region of the audio feature space broadly indicative of queenlessness. Simple operations such as thresholding predictions or subsetting by time of day improved model accuracy, but their utility differed between hives. Using mean feature importance values from a set of 24 models with high OOG AUC, we discovered that low frequencies were disproportionally relevant for detecting queenlessness, along with frequencies linked to waggle dancing and flight. These soundscape changes could indicate increased foraging after queen loss combined with a decrease in in-hive activity and inactive workers.

In Chapter 4, I applied my colony state prediction pipeline from Chapter 2 to a novel dataset to test the effect of task characteristics on model performance. I found that brood rearing capacity decreased with length of pollen starvation. However, even colonies deprived of pollen for 3 weeks recommenced brood rearing when pollen feeding was reinstated. Chalkbrood infection did not interact with pollen starvation to affect brood rearing. Models trained to predict colony nutritional state or colony strength from combined in-hive audio, microclimate data, and time of day performed substantially worse than expected. Multi-hive models showed very limited ability to make predictions on unseen hives. Single-hive models trained on parts of the data from one hive were unable to extrapolate to future or past unseen data. Colony strength prediction models performed better than nutritional state prediction models, indicating that the chosen feature set contained more signal for this prediction target.

In the following, I synthesise the results from my data chapters to identify likely determinants of model performance and possible avenues towards more commercially relevant models. In section 5.2, I discuss the effects of test set choice, prediction

label, task type, and dataset characteristics on the accuracy of colony state prediction models. In section 5.3, I draw on my experiences in creating and analysing these datasets across all three data chapters to suggest a set of best practices covering experiment design, data capture, model design, and model testing.

5.2 What factors influence the performance of colony state prediction models?

In Chapter 2, models performed very well when tested on a fully random test set taken from training colonies (WG test set). This was explained by high levels of autocorrelation within our audio time series, which can increase a models' ability to make predictions on unseen data if those data are similar to previously seen data — eg, a test recording taken in the same hive half an hour after a training recording.

In Chapter 4, I used a non-random WG test set to reduce temporal autocorrelation while maintaining the "same hive" advantage. Instead of choosing random samples for the WG test, I chose a set of WG test days. Data from these days was held out from training and only used for testing, but came from training hives.

Models for colony strength and nutritional state prediction performed better on these WG test sets than on unseen hive OOG test sets, but performance was not as good as I expected given the 90+% AUCs seen in queen state prediction models. Similarly, OOG performance was much worse for colony strength and nutritional state models compared to queen state models.

There are multiple possible reasons for this stark contrast in performance.

5.2.1 Test set

The switch from fully random test set to WG test days largely prevented autocorrelation between training set and test set samples. With random testing, most test samples were immediately preceded and followed by training samples with only 30min between samples. Samples from WG test days were as much as 12h removed from any training samples. The residual autocorrelation apparently did not confer a strong advantage to models.

5.2.2 Prediction label

My models from Chapters 2 and 4 were attempting to predict very different aspects of honeybee biology. Queenlessness is anecdotally known to affect colony sound strongly and is a frequent prediction target for audio-based hive monitoring studies. Colony strength has, to my knowledge, only twice been predicted from in-hive sound. One model classified hives as weak or strong with 75% accuracy (Luz et al., 2024), the other with 66% (Zhu et al., 2024). However, neither analysis used a hive-stratified test set, suggesting that the true OOG accuracy of the models would be lower and potentially comparable to the goodness-of-fit found in my own colony strength prediction models (OOG R^2 of up to 45%).

No hive monitoring studies to date have attempted to predict nutritional state. My own models for nutritional state prediction were largely unable to fit the data and make correct predictions. Lacking a baseline from the literature, it is hard to diagnose if the failure to learn was caused by the dataset, the model architecture, the chosen feature set, or a combination of factors. However, the relative success in colony strength prediction using the same features, data, and architecture suggests that nutritional strength prediction is inherently harder than the other prediction tasks I attempted.

5.2.3 Task type

The prediction tasks not only covered different colony conditions, but the tasks themselves were formulated in fundamentally different ways. While the models for queen state prediction performed binary classification, the models for nutritional state and colony strength prediction were predicting regression labels.

For binary classification tasks, models usually have a wide margin of error: with a standard threshold of 0.5, models can output anywhere from 0–0.49 to predict '0', or between 0.5–1 to predict '1'. Regression models on the other hand must output a value as close as possible to the true label. Model accuracy decreases the further the prediction value deviates from the ground truth. Baselines for these metrics also differ — the goodness of fit (R^2) of a random regression model is 0, whereas a random binary classifier has a baseline AUC of 50%.

Reformulating regression labels into multi-class or binary labels likely would improve model performance. For example, brood production could be turned into a binary label by classifying brood production as '0: seasonally appropriate', or '1: significantly lower than expected for the season'. Other studies attempting prediction of binarised colony strength labels support the notion that simplifying the target label can improve model performance (Zhu et al., 2024; Luz et al., 2024).

5.2.4 Datasets

The datasets tracking queen loss and pollen dearth differed in a few potentially critical ways that may have affected model performance. They were collected with differently-sized colonies, audio was recorded with slightly different settings (though using the same sensors), there were many fewer samples in total (180,000 vs. 8,000), and features were extracted at a lower resolution.

The queen loss dataset (Chapters 2–3) was collected with full colonies in normal-size hives, while the nutritional state dataset (Chapter 4) was collected with mini-colonies in mating nuc-type hives. On the one hand, smaller colonies with fewer bees are expected to respond to interventions quickly and therefore lead to fast and clear changes in the in-hive soundscape. On the other hand, mini-colonies might lack the critical mass of bees required to drown out background noise and produce a consistent soundscape in the first place. The physical size of the hive also likely played a role in reducing ambient noise: an IHS in the centre of a normal-size hive was further away from the hive walls and surrounded by more bees than in a mini-hive.

Audio recordings in the Chapter 4 dataset were recorded at a sampling rate of 8kHz and were 4s long. This high sampling rate allowed me to measure frequencies up to 4kHz[1], but examination of **pds!** from 0–4kHz showed very little activity above 500Hz. In keeping with Chapters 2–3, I therefore opted to truncate spectra at 500Hz. Contrary to previous chapters, where spectra were used as features at their native resolution ($\Delta \approx 3.9$Hz), I instead resampled the spectra to 10Hz resolution to reduce correlation between neighbouring Power Spectral Densities. This reduced the number of PSD features from 129 to 50 and therefore the amount of detail in the frequency features.

[1] following the Nyquist rule (Landau, 1967)

Overall, differences in model performance between chapters can be due to a range of factors. I estimate that the poorer performance of nutritional state and colony strength prediction models was mainly down to the task type (regression vs binary classification) and dataset size.

5.3 What constitutes best practice in hive monitoring?

5.3.1 Experiment design

Any model is only as good as the underlying data. Experiment design therefore needs to take into account human error and bias. The collection of monitoring data, eg in-hive sound and microclimate data, can be automated to avoid human error. Phenotypical measurements such as colony strength, brood counts, and disease levels, which require beekeeper assessment, must be recorded as objectively as possible. This also includes recording metrics for evaluating error, such as tracking who performed an inspection, and any uncontrolled factors that could impact colony health and behaviour, for example forage availability and weather conditions.

While maximising the number of hives in an experiment should be prioritised, the assemblage of hives should not include "singletons" — hives with any unique conditions. This is because any change in colony sound could now be due to this condition rather than the condition we intended to study, effectively removing the singleton hive from the population for statistical purposes.

Experimental datasets form the basis of most of the hive monitoring literature. This type of dataset is acquired as part of a controlled experiment, where colonies are manipulated to bring about the desired colony states while controlling other aspects of colony health and the surrounding conditions as much as possible. This can remove confounding factors, such as weather conditions, allowing the analysis to pin down changes in in-hive microclimate or sound resulting from interventions and treatments. Nevertheless, simultaneous interventions across all colonies can also be a disadvantage: On such datasets we can easily demonstrate extrapolation across colonies, but not across time.

Experiments can also introduce other confounding factors by placing honeybee colonies under unnatural conditions or performing manipulations that disrupt colony

homeostasis. In my experiments, examples are the confinement of queens to cages prior to removal (Chapter 2–3) and the placement of colonies into a semi-field environment with internal pollen provisioning, rather than free pollen foraging (Chapter 4).

A second type of data could be called phenomenological datasets. Recently, two new datasets of this type tracking audio, temperature, humidity, and colony phenology were released (Zhu et al., 2024; Abdollahi et al., 2025). The datasets have large numbers of hives and follow them over long periods of time (up to a year). The sensors are present in hives for extended periods of time, allowing for stable integration into the matrix of the hive and acceptance by the bees.

Experimental datasets are easier to analyse, since treatments are clearly defined, control colonies without interventions are present, and the environment is controlled. However, extrapolation of predictions outside of the experiment cannot be demonstrated. Phenomenological datasets in turn are less suitable for targeted analysis of a single aspect of colony health, but they can be valuable for the validation of models under real-world conditions.

The effect of hive dimensions on in-hive sound was demonstrated in Chapter 2. Any work on audio-based hive monitoring therefore should take into account hive size (number of brood boxes/honey supers) and material when comparing data between hives, since differences in resonant properties can strongly alter a colony's **pds!**. This poses a problem in the collection of phenomenological datasets. Even if all colonies start with the same hive configuration, differences in growth and productivity over time will mean that good beekeeping practice requires the addition of supers or brood boxes. Audio recordings from before and after the addition are now no longer comparable, as are recordings from hives with different configurations. Not adjusting the hive configuration to suit the colony is not an answer, since leaving a colony without adequate space would introduce stress. Regular replacement of filled frames with empty comb would alleviate pressure for space, but also artificially keep the colony small and defeat the purpose of monitoring the colony under natural conditions.

5.3.2 Data capture

Honeybee vibrational communication is complex and highly varied. Harmonics as high as 5.5kHz have been observed (Pratt et al., 1996), but most fundamental frequencies fall below 500Hz. My results (Chapter 3) also showed that sounds below 100Hz are impactful, at least in the detection of queenlessness. Therefore, microphones should show a flat frequency response in low frequencies. Though I was not able to test them as part of this work, low-frequency microphones with ranges down to 5Hz would be ideal. The sampling rate should be set to 1kHz or higher so that all frequencies up to 500Hz are captured. Increasing the sampling rate would allow capturing harmonics at the expense of increasing storage demands. This can be a major limiting factor when recordings are stored on edge devices or uploaded to cloud storage.

When setting sampling schedules, even coverage throughout the day should be prioritised over long continuous recordings. Honeybee in-hive sounds are not static, but recording for longer than ca. 30sec is unlikely to contribute more information while consuming storage space. In the literature, longer recordings are often subdivided into short snippets as short as 1sec which are then treated as independent samples. This constitutes pseudoreplication and should be avoided. Most likely, it is more valuable to record for 10sec every hour, capturing both short-term and diurnal variability, than to record a single sample of 4min[2].

5.3.3 Model design

In Chapter 2, we found that small ANNs with simple handcrafted audio features were sufficient to detect queenlessness in most unseen test hives. The same approach was unsuccessful for most hives when predicting colony nutritional state and, to a lesser extent, colony strength.

Honeybee hives are noisy environments and each individual sample may not carry the information required to make a confident prediction. By considering this property of in-hive sound in model design, prediction accuracy for difficult targets could be improved. MIL paradigms could be particularly promising (Foulds and Frank, 2010)

[2] 10sec * 24 = 240sec = 4min

in incorporating this uncertainty. Here, rather than giving a label to each individual sample, which may or may not have the necessary information for predicting that label, we can instead give a label to a group of samples ("bag of instances"). This group would include all samples from one hive on a given day, capturing that hive's dynamics during one full day-night cycle. The MIL model could learn to tell informative from non-informative, noisy samples and consider only the informative samples when making predictions.

Feature importance information, such as from SHAP analysis, also could be used to develop more streamlined monitoring approaches. By choosing only the most important features, the size and potentially the complexity of models could be reduced. Instead of compute-heavier CNNs, smaller MLPs or even statistical models may be sufficient to make usable predictions. Optimising the feature set and model size is especially important when models are developed to run directly on in-hive devices (Edge AI). Edge computing is more efficient than uploading data to be processed centrally, particularly for data-hungry modalities like audio, and could help to increase monitoring capacities.

5.3.4 Model testing

Models imperatively need to be tested on data from fully unseen colonies. We see repeatedly that model performance is higher WG than OOG. However, only OOG tests reflect the real-world applicability of a model. Ideally, models would be tested on hives from other apiaries, but most datasets only include hives in a single location. In these cases, test sets should include multiple hives reflecting all possible treatments or conditions. The balance of classes in the test set should also match that in the training set.

Overall, studies in precision apiculture and acoustic hive monitoring should prioritise dataset size and diversity to maximise impact and applicability. This includes not only the number of independent samples (not counting subsampled snippets of longer recordings), but the number of hives in the dataset. If resources allow, the length of the monitoring period can be extended, but not at the cost of reducing hive diversity. Hives should be uniform in size as much as possible, and there should be no singleton hives.

Data sampling should aim to evenly cover the full day-night cycle, even if this comes at the expense of individual sample length. The low fundamental frequencies of most honeybee sounds allow for reduction of the audio sampling rate and therefore lower data storage demands. Finally, models should be tested for their ability to generalise to multiple unseen hives. If possible, this can include generalisation across time or across locations.

5.4 Conclusion

AI has the potential to improve how we keep honeybees and study their behaviour, with roles to play in feature discovery as much as in direct hive monitoring. However, the development of hive monitoring models with real commercial applications and value for beekeepers will require an attitude change in the precision apiculture community. If we want our work to be adopted into beekeeping practice, those who work in hive monitoring have to hold themselves to the same standards as the precision medicine community. Here, testing on unseen patients and even unseen cohorts is the norm and strictly expected to pass peer review.

Testing a model on a random set of data from training hives simply does not tell us much about the model's capabilities. At best, it shows that the features and model architecture are sufficient to detect hive- and time-based differences in the sensor data. The literature contains many hive monitoring models with sophisticated feature extraction and data augmentation pipelines that might well be able to improve outcomes for beekeepers and honeybees alike — unfortunately, we cannot know unless we properly evaluate them.

With growing interest in AI applications in agriculture, attention on precision beekeeping will continue to increase. The release of new open-source, large-scale and long-term datasets will allow us to validate our models more carefully and opens the potential for truly generalisable colony state prediction.

Appendix A

Supplement to Chapter 2

A.1 Model size and dataset splitting

Fig. A.1.1 shows test performance of small SLPs (128 nodes in hidden layer) at all sensor positions and at a range of learning rates from 0.01 to 1e-6. Training, validation, and testing sets were allocated slightly differently than in the main analysis. Models were tested on 2 hives, one from the control group and one from the treatment group. Data from another treatment hive were used for validation and early stopping. This approach to splitting the dataset is illustrated in Fig. A.1.2a. The splitting scheme was designed to show whether models were using environmental sounds or other time-linked cues to make predictions, as this would be apparent in the prediction time series for the control hive. In further analyses, I used an alternative scheme where both control hives were used in the training set (Fig. A.1.2c). By increasing the number of control samples, I hoped to reduce the likelihood that models would use these cues in the first place.

Model performance varied with position, but was broadly similar across all positions and learning rates (Fig. A.1.1). 'Inside', 'Centre', and 'all sensors' appeared most promising. For balance and even sampling of the hive environment, 'Outside' was included in further analyses, while model performance on data from positions 'Inside 2' and 'Outside 2' was not tested separately.

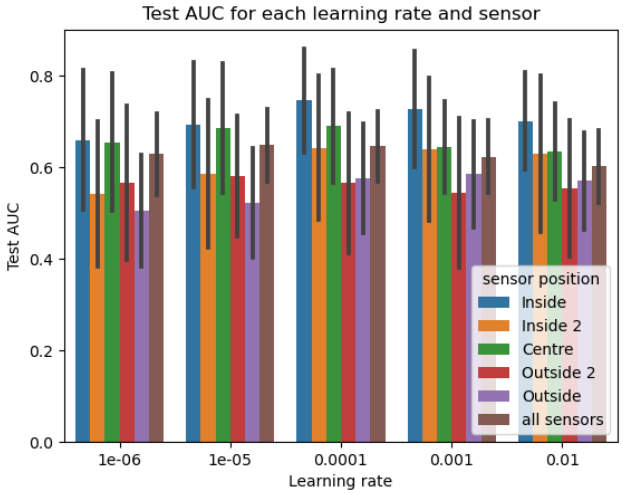


FIGURE A.1.1: Model performance at different sensor positions and learning rates, using a preliminary SLP architecture (128 nodes in hidden layer). Models were trained on 8 and tested on 2 hives (1 control and 1 treatment hive). Test hive assignment was rotated until every treatment hive had been used in the test set while alternating the control hive used. Results include models trained including and excluding hive C2, which showed high levels of Varroa infestation later in the experiment (after the end of the audio analysis period); however this was found to not have a significant effect on model performance. Black bars indicate standard deviation.

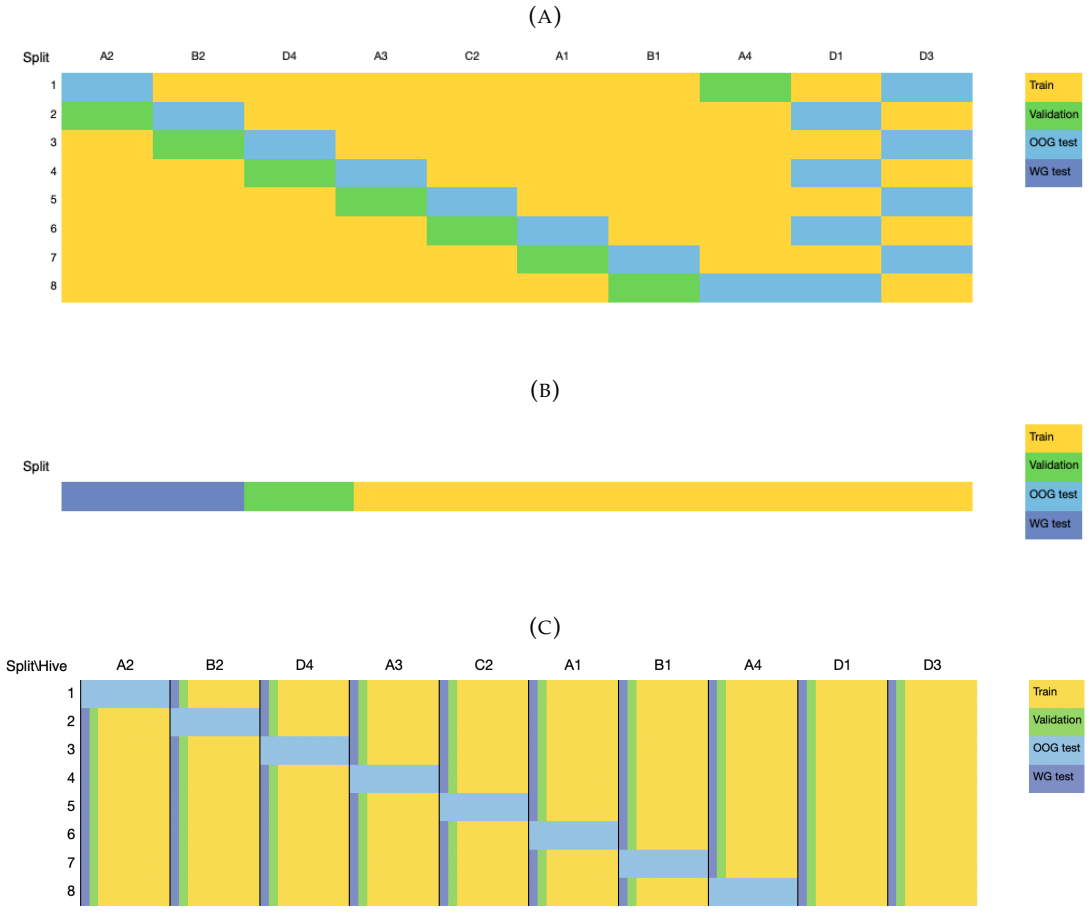


FIGURE A.1.2: Different schemes for splitting single-hive and multi-hive datasets into train, validation, and test sets. All schemes generate one fold (rows) per treatment hive (columns). (A) was used during preliminary analyses and included data from one control hive in each test fold. One treatment hive was used as the validation set. Assignments were rotated until each treatment hive had been in the test set and validation set once. Scheme (B) was used for the main analysis to split single-hive datasets. 20% of all data were split off to use as a test set. 20% of the remaining data were split off again and used as the validation set. Finally, scheme (C) was used to split multi-hive datasets. One fold was generated per treatment hive. Each treatment hive was used as the test set once. The remaining data were pooled and a random 20% were split off. 50% of this subset were used as the WG test set and the other 50% were used as the validation set.

A.2 Model performance by hive, feature set, and sensor position

Fig. A.2.1 shows MHM OOG test AUC distributions by sensor position facetted by hive and feature set for hives B1, B2, C2, and D4. The other four hives are presented in section 2.4.8.

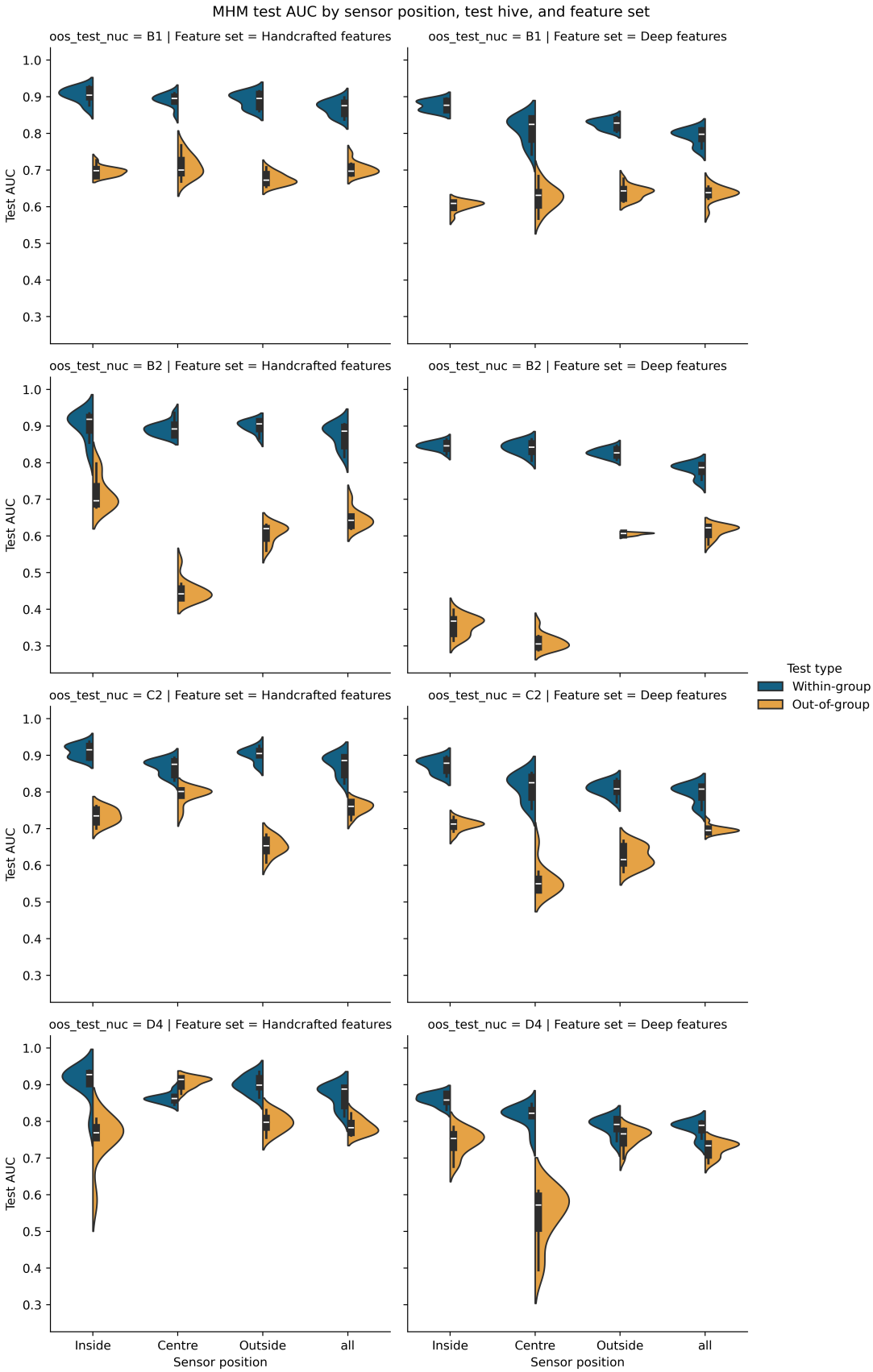


FIGURE A.2.1: OOG AUC of MHM by feature set, hive, and sensor position for hives B1, B2, C2, and D4. Model performance differed strongly by sensor position within each subset.

Appendix B

Supplement to Chapter 3

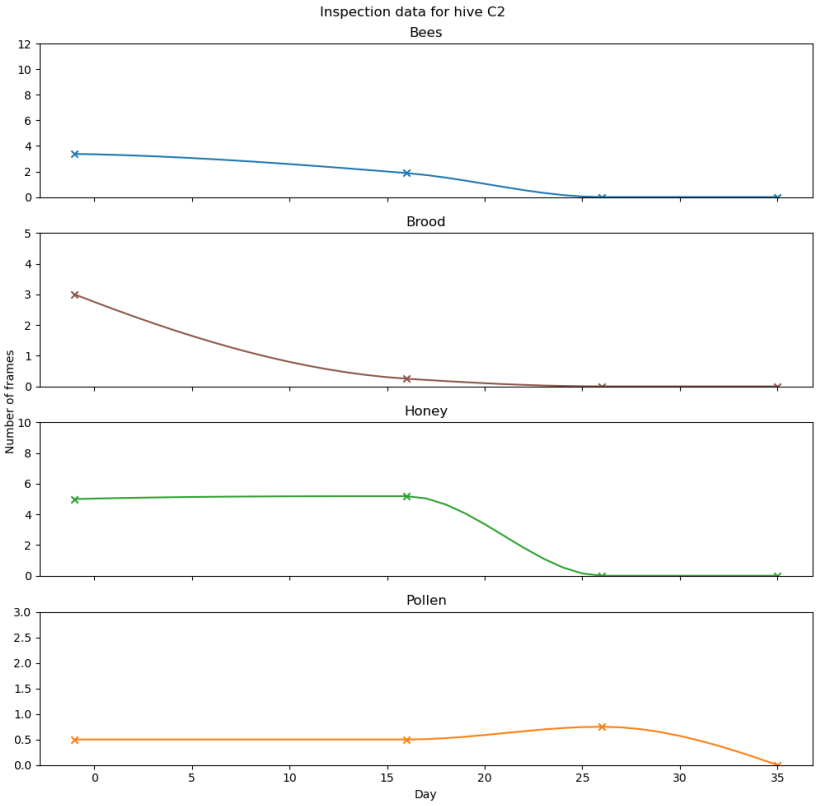
B.1 Inferring hive metrics

Daily queen status was inferred based on queen removal date and the average development time of queens from 1-day old larvae (Tab. B.1.1). Colony strength and amounts of brood, honey, and pollen stores were recorded every 1–2 weeks and interpolated between inspection dates using the PCHIP method.

Date	Hive									
	A1	A2	A3	A4	B1	B2	C2	D1	D3	D4
27-07-2023	0	0	0	0	0	0	0	0	0	0
28-07-2023	0	0	0	0	0	0	0	0	0	0
29-07-2023	0	0	0	0	0	0	0	0	0	0
30-07-2023	0	0	0	0	0	0	0	0	0	0
31-07-2023	0	0	0	0	0	0	0	0	0	0
01-08-2023	0	0	0	0	0	0	0	0	0	0
02-08-2023	0	0	0	0	0	0	0	0	0	0
03-08-2023	0	0	0	0	0	0	0	0	0	0
04-08-2023	0	0	0	0	0	0	0	0	0	0
05-08-2023	1	1	1	1	1	1	1	0	0	1
06-08-2023	1	1	1	1	1	1	1	0	0	1
07-08-2023	1	1	1	1	1	1	1	0	0	1
08-08-2023	1	1	1	1	1	1	1	0	0	1
09-08-2023	1	1	1	1	1	1	1	0	0	1
10-08-2023	1	1	1	1	1	1	1	0	0	1
11-08-2023	1	1	1	1	1	1	1	0	0	1
12-08-2023	1	1	1	1	1	1	1	0	0	1
13-08-2023	1	1	1	1	1	1	1	0	0	1
14-08-2023	1	1	1	1	1	1	1	0	0	1
15-08-2023	1	1	1	1	1	1	1	0	0	1
16-08-2023	1	1	1	1	1	1	1	0	0	1
17-08-2023	1	1	1	1	1	1	1	0	0	1

TABLE B.1.1: Queen status by day. 0 = Queenright (mated queen present); 1 = Queen-rearing/queenless.

(A) C2



(B) A3

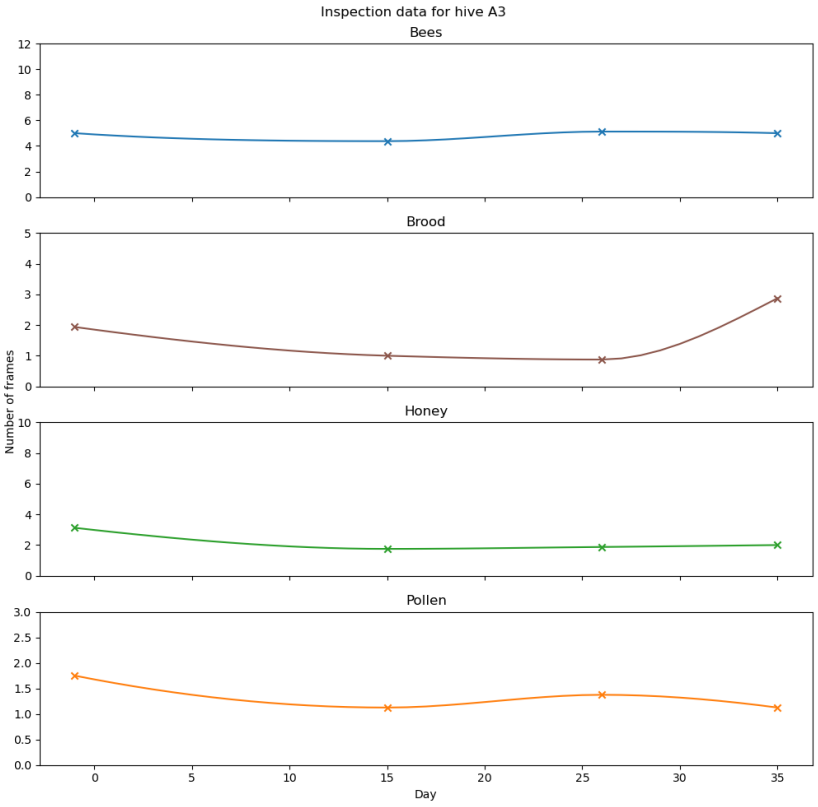
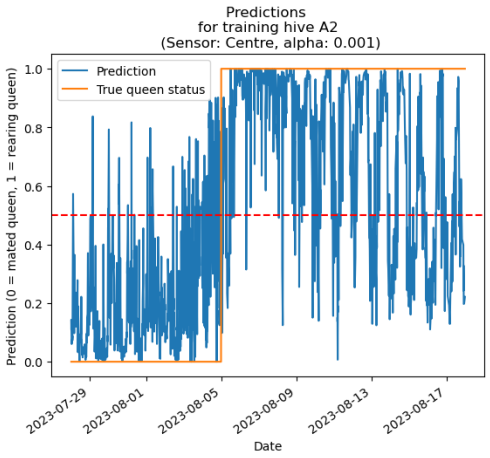


FIGURE B.1.1: Interpolated inspection data for hives C2 and A3 (Liebefeld grading).

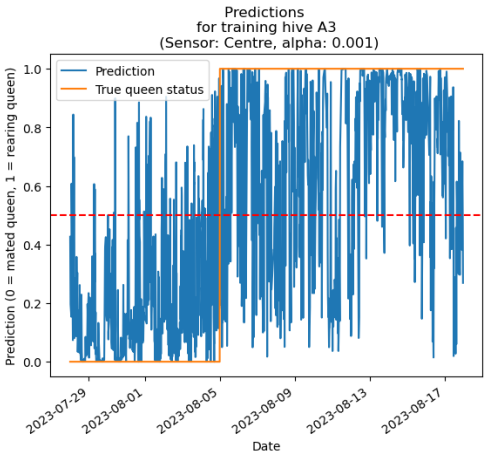
B.2 Model performance

Models were evaluated on training data to assess whether overfitting had occurred. Prediction time series showed similar levels of noise on training data compared to test data, indicating that models had not overfitted (Figs. B.2.1–B.2.2).

(A) C2 MHM — Predictions on hive A2 .



(B) C2 MHM — A3.



(C) C2 MHM — D1.

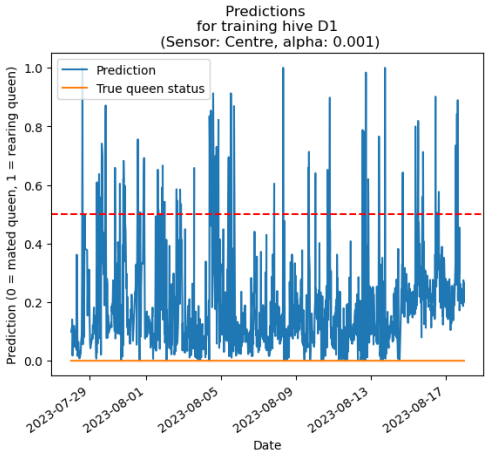
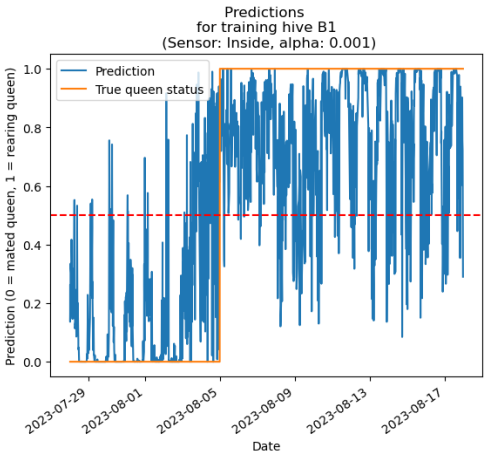
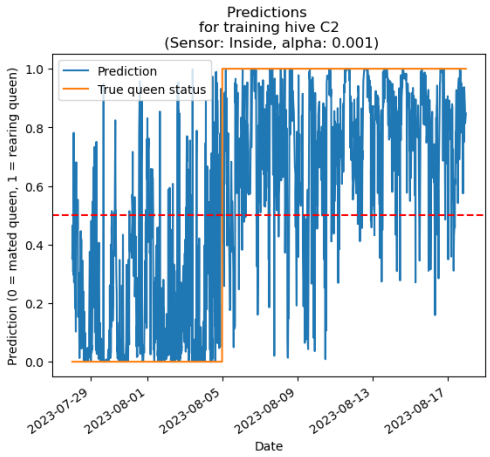


FIGURE B.2.1: Sample-level predictions of C2 Centre sensor MHM on training hives. D1 is a control hive, A2 and A3 are treatment hives. Red dashed: 0.5 threshold, orange: true queen state, blue: predicted queen state. 0=queenright, 1=queenless.

(A) A3 MHM — Predictions on hive B1.



(B) A3 MHM — C2.



(C) A3 MHM — D1.

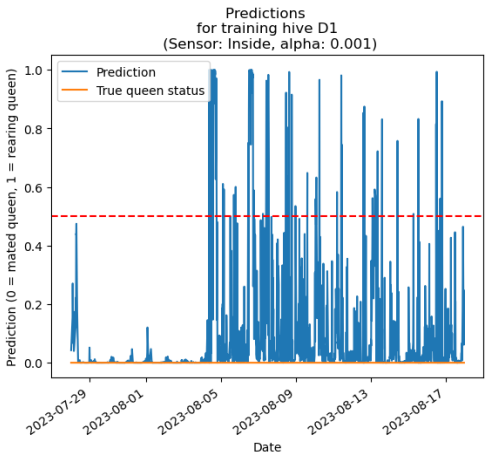


FIGURE B.2.2: Sample-level predictions of A3 Inside sensor MHM on training hives. D1 is a control hive, B1 and C2 are treatment hives. Red dashed: 0.5 threshold, orange: true queen state, blue: predicted queen state. 0=queenright, 1=queenless.

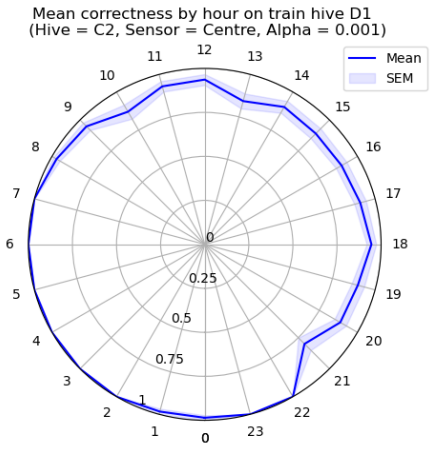
B.2.1 Daily variation in model correctness

Mean model accuracy is defined as the proportion of correct predictions over all predictions.

Model predictions for the test hive C2 showed a peak in accuracy between 8pm-1am and lowest correctness between 4am-1pm (Tab. B.2.1). Prediction correctness on the training control hive D1 was very high throughout the day, but highest between 10pm-7am (Fig. B.2.3a). Correctness for training hive A3 was stable throughout the day, with small peaks at 4-5am, 8am, 11am-12pm, 4pm, and 7-8pm. Correctness for training hive A2 shows a clearer temporal pattern, with a drop in correctness during the night from 9pm-8am (Fig. B.2.3b).

Model predictions for the test hive A3 show a peak in accuracy between 12-4pm, while predictions on the train control hive D1 are less accurate during the same time window (Tab. B.2.2 & Fig. B.2.4a). Predictions on training hive C2 are most accurate at 6-7pm, while predictions for hive B1 are high throughout with a slight dip at 2-5am (Fig. B.2.4b).

(A) Training hive D1 (ctrl)



(B) Training hive A2

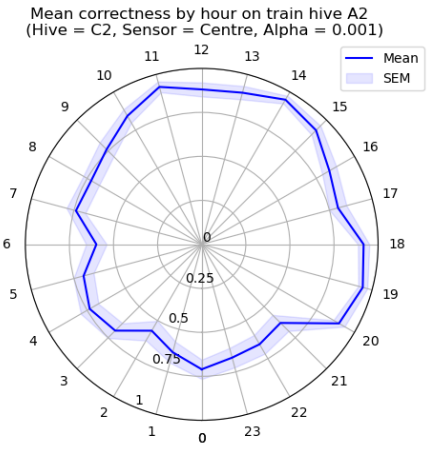
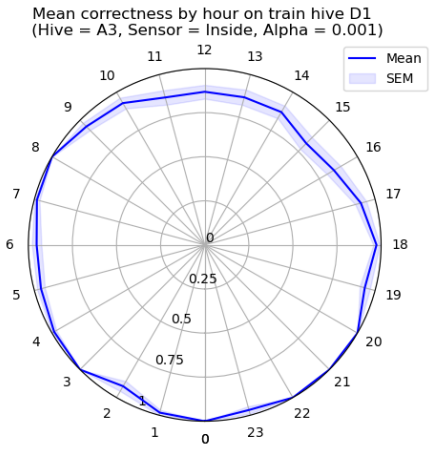


FIGURE B.2.3: Mean accuracy by hour as polar plot. Predictions of C2 MHM on training hives A2 and D1. Blue line: mean accuracy (proportion of correct predictions per hour); light blue band: standard error.

Hour	Mean accuracy (%)			
	C2 (test)	D1 (train ctrl)	A3 (train)	A2 (train)
0	87	99	76	71
1	80	98	73	63
2	74	100	67	57
3	75	100	69	69
4	67	100	80	73
5	62	100	83	69
6	62	100	78	60
7	69	100	72	74
8	69	97	81	72
9	67	95	67	76
10	78	87	77	84
11	63	93	88	93
12	63	94	90	88
13	60	84	80	89
14	71	90	77	95
15	77	89	82	92
16	77	90	90	84
17	81	91	81	80
18	73	95	75	92
19	73	90	84	95
20	79	89	82	90
21	81	80	76	63
22	80	100	79	66
23	78	100	77	67

TABLE B.2.1: Mean model accuracy by time of day for C2 model (MHM for test hive C2, $\alpha = 0.001$, repeat 1, sensor position = Centre).

(A) Training hive D1 (ctrl)



(B) Training hive B1

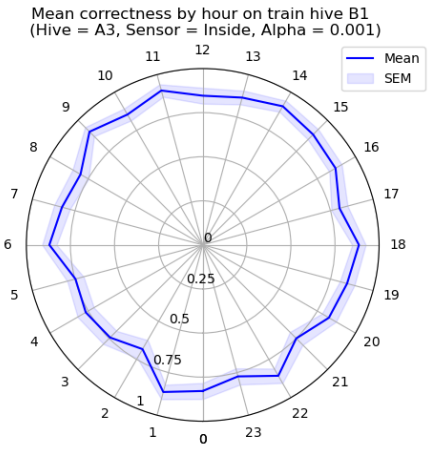


FIGURE B.2.4: Mean correctness by hour as polar plot. Predictions of A3 MHM on training hives B1 and D1. Blue line: mean correctness (proportion of correct predictions per hour); light blue band: standard error.

Hour	Mean Correctness (%)			
	A3 (test)	D1 (train ctrl)	C2 (train)	B1 (train)
0	53	100	91	83
1	52	98	88	86
2	53	93	80	68
3	57	100	80	74
4	55	99	81	77
5	46	96	74	75
6	49	95	82	87
7	58	99	77	83
8	55	100	80	80
9	58	95	88	91
10	59	93	86	85
11	77	87	83	91
12	85	87	79	85
13	81	87	87	86
14	95	87	90	91
15	87	81	89	89
16	75	85	89	87
17	64	92	89	80
18	58	97	96	89
19	63	94	99	85
20	53	100	94	83
21	49	100	83	75
22	51	100	81	86
23	48	97	84	77

TABLE B.2.2: Mean model correctness by time of day for A3 model (MHM for test hive A3, $\alpha = 0.001$, repeat 3, sensor position = Inside)

B.3 Mean SHAP values in single-hive and multi-hive models across hives

Using SHAP analysis, I calculated the contribution of each handcrafted feature to each model's output. I then averaged SHAP values over the 24 models with highest OOG performance (SHMs and MHMs) to estimate the overall importance of each feature to each model type. Plotting SHAP value distributions in frequency order shows correlation in feature importance between neighbouring frequencies (Fig. B.3.1).

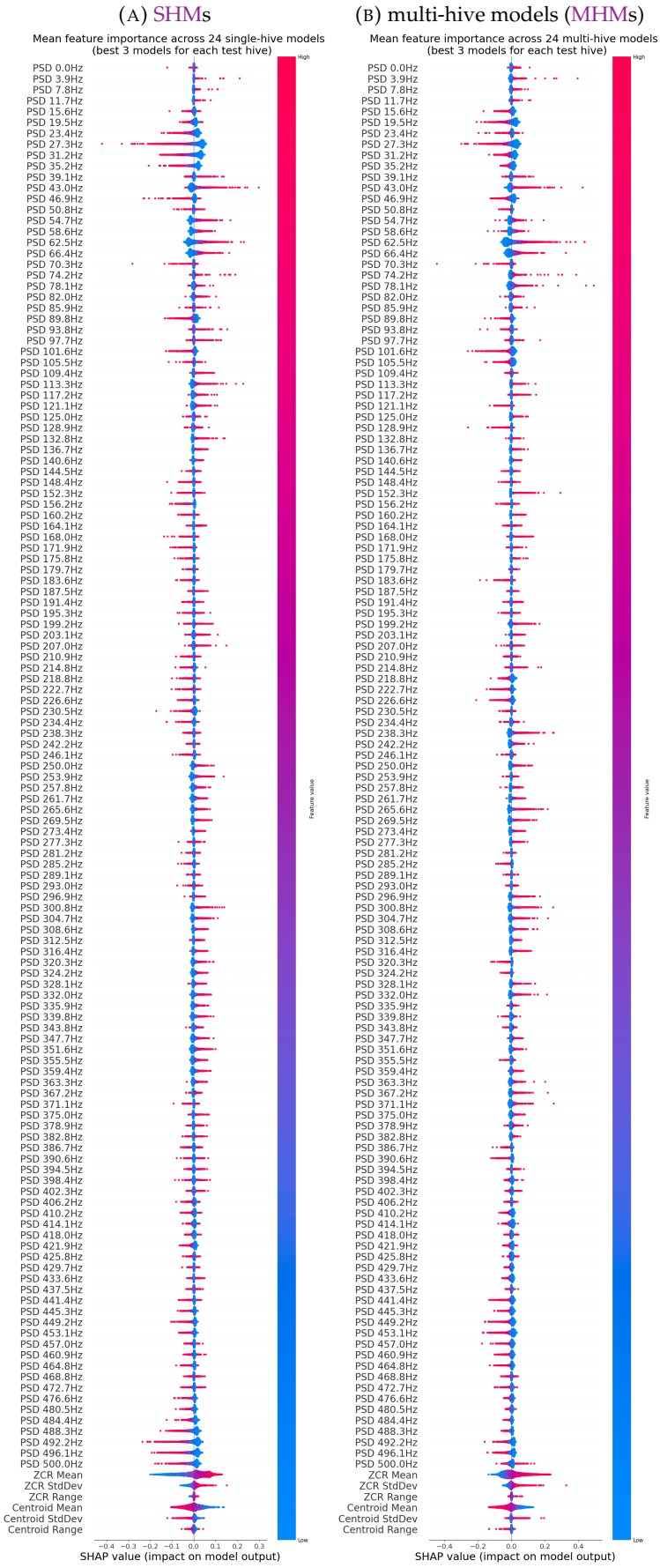


FIGURE B.3.1: Mean impacts of handcrafted features on prediction value for single-hive and multi-hive models, estimated using SHAP values (using `PermutationExplainer` and 10,000 samples). All features are presented in frequency order. Point colour indicates feature value (red = high, blue = low) and x-axis value indicates contribution to prediction value.

Appendix C

Supplement to Chapter 4

C.1 Experimental

C.1.1 Assignment of nucs to treatment groups

Colonies were assigned into experiment groups based on most recent hive inspections. I balanced groups as much as possible to evenly distribute colonies by different hive metrics: colony strength (number of beeseams), capped brood estimates, hive orientation, polytunnel side, and "readiness".

"Readiness" was judged based on the presence and laying activity of the queen. Almost all colonies were known or presumed queenright at the start of the experiment based on the presence of fresh eggs or direct observation of the queen. Five colonies had fewer than 4 beeseams and no laying queen and were omitted from the experiment at this stage (readiness score = 0). Four colonies, 6, 16, 18, and 30, were known or presumed queenless and judged to be unlikely to survive (readiness = 0.25), but were sufficiently healthy to allow for requeening. Mated queens from healthy nucleus colonies outside of the experiment were placed in these colonies on 01.09.2023. The colonies were included in the experiment going forward, and all except nuc 6 survived long enough to contribute to the analysis. The remaining 27 colonies were given readiness scores of 0.5, 0.75, or 1 depending on the certainty of queen presence and a subjective assessment of overall colony condition.

I assembled inspection data on the remaining 31 colonies to assess their current nutritional health (using colony strength and capped brood estimates) as well as their placement in the polytunnel (orientation and polytunnel side) (Table C.1.1). Colony strength and capped brood estimate data were taken from the most recent inspection

of each nuc. Nucs that had been topped up recently were not opened to avoid disruption and I assumed a strength of 5 beeseams. Hive entrance orientations were described by numerically encoded compass direction (1:SSE, 2:N, 3:NNW). The nucs were placed on small tables in two rows along the east and west side of the polytunnel (1:W, 2:E). Based on these combined hive metrics and personal assessment of readiness, colonies were grouped using `blockTools` in `R`. Readiness was used as the main grouping variable. Within each group, colonies were further grouped into four 'blocks' by the other parameters. Final groups were assigned by taking a block from each readiness group. This ensured that all treatment groups contained even numbers of colonies at each readiness level while balancing the other factors as fairly as possible.

C.1.2 Relationship between brood estimates and brood counts

Brood amounts were recorded in two ways. First, I recorded the amount of each brood stage by inspecting all frames and estimating the area of the frame side covered by each brood type. This was done to a precision of 0.5 frame sides. From D0 on, I additionally photographed each frame side and counted the number of cells with each brood stage.

I calculated the regression coefficient between the frame estimate and count for each brood stage (Tab. C.1.2).

C.1.3 Relationship between nutritional state and chalkbrood severity

I did not find any correlation of chalkbrood severity (measured as total count of chalkbrood mummies in cells, in the bottom of the hive, and at hive entrances) with colony strength (R^2 = 0.05), larvae counts (R^2 = 0.00), capped brood counts (R^2 = 0.01), or weighted brood deviation (R^2 = 0.05) (Fig. C.1.1).

C.2 Soundscape differences

Soundscapes differed between groups, but differences remained relatively stable throughout the experiment and did not show a clear shift after D0 (Figs. C.2.1 – C.2.3).

TABLE C.1.1: Colonies were assigned to groups based on judgement of readiness, colony size (beeseams), capped brood amounts, hive orientation, and placement within the polytunnel (side).

Nuc	Readiness	Beeseams	Capped	Orientation	Side	Group
1	0.5	5.5	2	1	1	1
2	0.5	6	3	2	1	1
3	1	5	1	3	1	C
5	1	5	1	2	1	1
6	0.25	5	0.5	3	1	2
8	0.5	5.5	3	2	1	3
9	0.5	5.5	1.5	3	1	C
10	1	6	3	1	1	C
11	1	6	2.5	2	1	1
12	0.75	5.5	2.5	3	1	C
13	0.75	6	2.5	1	1	3
14	1	6	3	2	1	2
15	1	5	1.5	3	1	2
16	0.25	5.5	2.5	1	1	3
17	1	5.5	2	2	1	3
18	0.25	6	3	3	1	C
19	1	5	2	1	2	1
20	0.75	5	2	2	2	2
22	0.75	6	2.5	1	2	C
23	1	5	2.5	2	2	2
24	0.75	4	1	3	2	1
25	1	5	1.5	1	2	C
26	0.75	5.5	1	2	2	3
27	1	5	1	3	2	3
28	1	5	1.5	1	2	3
29	0.5	5.5	3	2	2	2
30	0.25	5	2	3	2	1
32	0.5	4.5	2	2	2	3
33	0.75	5	0.5	3	2	1
34	0.5	6	2	1	2	C
36	0.5	4.5	1.5	3	2	2

TABLE C.1.2: Regression between brood frame estimate and cell count.

Brood stage	Intercept	Slope	R2 (%)
Eggs	6.06	36.40	60.77
Larvae	8.82	87.18	74.65
Capped brood	4.03	238.58	79.48

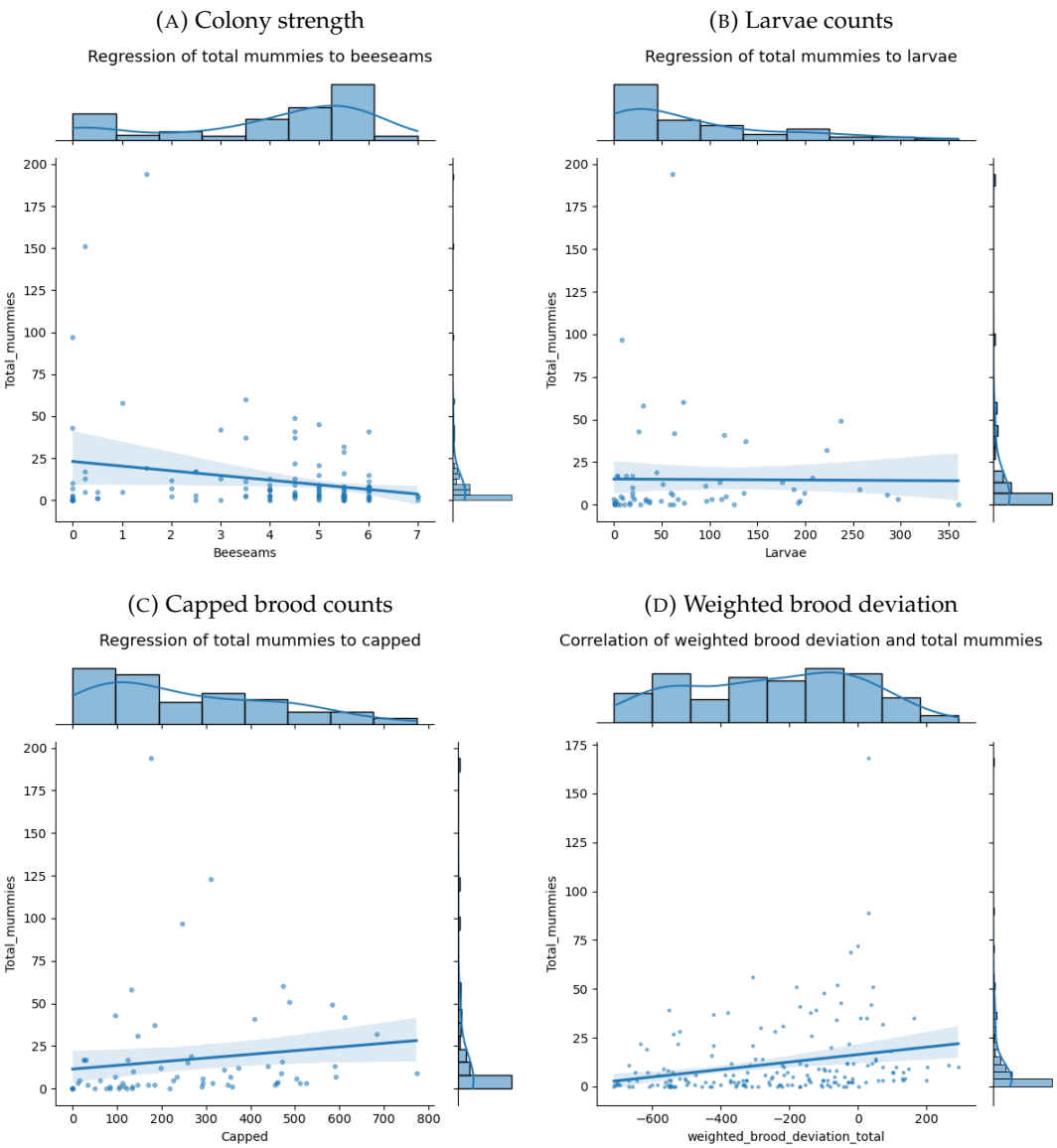
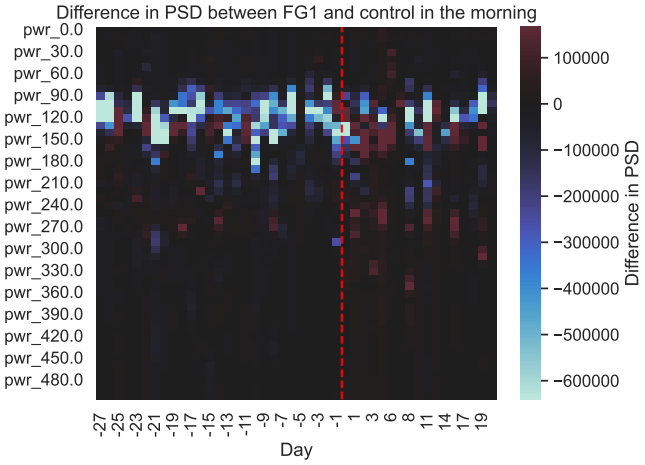
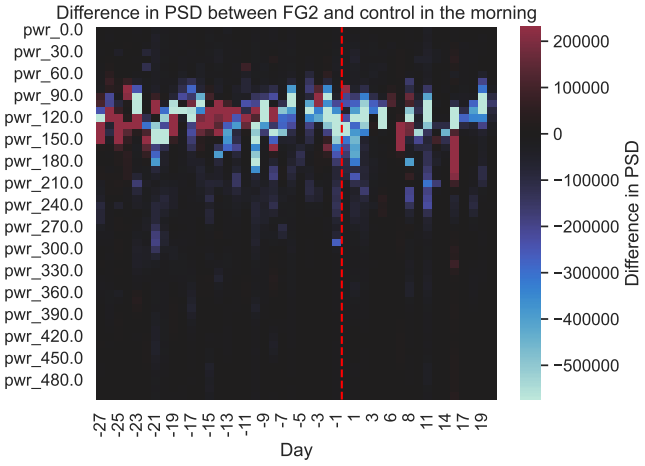


FIGURE C.1.1: Regression of chalkbrood severity (total mummy count) with population health metrics: Colony strength (A), larvae count (B), capped brood count (C), and weighted brood count deviation (larvae + capped brood) from control mean (D). N = 27 colonies. Error envelope represents 95% confidence interval.

(A)



(B)



(C)

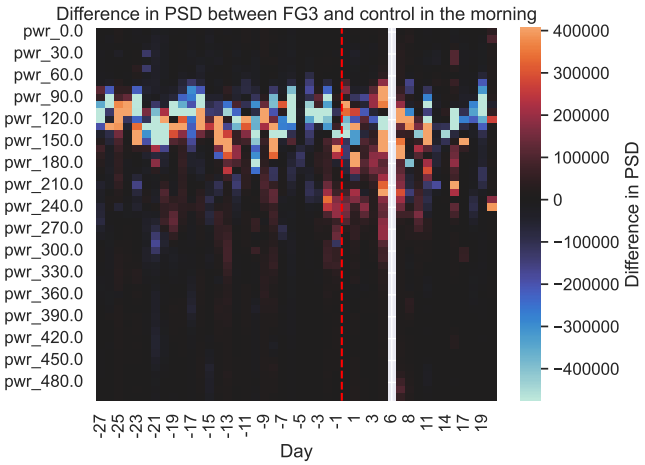
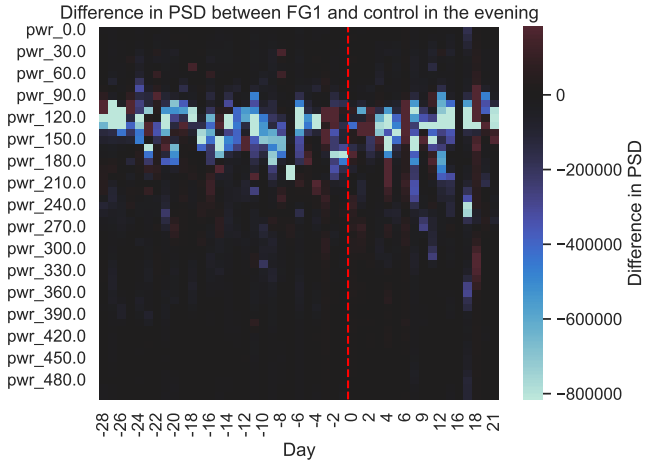
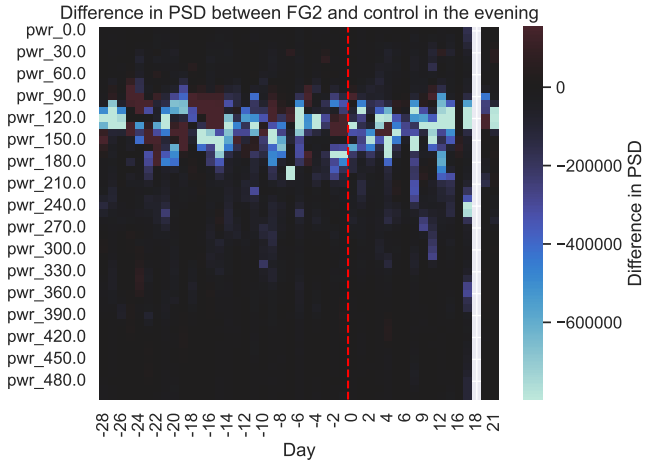


FIGURE C.2.1: Soundscape differences between each treatment group and control in the morning (04:00 – 09:59).

(A)



(B)



(C)

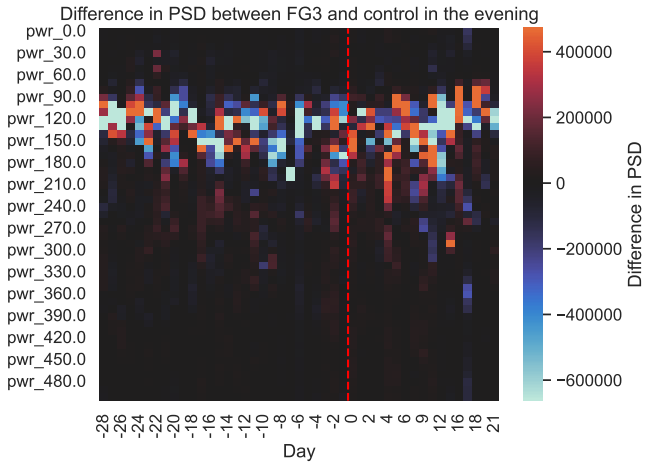
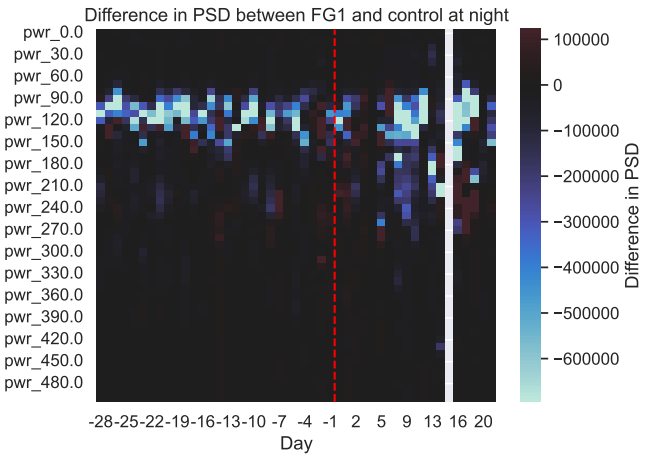
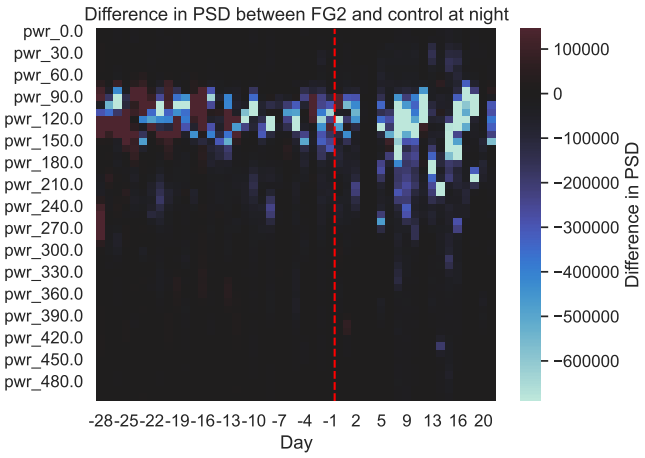


FIGURE C.2.2: Soundscape differences between each treatment group and control in the evening (16:00 – 21:59).

(A)



(B)



(C)

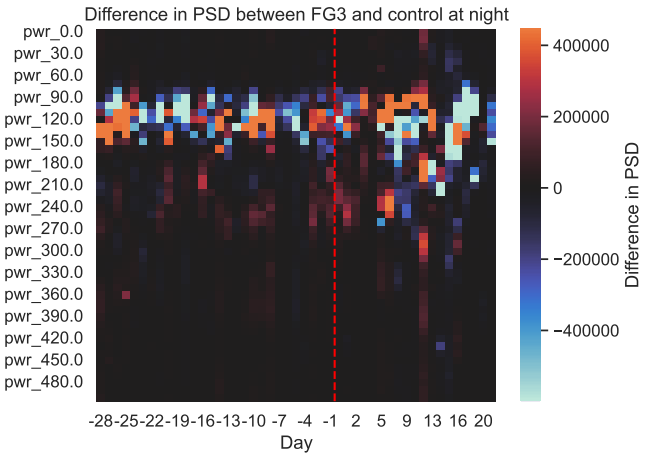


FIGURE C.2.3: Soundscape differences between each treatment group and control at night (22:00 – 03:59).

C.3 Modelling

C.3.1 Brood count weighting

Brood counts were weighted to account for the different ages of the brood. Older larvae have received more feeding than younger larvae and therefore are given a higher weighting factor (Tab. C.3.1). Both linear and growth-derived weighting were tested and were found to yield nearly identical brood deviation curves (Fig.C.3.1). The greatest difference between C.3.1 a) and b) can be seen between 22.08–01.09.2023 where the brood deviation increase in the control group and the decrease in other groups is slowed.

Capped brood are also weighted to reflect the age distribution and the overlap of feeding days between cohorts. However, since the weights tail off evenly at both ends of the age spectrum, the weighted mean here is equivalent to the unweighted mean.

TABLE C.3.1: Cohort weighting factors for calculation of colony nutritional state label from brood production.

Brood stage	Cohort age (days)	Weighting factor	
		Linear	Growth-derived
Larva	4	0.2	0.003
Larva	5	0.4	0.020
Larva	6	0.6	0.135
Larva	7	0.8	0.526
Larva	8	1.0	1.000
Capped	9	0.2	
Capped	10	0.4	
Capped	11	0.6	
Capped	12	0.8	
Capped	13	1.0	
Capped	14	1.0	
Capped	15	1.0	
Capped	16	1.0	
Capped	17	1.0	
Capped	18	0.8	
Capped	19	0.6	
Capped	20	0.4	
Capped	21	0.2	

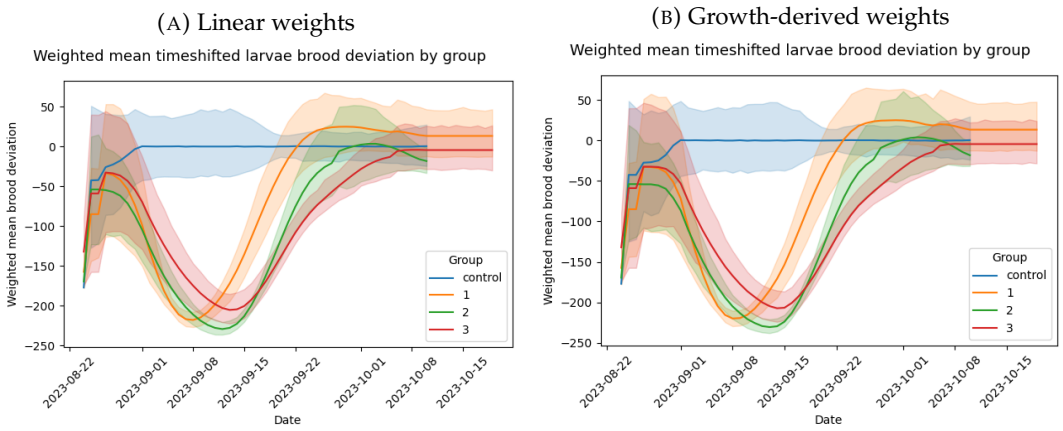


FIGURE C.3.1: Impact of weighting function on larval brood deviation metric. Using linear or growth-derived weighting factors yields nearly identical larval brood deviation curves.

C.3.2 Model checkpoint

During each model run up to 150 epochs, I checkpointed the model with the lowest validation loss as well as the final model. I observed while monitoring training that the final model often had a higher OOG test performance than the model with the lowest validation loss ("best" model). Overall, the effect was very small (Fig. C.3.2) but I chose to use only the final model when producing figures of model performance.

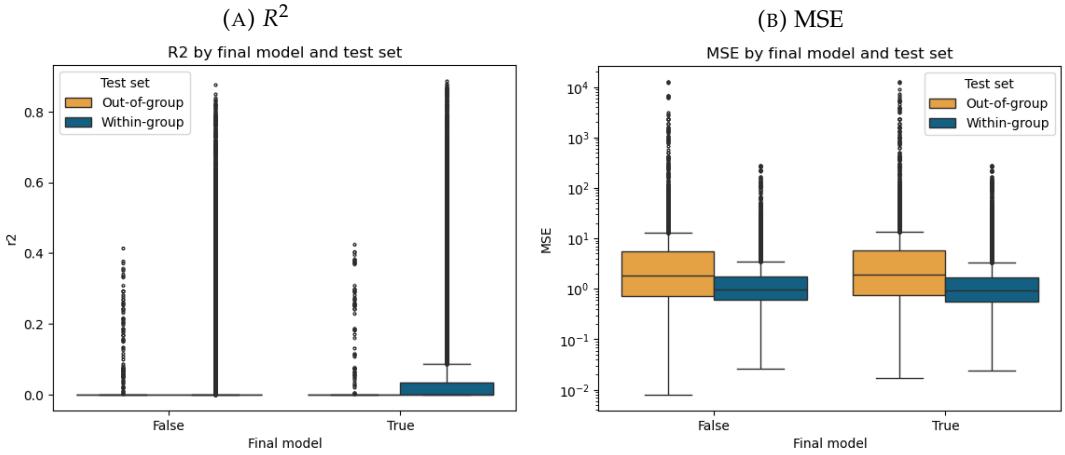


FIGURE C.3.2: Impact of model checkpoint on model performance. "False": using model with lowest validation loss; "True": using final model (epoch 150).

C.3.3 Learning rate

I tested learning rates (α) from 0.01 – 0.00001 (1e-2 – 1e-5). For single-hive model, the two highest learning rates ($\alpha = \{1e-2; 1e-3\}$) gave the best results (WG R^2 score) for both target labels (Fig. C.3.3, left panels). OOG performance was very poor no matter the learning rate.

For multi-hive model (MHM), the best OOG model performance was achieved with learning rates $\alpha = \{0.001; 1e-5\}$ for the Beeseams label, and with $\alpha = \{0.001, 0.0001, 1e-5\}$ for the brood production label (OOG R^2 score). However, WG performance was highest with $\alpha = 0.001$.

When making comparison figures of model performance, I included SHMs with $\alpha = \{0.01, 0.001\}$ and MHMs with $\alpha = \{0.001, 0.0001, 1e-5\}$.

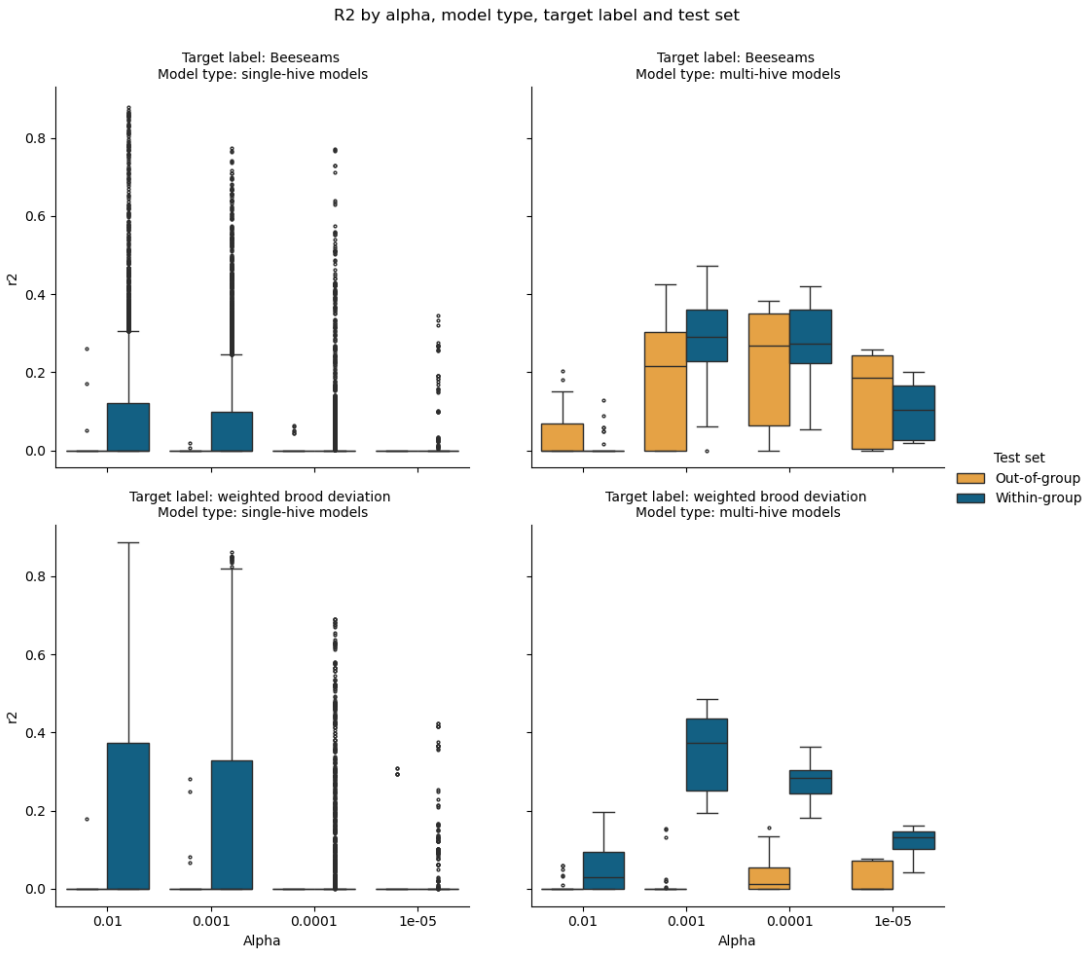
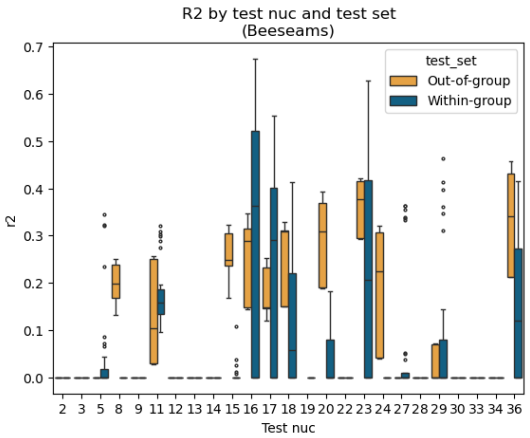


FIGURE C.3.3: Model performance (R^2 score) by model type (left – right) and target label (top – bottom).

C.3.4 Differences by test hive

Performance of multi-hive models varied strongly between test hives, both WG and OOG (Fig. C.3.4). Further, the hives with positive R^2 scores differed between target labels. Only a single hive (11) had a non-zero R^2 median for both labels.

(A) Colony strength



(B) Brood production

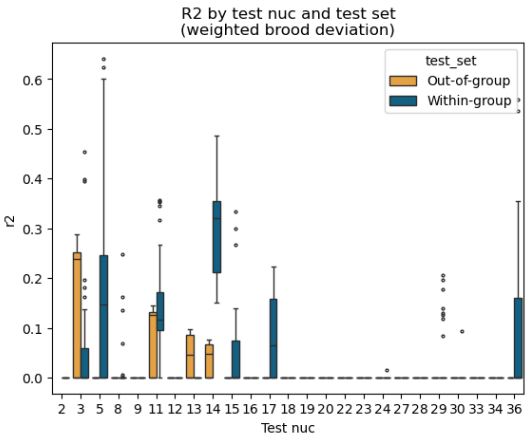


FIGURE C.3.4: MHM prediction performance by unseen test hive for colony strength label ("Beeseams", A) and colony nutritional state label ("weighted brood deviation", B).

Bibliography

Abadi, Martín et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Computer Program. URL: https://www.tensorflow.org..

Abdollahi, M. et al. (2025). "UrBAN: Urban Beehive Acoustics and PheNotyping Dataset". In: *Sci Data* 12.1, p. 536. ISSN: 2052-4463 (Electronic) 2052-4463 (Linking). DOI: 10.1038/s41597-025-04869-1. URL: https://www.ncbi.nlm.nih.gov/pubmed/40164648.

Abdollahi, Mahsa, Pierre Giovenazzo, and Tiago H. Falk (2022). "Automated Beehive Acoustics Monitoring: A Comprehensive Review of the Literature and Recommendations for Future Work". In: *Applied Sciences* 12.8. ISSN: 2076-3417. DOI: 10.3390/app12083920.

Agarwal, Shivang, Ajita Rattani, and C. Ravindranath Chowdary (2021). "A comparative study on handcrafted features v/s deep features for open-set fingerprint liveness detection". In: *Pattern Recognition Letters* 147, pp. 34–40. ISSN: 01678655. DOI: 10.1016/j.patrec.2021.03.032.

Ahn, Kiheung et al. (2012). "Effects of Long Distance Transportation on Honey Bee Physiology". In: *Psyche: A Journal of Entomology* 2012, pp. 1–9. ISSN: 0033-2615 1687-7438. DOI: 10.1155/2012/193029.

Alaux, Cédric et al. (2010). "Diet effects on honeybee immunocompetence". In: *Biology Letters* 6.4, pp. 562–5. ISSN: 1744-957X (Electronic) 1744-9561 (Print) 1744-9561 (Linking). DOI: 10.1098/rsbl.2009.0986. URL: https://www.ncbi.nlm.nih.gov/pubmed/20089536.

Allen, M. Delia (1955). "Observations on honeybees attending their queen". In: *The British Journal of Animal Behaviour* 3.2, pp. 66–69.

— (1956). "The behaviour of honeybees preparing to swarm". In: 4.1, pp. 14–22. ISSN: 0950-5601. DOI: 10.1016/S0950-5601(56)80011-7.

Althnian, Alhanoof et al. (2021). "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain". In: *Applied Sciences* 11.2. ISSN: 2076-3417. DOI: 10.3390/app11020796.

Alías, Francesc, Joan Socoró, and Xavier Sevillano (2016). "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sounds". In: *Applied Sciences* 6.5, pp. 1–44. ISSN: 2076-3417. DOI: 10.3390/app6050143.

Amdam, Gro Vang and Stig W. Omholt (2002). "The regulatory anatomy of honeybee lifespan". In: *J Theor Biol* 216.2, pp. 209–28. ISSN: 0022-5193 (Print) 0022-5193 (Linking). DOI: 10.1006/jtbi.2002.2545. URL: https://www.ncbi.nlm.nih.gov/pubmed/12079372.

— (2003). "The hive bee to forager transition in honeybee colonies: the double repressor hypothesis". In: *Journal of theoretical Biology* 223.4, pp. 451–64. ISSN: 0022-5193 (Print) 0022-5193 (Linking). DOI: 10.1016/s0022-5193(03)00121-8. URL: https://www.ncbi.nlm.nih.gov/pubmed/12875823.

Amiri, Esmaeil et al. (2020). "Egg-size plasticity in Apis mellifera: Honey bee queens alter egg size in response to both genetic and environmental factors". In: *J Evol Biol* 33.4, pp. 534–543. ISSN: 1420-9101 (Electronic) 1010-061X (Linking). DOI: 10.1111/jeb.13589. URL: https://www.ncbi.nlm.nih.gov/pubmed/31961025.

Archer, C. Ruth et al. (2014). "Nutrition affects survival in African honeybees exposed to interacting stressors". In: *Functional Ecology* 28.4, pp. 913–923. ISSN: 0269-8463 1365-2435. DOI: 10.1111/1365-2435.12226.

Aronstein, K. A. and K. D. Murray (2010). "Chalkbrood disease in honey bees". In: *J Invertebr Pathol* 103 Suppl 1, S20–9. ISSN: 1096-0805 (Electronic) 0022-2011 (Linking). DOI: 10.1016/j.jip.2009.06.018. URL: https://www.ncbi.nlm.nih.gov/pubmed/19909969.

Avni, Dorit, Arnon Dag, and Sharoni Shafir (2015). "The effect of surface area of pollen patties fed to honey bee (Apis mellifera) colonies on their consumption, brood production and honey yields". In: *Journal of Apicultural Research* 48.1, pp. 23–28. ISSN: 0021-8839 2078-6913. DOI: 10.3896/ibra.1.48.1.06.

Baker, Ellen Carmanah (2024). "Ecology and evolution of sterols in pollen and wild bees". Thesis. DOI: 10.5287/ora-9ojr1g52o.

Beekman, Madeleine et al. (2004). "Comparing Foraging Behaviour of Small and Large Honey-Bee Colonies by Decoding Waggle Dances Made by Foragers". In: *Functional Ecology* 18.6, pp. 829–835. URL: https://www.jstor.org/stable/3599110.

Bencsik, Martin and Michael I. Newton (2018). *Honey Bee Vibration Monitoring Using the 805M1 Accelerometer*. Conference Paper. DOI: 10.3390/ecsa-5-05637.

Biswas, Sumit (2020). "Intelligent beehive status monitoring in noisy environment". Thesis.

Bonnaffé, Willem, Ben C. Sheldon, and Tim Coulson (2021). "Neural ordinary differential equations for ecological and evolutionary time-series analysis". In: *Methods in Ecology and Evolution* 12.7, pp. 1301–1315. ISSN: 2041-210X 2041-210X. DOI: 10.1111/2041-210x.13606.

Bricout, A. et al. (2024). "Bee Together: Joining Bee Audio Datasets for Hive Extrapolation in AI-Based Monitoring". In: *Sensors (Basel)* 24.18. ISSN: 1424-8220 (Electronic) 1424-8220 (Linking). DOI: 10.3390/s24186067. URL: https://www.ncbi.nlm.nih.gov/pubmed/39338812.

Buckemüller, Christina et al. (2017). "Octopamine Underlies the Counter-Regulatory Response to a Glucose Deficit in Honeybees (Apis mellifera)". In: *Frontiers in Systems Neuroscience* 11, p. 63. ISSN: 1662-5137 (Print) 1662-5137 (Electronic) 1662-5137 (Linking). DOI: 10.3389/fnsys.2017.00063. URL: https://www.ncbi.nlm.nih.gov/pubmed/28912693.

Butler, C. G. (1954). "The Method and Importance of the Recognition by a Colony of Honeybees (A. Mellifera) of the Presence of Its Queen". In: *Transactions of the Royal Entomological Society of London* 105.2, pp. 11–29. ISSN: 0035-8894. DOI: 10.1111/j.1365-2311.1954.tb00773.x.

Camazine, Scott (1991). "Self-organizing pattern formation on the combs of honey bee colonies". In: *Behavioral Ecology and Sociobiology* 28.1, pp. 61–76. ISSN: 0340-5443. DOI: 10.1007/BF00172140.

— (1993). "The regulation of pollen foraging by honey bees: how foragers assess the colony's need for pollen". In: *Behavioral Ecology and Sociobiology* 32.4, pp. 265–272. ISSN: 0340-5443. DOI: 10.1007/BF00166516.

Cao, Tuan T. et al. (2007). "The influence of the vibration signal on worker interactions with the nest and nest mates in established and newly founded colonies of the

honey bee, Apis mellifera". In: *Insectes Sociaux* 54.2, pp. 144–149. ISSN: 0020-1812 1420-9098. DOI: 10.1007/s00040-007-0921-1.

Cavigli, I. et al. (2016). "Pathogen prevalence and abundance in honey bee colonies involved in almond pollination". In: *Apidologie* 47.2, pp. 251–266. ISSN: 0044-8435 (Print) 1297-9678 (Electronic) 0044-8435 (Linking). DOI: 10.1007/s13592-015-0395-5. URL: https://www.ncbi.nlm.nih.gov/pubmed/27053820.

Chazette, Larissa, Matthias Becker, and Helena Szczerbicka (2016). *Basic algorithms for bee hive monitoring and laser-based mite control*. Conference Paper. DOI: 10.1109/SSCI.2016.7850001.

Chollet, François et al. (2015). *Keras*. Computer Program. URL: https://keras.io.

Cook, D. et al. (2022). "Temperature Sensing and Honey Bee Colony Strength". In: *J Econ Entomol* 115.3, pp. 715–723. ISSN: 1938-291X (Electronic) 0022-0493 (Print) 0022-0493 (Linking). DOI: 10.1093/jee/toac034. URL: https://www.ncbi.nlm.nih.gov/pubmed/35522232.

Corby-Harris, Vanessa, Lucy Snyder, and Charlotte Meador (2019). "Fat body lipolysis connects poor nutrition to hypopharyngeal gland degradation in Apis mellifera". In: *J Insect Physiol* 116, pp. 1–9. ISSN: 1879-1611 (Electronic) 0022-1910 (Linking). DOI: 10.1016/j.jinsphys.2019.04.001. URL: https://www.ncbi.nlm.nih.gov/pubmed/30953617.

Crailsheim, K. et al. (2003). "Collective and individual nursing investment in the queen and in young and old honeybee larvae during foraging and non-foraging periods". In: *Insectes Sociaux* 50.2, pp. 174–184. ISSN: 0020-1812 1420-9098. DOI: 10.1007/s00040-003-0644-x.

Crailsheim, Karl (1990). "The protein balance of the honey bee worker". In: *Apidologie* 21, pp. 417–429.

— (1991). "Interadult feeding of jelly in honeybee (Apis mellifera L.) colonies". In: *Journal of Comparative Physiology B* 161.1, pp. 55–60. ISSN: 0174-1578. DOI: 10.1007/BF00258746.

Dainat, Benjamin et al. (2020). "A scientific note on the 'Liebefeld Method' to estimate honey bee colony strength: its history, use, and translation". In: *Apidologie* 51.3, pp. 422–427. ISSN: 0044-8435 1297-9678. DOI: 10.1007/s13592-019-00728-2.

Deng, Jia et al. (2009). *ImageNet: A Large-Scale Hierarchical Image Database*. Conference Paper.

Di Pasquale, G. et al. (2013). "Influence of pollen nutrition on honey bee health: do pollen quality and diversity matter?" In: *PLoS One* 8.8, e72016. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0072016. URL: https://www.ncbi.nlm.nih.gov/pubmed/23940803.

Dolezal, A. G. et al. (2019). "Interacting stressors matter: diet quality and virus infection in honeybee health". In: *R Soc Open Sci* 6.2, p. 181803. ISSN: 2054-5703 (Print) 2054-5703 (Electronic) 2054-5703 (Linking). DOI: 10.1098/rsos.181803. URL: https://www.ncbi.nlm.nih.gov/pubmed/30891288.

Dreller, Claudia and W.H. Kirchner (1993). "Hearing in honeybees: localization of the auditory sense organ". In: *Journal of Comparative Physiology A* 173.3, pp. 175–279. ISSN: 0340-7594. DOI: 10.1007/BF00212691.

Duarte, Adam et al. (2024). "Passive acoustic monitoring and convolutional neural networks facilitate high-resolution and broadscale monitoring of a threatened species". In: *Ecological Indicators* 162. ISSN: 1470160X. DOI: 10.1016/j.ecolind.2024.112016.

Dyer, Fred C. (2002). "The biology of the dance language". In: *Annual Review of Entomology* 47.1, pp. 917–949. ISSN: 0066-4170. DOI: 10.1146/annurev.ento.47.091201.145306.

Dynes, Travis L. et al. (2019). "Reduced density and visually complex apiaries reduce parasite load and promote honey production and overwintering survival in honey bees". In: *PLoS One* 14.5, e0216286. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0216286. URL: https://www.ncbi.nlm.nih.gov/pubmed/31120911.

Ellis, Rebecca A. et al. (2020). "From a free gift of nature to a precarious commodity: Bees, pollination services, and industrial agriculture". In: *Journal of Agrarian Change* 20.3, pp. 437–459. ISSN: 1471-0358 1471-0366. DOI: 10.1111/joac.12360.

Esch, Harald (1961a). "Ein neuer Bewegungstyp im Schwänzeltanz der Bienen". In: *Die Naturwissenschaften* 48.5, pp. 140–141. ISSN: 0028-1042. DOI: 10.1007/BF00631947.

— (1961b). "Über die Schallerzeugung beim Werbetanz der Honigbiene". In: *Zeitschrift für Vergleichende Physiologie* 45, pp. 1–11. ISSN: 0340-7594. DOI: 10.1007/BF00297754.

Esch, Harald (1964). "Beiträge zum Problem der Entfernungsweisung in den Schwänzeltänzen der Honigbiene". In: *Zeitschrift für Vergleichende Physiologie* 48, pp. 534–546. ISSN: 0340-7594. DOI: 10.1007/BF00348847.

— (1967). "The evolution of bee language". In: *Scientific American* 216.4, pp. 96–105. URL: https://www.jstor.org/stable/10.2307/24931472.

Evans, Huw (2021). Personal Communication.

Evans, S. K. et al. (2024). "Hive Orientation and Colony Strength Affect Honey Bee Colony Activity during Almond Pollination". In: *Insects* 15.2. ISSN: 2075-4450 (Print) 2075-4450 (Electronic) 2075-4450 (Linking). DOI: 10.3390/insects15020112. URL: https://www.ncbi.nlm.nih.gov/pubmed/38392532.

Fahrenholz, L., I. Lamprecht, and B. Schricker (1989). "Thermal investigations of a honey bee colony: thermoregulation of the hive during summer and winter and heat production of members of different bee castes". In: *J Comp Physiol B* 159, pp. 551–560.

Farina, Almo (2023). "Discovering ecoacoustic codes in beehives: First evidence and perspectives". In: *Biosystems* 234, p. 105041. ISSN: 1872-8324 (Electronic) 0303-2647 (Linking). DOI: 10.1016/j.biosystems.2023.105041. URL: https://www.ncbi.nlm.nih.gov/pubmed/37806648.

Favaro, Livio et al. (2021). "Passive acoustic monitoring of the endangered African Penguin (Spheniscus demersus) using autonomous recording units and ecoacoustic indices". In: *Ibis* 163.4, pp. 1472–1480. ISSN: 0019-1019 1474-919X. DOI: 10.1111/ibi.12970.

Favre, Daniel (2011). "Mobile phone-induced honeybee worker piping". In: *Apidologie* 42.3, pp. 270–279. ISSN: 0044-8435 1297-9678. DOI: 10.1007/s13592-011-0016-x.

Fell, R.D. and R. A. Morse (1984). "Emergency queen cell production in the honey bee colony". In: *Insectes Soc* 31.3, pp. 221–237.

Ferrari, Sara et al. (2008). "Monitoring of swarming sounds in bee hives for early detection of the swarming period". In: *Computers and Electronics in Agriculture* 64.1, pp. 72–77. ISSN: 01681699. DOI: 10.1016/j.compag.2008.05.010.

Fewell, Jennifer H. and Mark L. Winston (1992). "Colony state and regulation of pollen foraging in the honey bee, Apis mellifera L." In: *Behavioral Ecology and Sociobiology* 30.6, pp. 387–393. ISSN: 0340-5443. DOI: 10.1007/BF00176173.

Fewell, Jennifer H., Ronald C. Ydenberg, and Mark L. Winston (1991). "Individual foraging effort as a function of colony population in the honey bee, Apis mellifera L." In: *Animal Behaviour* 42, pp. 153–155.

Foulds, James and Eibe Frank (2010). "A review of multi-instance learning assumptions". In: *The Knowledge Engineering Review* 25.1, pp. 1–25. ISSN: 0269-8889 1469-8005. DOI: 10.1017/s026988890999035x.

Fourer, Dominique and Agnieszka Orlowska (2022). *Detection and identification of beehive piping audio signals*. Conference Paper.

Frisch, Karl von (1967). *The Dance Language and Orientation of Bees*. Harvard University Press.

Gahl, Robert A. (1975). "The shaking dance of honey bee workers: Evidence for age discrimination". In: *Animal Behaviour* 23.1, pp. 230–232. ISSN: 0003-3472. DOI: 10.1016/0003-3472(75)90068-8.

Gallus, Anja et al. (2012). "Use of static passive acoustic monitoring to assess the status of the 'Critically Endangered' Baltic harbour porpoise in German waters". In: *Endangered Species Research* 18.3, pp. 265–278. ISSN: 1863-5407 1613-4796. DOI: 10.3354/esr00448.

Gonçalves, Rui (2021). Personal Communication.

Goodrich, Brittney K. (2019). "Do more bees imply higher fees? Honey bee colony strength as a determinant of almond pollination fees". In: *Food Policy* 83, pp. 150–160. ISSN: 03069192. DOI: 10.1016/j.foodpol.2018.12.008.

Gould, James L. (1976). "The Dance-Language Controversy". In: *The Quarterly Review of Biology* 51.2, pp. 211–244. URL: https://www.jstor.org/stable/2823629.

Gouyon, Fabien, François Pachet, and Olivier Delerue (2000). "On the use of zero-crossing rate for an application of classification of percussive sounds". In: *COST G-6 Conference on Digital Audio Effects (DAFX-00)*.

Grodzicki, Przemysław, Bartosz Piechowicz, and Michał Caputa (2020). "The Effect of the Queen's Presence on Thermal Behavior and Locomotor Activity of Small Groups of Worker Honey Bees". In: *Insects* 11.8. ISSN: 2075-4450 (Print) 2075-4450 (Electronic) 2075-4450 (Linking). DOI: 10.3390/insects11080464. URL: https://www.ncbi.nlm.nih.gov/pubmed/32717876.

Grooters, H.J. (1987). "Influences of queen piping and worker behaviour on the timing of emergence of honey bee queens". In: *Insectes Sociaux* 34.3, pp. 181–193. ISSN: 0020-1812. DOI: 10.1007/BF02224083.

Harris, Charles R. et al. (2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Harrison, Jon M. (1986). "Caste-specific changes in honeybee flight capacity". In: *Physiological Zoology* 59.2, pp. 175–187. URL: https://www.jstor.org/stable/30156031.

Hasegawa, Yuji and Hidetoshi Ikeno (2011). "How do honeybees attract nestmates using waggle dances in dark and noisy hives?" In: *PLoS One* 6.5, e19619. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0019619. URL: https://www.ncbi.nlm.nih.gov/pubmed/21603608.

He, Xu Jiang et al. (2016). "Starving honey bee (Apis mellifera) larvae signal pheromonally to worker bees". In: *Scientific Reports* 6, p. 22359. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/srep22359. URL: https://www.ncbi.nlm.nih.gov/pubmed/26924295.

Hrassnigg, Norbert and Karl Crailsheim (1998). "The influence of brood on the pollen consumption of worker bees (Apis mellifera L.)" In: *Journal of Insect Physiology* 44, pp. 393–404.

Hrncir, Michael, Friedrich G. Barth, and Jürgen Tautz (2006). "Vibratory and Airborne-Sound Signals in Bee Communication (Hymenoptera)". In: *Insect Sounds and Communication: Physiology, Behaviour, Ecology and Evolution*. Boca Raton, London. ISBN: 9780849320606.

Hrncir, Michael, Camila Maia-Silva, and Walter M. Farina (2019). "Honey bee workers generate low-frequency vibrations that are reliable indicators of their activity level". In: *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 205.1, pp. 79–86. ISSN: 1432-1351 (Electronic) 0340-7594 (Linking). DOI: 10.1007/s00359-018-1305-x. URL: https://www.ncbi.nlm.nih.gov/pubmed/30465281.

Hrncir, Michael et al. (2004). "Thorax vibrations of a stingless bee (Melipona seminigra). II. Dependence on sugar concentration". In: *Journal of Comparative Physiology A* 190.7. ISSN: 0340-7594 1432-1351. DOI: 10.1007/s00359-004-0515-6.

Hrncir, Michael et al. (2011). "The recruiter's excitement — features of thoracic vibrations during the honey bee's waggle dance related to food source profitability". In: *J Exp Biol* 214.Pt 23, pp. 4055–64. ISSN: 1477-9145 (Electronic) 0022-0949 (Linking). DOI: 10.1242/jeb.063149. URL: https://www.ncbi.nlm.nih.gov/pubmed/22071198.

Huang, Zhi-Yong and Gard W. Otis (1991). "Inspection and Feeding of Larvae by Worker Honey Bees (Hymenoptera: Apidae): Effect of Starvation and Food Quantity". In: *Journal of Insect Behavior* 4.3, pp. 305–317. ISSN: 0892-7553. DOI: 10.1007/BF01048280.

Huang, Zhi-Yong et al. (1991). "Hormonal regulation of behavioural development in the honey bee is based on changes in the rate of juvenile hormone biosynthesis". In: *Journal of Insect Physiology* 37.10, pp. 733–741.

Hunt, J. H. and F. J. Richard (2013). "Intracolony vibroacoustic communication in social insects". In: *Insectes Sociaux* 60.4, pp. 403–417. ISSN: 0020-1812 1420-9098. DOI: 10.1007/s00040-013-0311-9.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Imdorf, Anton et al. (1987). "Überprüfung der Schätzmethode zur Ermittlung der Brutfläche und der Anzahl Arbeiterinnen in freifliegenden Bienenvölkern". In: *Apidologie* 18.2, pp. 137–146.

iPollinate (2025). Web Page. URL: https://ipollinate.eu.

Johnson, Brian R. (2003). "Organization of work in the honeybee: a compromise between division of labour and behavioural flexibility". In: *Proc Biol Sci* 270.1511, pp. 147–52. ISSN: 0962-8452 (Print) 1471-2954 (Electronic) 0962-8452 (Linking). DOI: 10.1098/rspb.2002.2207. URL: https://www.ncbi.nlm.nih.gov/pubmed/12590752.

Junge, M. (2006). "Fanning in honey bees—a comparison between measurement and calculation of non-stationary aerodynamic forces". In: *Flow Phenomena in Nature Volume 1*. WIT Transactions on State of the Art in Science and Engineering, pp. 219–243. ISBN: 1-84564-001-2. DOI: 10.2495/1-84564-001-2/3d.

Kaatz, Hannes, Stefan Eichmüller, and Sabine Kreissl (1994). "Stimulatory effect of octopamine on juvenile hormone biosynthesis in honey bees (Apis mellifera): Physiological and immunocytochemical evidence". In: *Journal of Insect Physiology* 40.10, pp. 865–872. DOI: 10.1016/0022-1910(94)90020-5. URL: https://nbn-resolving.org/urn:nbn:de:bsz:352-163244.

Kanelis, D. et al. (2023). "Decoding the Behavior of a Queenless Colony Using Sound Signals". In: *Biology (Basel)* 12.11, p. 1392. ISSN: 2079-7737 (Print) 2079-7737 (Electronic) 2079-7737 (Linking). DOI: 10.3390/biology12111392. URL: https://www.ncbi.nlm.nih.gov/pubmed/37997991.

Katzav-Gozansky, Tamar et al. (2004). "Queen-signal modulation of worker pheromonal composition in honeybees". In: *Proceedings of the Royal Society. B* 271.1552, pp. 2065–2069. ISSN: 0962-8452 (Print) 1471-2954 (Electronic) 0962-8452 (Linking). DOI: 10.1098/rspb.2004.2839. URL: https://www.ncbi.nlm.nih.gov/pubmed/15451697.

Kilpinen, Ole and Jesper Storm (1997). "Biophysics of the subgenual organ of the honeybee, Apis mellifera". In: *Journal of Comparative Physiology A* 181.4, pp. 309–318.

King, Marcus J., Stephen L. Buchmann, and Hayward Spangler (1996). "Activity of asynchronous flight muscle from two bee families during sonication (buzzing)". In: *The Journal of Experimental Biology* 199.10, pp. 2317–2321. ISSN: 0022-0949. DOI: 10.1242/jeb.199.10.2317.

Kingma, Diederik P. and Jimmy Lei Ba (2015). *Adam: A Method for Stochastic Optimization*. Conference Paper.

Kirchner, Wolfgang H. (1993). "Acoustical Communication in Honeybees". In: *Apidologie* 24, pp. 297–307.

Kirchner, Wolfgang H. and Kathrin Sommer (1992). "The dance language of the honeybee mutant diminutive wings". In: *Behavioral Ecology and Sociobiology* 30.3/4, pp. 181–184. ISSN: 0340-5443. DOI: 10.1007/BF00166701.

Klatt, Björn K. et al. (2014). "Bee pollination improves crop quality, shelf life and commercial value". In: *Proc Biol Sci* 281.1775, p. 20132440. ISSN: 1471-2954 (Electronic) 0962-8452 (Print) 0962-8452 (Linking). DOI: 10.1098/rspb.2013.2440. URL: https://www.ncbi.nlm.nih.gov/pubmed/24307669.

Klein, B. A. et al. (2008). "Caste-dependent sleep of worker honey bees". In: *J Exp Biol* 211.Pt 18, pp. 3028–40. ISSN: 0022-0949 (Print) 0022-0949 (Linking). DOI: 10.1242/jeb.017426. URL: https://www.ncbi.nlm.nih.gov/pubmed/18775940.

Kritzky, Gene (2015). *The tears of Re: Beekeeping in Ancient Egypt*. 1st ed. Oxford University Press, p. 159. ISBN: 9780199361380.

Kronenberg, Fredi and H. Craig Heller (1982). "Colonial thermoregulation in honey bees (Apis mellifera)". In: *J Comp Physiol B* 148, pp. 65–76.

Krupke, C. H. et al. (2012). "Multiple routes of pesticide exposure for honey bees living near agricultural fields". In: *PLoS One* 7.1, e29268. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0029268. URL: https://www.ncbi.nlm.nih.gov/pubmed/22235278.

Kulyukin, Vladimir, Sarbajit Mukherjee, and Prakhar Amlathe (2018). "Toward Audio Beehive Monitoring: Deep Learning vs. Standard Machine Learning in Classifying Beehive Audio Samples". In: *Applied Sciences* 8.9. ISSN: 2076-3417. DOI: 10.3390/app8091573.

Landau, H. J. (1967). "Sampling, Data Transmission, and the Nyquist Rate". In: *Proceedings of the IEEE* 55.10, pp. 1701–1706.

Lewis, Lee Anne, Stanley Scott Schneider, and Gloria Degrandi-Hoffman (2002). "Factors influencing the selection of recipients by workers performing vibration signals in colonies of the honeybee, Apis mellifera". In: *Animal Behaviour* 63.2, pp. 361–367. ISSN: 00033472. DOI: 10.1006/anbe.2001.1894.

Li, Z. et al. (2016). "Drone and Worker Brood Microclimates Are Regulated Differentially in Honey Bees, Apis mellifera". In: *PLoS One* 11.2, e0148740. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0148740. URL: https://www.ncbi.nlm.nih.gov/pubmed/26882104.

Liang, Andrew (2024). "Developing an AI-based Integrated System for Bee Health Evaluation". In: *arXiv*. DOI: arXiv:2401.09988v1.

Lin, W. et al. (2020). "Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment". In: *Sci Rep* 10.1, p. 20336. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/s41598-020-77264-y. URL: https://www.ncbi.nlm.nih.gov/pubmed/33230152.

Lindauer, Martin (1952). "Ein Beitrag zur Frage der Arbeitsteilung im Bienenstaat". In: *Zeitschrift für Vergleichende Physiologie* 34.4, pp. 299–345. ISSN: 0340-7594. DOI: 10.1007/BF00298048.

Liu, Bo et al. (2017). *Deep neural networks for high dimension, low sample size data*. Conference Paper.

Lund, Brady D. et al. (2023). "ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing". In: *Journal of the Association for Information Science and Technology* 74.5, pp. 570–581. ISSN: 2330-1635 2330-1643. DOI: 10.1002/asi.24750.

Lundberg, Scott and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems*.

Luz, Jederson S. et al. (2024). "Cepstral and Deep Features for Apis mellifera Hive Strength Classification". In: *Journal of Internet Services and Applications* 15.1, pp. 548–560. ISSN: 1869-0238. DOI: 10.5753/jisa.2024.4015.

Maralit, Alexander D. et al. (2023). *QueenBuzz: A CNN-based architecture for Sound Processing of Queenless Beehive Towards European Apis Mellifera Bee Colonies' Survivability*. Conference Paper. DOI: 10.1109/ICCoSITE57641.2023.10127739.

Mayack, C., M. E. Natsopoulou, and D. P. McMahon (2015). "Nosema ceranae alters a highly conserved hormonal stress pathway in honeybees". In: *Insect Mol Biol* 24.6, pp. 662–70. ISSN: 1365-2583 (Electronic) 0962-1075 (Linking). DOI: 10.1111/imb.12190. URL: https://www.ncbi.nlm.nih.gov/pubmed/26335565.

Mayack, C. and D. Naug (2009). "Energetic stress in the honeybee Apis mellifera from Nosema ceranae infection". In: *J Invertebr Pathol* 100.3, pp. 185–8. ISSN: 1096-0805 (Electronic) 0022-2011 (Linking). DOI: 10.1016/j.jip.2008.12.001. URL: https://www.ncbi.nlm.nih.gov/pubmed/19135448.

McFee, B. et al. (2024). *librosa/librosa: 0.10.2.post1 (0.10.2.post1)*. Computer Program. DOI: 10.5281/zenodo.4923181. URL: https://doi.org/10.5281/zenodo.4923181.

McFee, Brian et al. (2023). *librosa/librosa: 0.10.1 (0.10.1)*. Computer Program. DOI: 10.5281/zenodo.8252662. URL: https://doi.org/10.5281/zenodo.8252662.

McInnes, Leland (2018). *Frequently Asked Questions - Uniform Manifold Approximation and Projection*. Web Page.

McMinn-Sauder, Harper B. G. et al. (2024). "Next-generation colony weight monitoring: a review and prospectus". In: *Apidologie* 55.1. ISSN: 0044-8435 1297-9678. DOI: 10.1007/s13592-023-01050-8.

Meikle, W. G. and N. Holst (2015). "Application of continuous monitoring of honeybee colonies". In: *Apidologie* 46.1, pp. 10–22. ISSN: 0044-8435 1297-9678. DOI: 10.1007/s13592-014-0298-x.

Michelsen, A. (2003). "Karl von Frisch lecture. Signals and flexibility in the dance communication of honeybees". In: *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 189.3, pp. 165–74. ISSN: 0340-7594 (Print) 0340-7594 (Linking). DOI: 10.1007/s00359-003-0398-y. URL: https://www.ncbi.nlm.nih.gov/pubmed/12664092.

Michelsen, Axel, Wolfgang H. Kirchner, and Martin Lindauer (1986). "Sound and vibrational signals in the dance language of the honeybee, Apis mellifera". In: *Behavioural Ecology and Sociobiology* 18, pp. 206–212.

Michelsen, Axel et al. (1986). "The tooting and quacking vibration signals of honeybee queens: a quantitative analysis". In: *Journal of Comparative Physiology A* 158, pp. 605–611.

Michelsen, Axel et al. (1987). "The acoustic near field of a dancing honeybee". In: *Journal of Comparative Physiology A* 161, pp. 633–643.

Moore, D. (2001). "Honey bee circadian clocks: behavioral control from individual workers to whole-colony rhythms". In: *Journal of Insect Physiology* 47.8, pp. 843–857. ISSN: 0022-1910. DOI: 10.1016/S0022-1910(01)00057-9.

Murphy, Fiona Edwards et al. (2015). *Big brother for bees (3B) - Energy neutral platform for remote monitoring of beehive imagery and sound*. Conference Paper.

NASS, National Agricultural Statistics Service (2024a). *Honey Bee Colonies (August 2024)*. Government Document.

— (2024b). *Honey (March 2024)*. Government Document.

— (2025). *Honey (March 2025)*. Government Document.

Naug, Dhruba and Ann Gibbs (2009). "Behavioral changes mediated by hunger in honeybees infected with Nosema ceranae". In: *Apidologie* 40.6, pp. 595–599. ISSN: 0044-8435 1297-9678. DOI: 10.1051/apido/2009039.

Naumann, Ken, Mark L. Winston, and Keith N. Slessor (1993). "Movement of Honey Bee (Apis mellifera L.) Queen Mandibular Gland Pheromone in Populous and Unpopulous Colonies". In: *Journal of Insect Behavior* 6.2, pp. 211–223.

Naumann, Ken et al. (1991). "Production and transmission of honey bee queen (Apis mellifera L.) mandibular gland pheromone". In: *Behavioral Ecology and Sociobiology* 29.5, pp. 321–332. ISSN: 0340-5443. DOI: 10.1007/BF00165956.

Nearman, A. and D. vanEngelsdorp (2022). "Water provisioning increases caged worker bee lifespan and caged worker bees are living half as long as observed 50 years ago". In: *Sci Rep* 12.1, p. 18660. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/s41598-022-21401-2. URL: https://www.ncbi.nlm.nih.gov/pubmed/36376353.

Nieh, James C. (1993). "The stop signal of honey bees: reconsidering its message". In: *Behavioral Ecology and Sociobiology* 33.1, pp. 51–56. ISSN: 0340-5443. DOI: 10.1007/BF00164346.

— (1998). "The honey bee shaking signal: function and design of a modulatory communication signal". In: *Behavioral Ecology and Sociobiology* 42.1, pp. 23–36. ISSN: 0340-5443. DOI: 10.1007/s002650050408.

Nieh, James C. and Jürgen Tautz (2000). "Behaviour-locked signal analysis reveals weak 200-300 Hz comb vibrations during the honeybee waggle dance". In: *The Journal of Experimental Biology* 203.10, pp. 1573–1579. ISSN: 0022-0949. DOI: 10.1242/jeb.203.10.1573.

Nolasco, Inês et al. (2019). *Audio-based identification of beehive states*. Conference Paper. DOI: 10.1109/ICASSP.2019.8682981.

Ohtani, Takeshi and Tsutomu Kamada (1980). "Worker Piping: The Piping Sounds Produced by Laying and Guarding Worker Bees". In: *Journal of Agricultural Research* 19.3, pp. 154–163.

Okawa, Masaki et al. (2019). *Audio classification of bit-representation waveform*. Conference Paper. DOI: 10.48550/arXiv.1904.04364.

Orlowska, Agnieszka et al. (2021). *Honey Bee Queen Presence Detection from Audio Field Recordings using Summarized Spectrogram and Convolutional Neural Networks*. Conference Paper. DOI: 10.1007/978-3-030-96308-8_8.

Paffhausen, Benjamin H. et al. (2021). "The Electronic Bee Spy: Eavesdropping on Honeybee Communication via Electrostatic Field Recordings". In: *Front Behav Neurosci* 15, p. 647224. ISSN: 1662-5153 (Print) 1662-5153 (Electronic) 1662-5153 (Linking). DOI: 10.3389/fnbeh.2021.647224. URL: https://www.ncbi.nlm.nih.gov/pubmed/33994968.

Page, Robert E. (1992). "Genotypic variability for rates of behavioural development in worker honeybees (Apis mellifera L.)" In: *Behavioural Ecology* 3.2, pp. 173–180. DOI: https://doi.org/10.1093/beheco/3.2.173.

Pankiw, Tanya et al. (1998). "Queen mandibular gland pheromone influences worker honey bee (Apis mellifera L.) foraging ontogeny and juvenile hormone titers". In: *Journal of Insect Physiology* 44, pp. 685–692. DOI: 10.1016/S0022-1910(98)00040-7.

Papachristoforou, Alexandros et al. (2008). "High frequency sounds produced by Cyprian honeybeesApis mellifera cypriawhen confronting their predator, the Oriental hornetVespa orientalis". In: *Apidologie* 39.4, pp. 468–474. ISSN: 0044-8435 1297-9678. DOI: 10.1051/apido:2008027.

Paszke, Adam et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Conference Paper.

Pavlovic, R. et al. (2024). "Micronutrient Deficiency May Be Associated with the Onset of Chalkbrood Disease in Honey Bees". In: *Insects* 15.4. ISSN: 2075-4450 (Print) 2075-4450 (Electronic) 2075-4450 (Linking). DOI: 10.3390/insects15040269. URL: https://www.ncbi.nlm.nih.gov/pubmed/38667399.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pettis, J. S. et al. (2012). "Pesticide exposure in honey bees results in increased levels of the gut pathogen Nosema". In: *Naturwissenschaften* 99.2, pp. 153–8. ISSN: 1432-1904 (Electronic) 0028-1042 (Print) 0028-1042 (Linking). DOI: 10.1007/s00114-011-0881-1. URL: https://www.ncbi.nlm.nih.gov/pubmed/22246149.

Pettis, J. S. et al. (2013). "Crop pollination exposes honey bees to pesticides which alters their susceptibility to the gut pathogen Nosema ceranae". In: *PLoS One* 8.7, e70182. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0070182. URL: https://www.ncbi.nlm.nih.gov/pubmed/23894612.

Pratt, Stephen C. et al. (1996). "Worker piping associated with foraging in undisturbed queenright colonies of honey bees". In: *Apidologie* 27, pp. 13–20.

Ramsey, Michael, Martin Bencsik, and Michael I. Newton (2017). "Long-term trends in the honeybee 'whooping signal' revealed by automated detection". In: *PLoS One* 12.2, e0171162. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0171162. URL: https://www.ncbi.nlm.nih.gov/pubmed/28178291.

Ramsey, Michael-Thomas et al. (2020). "The prediction of swarming in honeybee colonies using vibrational spectra". In: *Sci Rep* 10.1, p. 9798. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/s41598-020-66115-5. URL: https://www.ncbi.nlm.nih.gov/pubmed/32546693.

Robinson, Gene E. (1986). "Regulation of honey bee age polyethism by juvenile hormone". In: *Behavioral Ecology and Sociobiology* 20, pp. 329–338.

Robinson, Gene E. et al. (1992). "Colony Integration in Honey Bees: Mechanisms of Behavioral Reversion". In: *Ethology* 90, pp. 336–348. ISSN: 0179-1613 1439-0310. DOI: 10.1111/j.1439-0310.1992.tb00844.x.

Robles-Guerrero, Antonio et al. (2023). "Comparative Study of Machine Learning Models for Bee Colony Acoustic Pattern Classification on Low Computational Resources". In: *Sensors* 23.1. ISSN: 1424-8220. DOI: 10.3390/s23010460.

Robles-Guerrero, Antonio et al. (2024). "Convolutional Neural Networks for Real Time Classification of Beehive Acoustic Patterns on Constrained Devices". In: *Sensors (Basel)* 24.19. ISSN: 1424-8220 (Electronic) 1424-8220 (Linking). DOI: 10.3390/s24196384. URL: https://www.ncbi.nlm.nih.gov/pubmed/39409424.

Rodríguez Ballesteros, Alberto et al. (2024). "Towards acoustic monitoring of bees: wingbeat sounds are related to species and individual traits". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 379.1904. ISSN: 0962-8436 1471-2970. DOI: 10.1098/rstb.2023.0111.

Roe, Paul et al. (2021). "The Australian Acoustic Observatory". In: *Methods in Ecology and Evolution* 12.10, pp. 1802–1808. ISSN: 2041-210X 2041-210X. DOI: 10.1111/2041-210x.13660.

Roffet-Salque, Mélanie et al. (2015). "Widespread exploitation of the honeybee by early Neolithic farmers". In: *Nature* 527.7577, pp. 226–30. ISSN: 1476-4687 (Electronic)

0028-0836 (Linking). DOI: 10.1038/nature15757. URL: https://www.ncbi.nlm.nih.gov/pubmed/26560301.

Ruvinga, Stenford et al. (2023). "Identifying Queenlessness in Honeybee Hives from Audio Signals Using Machine Learning". In: *Electronics* 12.7. ISSN: 2079-9292. DOI: 10.3390/electronics12071627.

Rösch, Gustav Adolf (1930). "Untersuchungen über die Arbeitsteilung im Bienenstaat. 2. Teil: Die Tätigkeiten der Arbeitsbienen unter experimentell veränderten Bedingungen". In: *Zeitschrift für Vergleichende Physiologie* 12.1, pp. 1–71. DOI: 10.1007/BF00339476.

Sagili, R. R. et al. (2018). "Honey bees consider larval nutritional status rather than genetic relatedness when selecting larvae for emergency queen rearing". In: *Sci Rep* 8.1, p. 7679. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking). DOI: 10.1038/s41598-018-25976-7. URL: https://www.ncbi.nlm.nih.gov/pubmed/29769574.

Salomon, Mor et al. (2012). "The role of tyramine and octopamine in the regulation of reproduction in queenless worker honeybees". In: *Naturwissenschaften* 99.2, pp. 123–131. ISSN: 1432-1904 (Electronic) 0028-1042 (Linking). DOI: 10.1007/s00114-011-0877-x. URL: https://www.ncbi.nlm.nih.gov/pubmed/22205342.

Saltykova, E. S., L. R. Gaifullina, and A. G. Nikolenko (2008). "A change of sensitivity threshold of Apis mellifera to action of pathogen at different periods of starvation". In: *Journal of Evolutionary Biochemistry and Physiology* 44.4, pp. 484–493. ISSN: 0022-0930 1608-3202. DOI: 10.1134/s0022093008040091.

Sandeman, David C., Jürgen Tautz, and Martin Lindauer (1996). "Transmission of Vibration across honeycombs and its detection by bee leg receptors". In: *The Journal of Experimental Biology* 199, pp. 2585–2594. DOI: 10.1242/jeb.199.12.2585.

Sarker, I. H. (2021). "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions". In: *SN Comput Sci* 2.6, p. 420. ISSN: 2661-8907 (Electronic) 2662-995X (Print) 2661-8907 (Linking). DOI: 10.1007/s42979-021-00815-1. URL: https://www.ncbi.nlm.nih.gov/pubmed/34426802.

Sasaki, K. and Y. Obara (2001). "Nutritional factors affecting the egg sex ratio adjustment by a honeybee queen". In: *Insectes Sociaux* 48.4, pp. 355–359. ISSN: 0020-1812. DOI: 10.1007/PL00001790.

Schmickl, T. and K. Crailsheim (2001). "Cannibalism and early capping: strategy of honeybee colonies in times of experimental pollen shortages". In: *J Comp Physiol A* 187.7, pp. 541–7. DOI: 10.1007/s003590100226. URL: https://www.ncbi.nlm.nih.gov/pubmed/11730301.

— (2002). "How honeybees (Apis mellifera L.) change their broodcare behaviour in response to non-foraging conditions and poor pollen conditions". In: *Behavioral Ecology and Sociobiology* 51.5, pp. 415–425. ISSN: 0340-5443 1432-0762. DOI: 10.1007/s00265-002-0457-3.

Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri (2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nat Methods* 9.7, pp. 671–5. ISSN: 1548-7105 (Electronic) 1548-7091 (Print) 1548-7091 (Linking). DOI: 10.1038/nmeth.2089. URL: https://www.ncbi.nlm.nih.gov/pubmed/22930834.

Schulz, David J., Zhi-Yong Huang, and Gene E. Robinson (1998). "Effects of colony food shortage on behavioral development in honey bees". In: *Behavioral Ecology and Sociobiology* 42.5, pp. 295–303. ISSN: 0340-5443. DOI: 10.1007/s002650050442.

Schulz, David J. et al. (2002). "Effects of colony food shortage on social interactions in honey bee colonies". In: *Insectes Sociaux* 49.1, pp. 50–55. ISSN: 0020-1812. DOI: 10.1007/s00040-002-8279-x.

Scofield, Hailey N. and Heather R. Mattila (2015). "Honey bee workers that are pollen stressed as larvae become poor foragers and waggle dancers as adults". In: *PLoS One* 10.4, e0121731. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0121731. URL: https://www.ncbi.nlm.nih.gov/pubmed/25853902.

Seeley, Thomas D. (1979). "Queen Substance Dispersal by Messenger Workers in Honeybee Colonies". In: *Behavioral Ecology and Sociobiology* 5, pp. 391–415.

— (1989a). "Social foraging in honey bees: how nectar foragers assess their colony's nutritional status". In: *Behavioral Ecology and Sociobiology* 24.3, pp. 181–199. ISSN: 0340-5443. DOI: 10.1007/BF00292101.

— (1989b). "The honey bee colony as a superorganism". In: *American Scientist* 77.6, pp. 546–553. ISSN: 0003-0996. URL: https://www.jstor.org/stable/27856005.

— (1992). "The tremble dance of the honey bee: message and meanings". In: *Behavioural Ecology and Sociobiology* 31, pp. 375–383.

Seeley, Thomas D. and Jürgen Tautz (2001). "Worker piping in honey bee swarms and its role in preparing for liftoff". In: *Journal of Comparative Physiology A* 187.8, pp. 667–676. DOI: 10.1007/s00359-001-0243-0. URL: https://www.ncbi.nlm.nih.gov/pubmed/11763965.

Seeley, Thomas D., Anja Weidenmüller, and Susanne Kühnholz (1998). "The shaking signal of the honey bee informs workers to prepare for greater activity". In: *Ethology* 104.1, pp. 10–26. ISSN: 0179-1613. DOI: 10.1111/j.1439-0310.1998.tb00026.x.

Sejdić, Ervin, Igor Djurović, and Jin Jiang (2009). "Time–frequency feature representation using energy concentration: An overview of recent advances". In: *Digital Signal Processing* 19.1, pp. 153–183. ISSN: 10512004. DOI: 10.1016/j.dsp.2007.12.004.

Shanahan, Maggie (2022). "Honey Bees and Industrial Agriculture: What Researchers are Missing, and Why it's a Problem". In: *J Insect Sci* 22.1. ISSN: 1536-2442 (Electronic) 1536-2442 (Linking). DOI: 10.1093/jisesa/ieab090. URL: https://www.ncbi.nlm.nih.gov/pubmed/35137135.

Sharif, Muhammad Zahid et al. (2020). "Soundscape Indices: New Features for Classifying Beehive Audio Samples". In: *Sociobiology* 67.4, pp. 566–571. ISSN: 2447-8067 0361-6525. DOI: 10.13102/sociobiology.v67i4.5860.

Simpson, J. (1964). "The mechanism of honey-bee queen piping". In: *Zeitschrift für vergleichende Physiologie* 48, pp. 277–282.

Skaggs, R. et al. (2014). "The possible role of ritualized aggression in the vibration signal of the honeybee, Apis mellifera". In: *Animal Behaviour* 98, pp. 103–111. ISSN: 00033472. DOI: 10.1016/j.anbehav.2014.09.030.

Smith, Michael L. and Po-Cheng Chen (2017). "Larger but not louder: bigger honey bee colonies have quieter combs". In: *Behavioral Ecology and Sociobiology* 71.11. ISSN: 0340-5443 1432-0762. DOI: 10.1007/s00265-017-2399-9.

Song, Y. et al. (2024). "Experimental Investigation on the Acoustic Insulation Properties of Filled Paper Honeycomb-Core Wallboards". In: *Biomimetics (Basel)* 9.9. ISSN: 2313-7673 (Electronic) 2313-7673 (Linking). DOI: 10.3390/biomimetics9090528. URL: https://www.ncbi.nlm.nih.gov/pubmed/39329550.

Stabe, Henry A. (1929). "The rate of growth of worker, drone and queen larvae of the honeybee, Apis mellifera Linn." Thesis.

Stabler, Daniel et al. (2021). "Regulation of dietary intake of protein and lipid by nurse-age adult worker honeybees". In: *J Exp Biol* 224.Pt 3. ISSN: 1477-9145 (Electronic) 0022-0949 (Print) 0022-0949 (Linking). DOI: 10.1242/jeb.230615. URL: https://www.ncbi.nlm.nih.gov/pubmed/33443043.

Sugai, Larissa Sayuri Moreira et al. (2019). "Terrestrial Passive Acoustic Monitoring: Review and Perspectives". In: *BioScience* 69.1, pp. 15–25. DOI: 10.1093/biosci/biy147.

Sullivan, Joseph P. et al. (2000). "Juvenile hormone paces behavioral development in the adult worker honey bee". In: *Hormones and Behaviour* 37.1, pp. 1–14. ISSN: 0018-506X (Print) 0018-506X (Linking). DOI: 10.1006/hbeh.1999.1552. URL: https://www.ncbi.nlm.nih.gov/pubmed/10712853.

Szentgyörgyi, Hajnalka, Krystyna Czekońska, and Adam Tofilski (2017). "The Effects of Starvation of Honey Bee Larvae on Reproductive Quality and Wing Asymmetry of Honey Bee Drones". In: *Journal of Apicultural Science* 61.2, pp. 233–243. ISSN: 2299-4831. DOI: 10.1515/jas-2017-0018.

Tan, Ken et al. (2016). "Honey Bee Inhibitory Signaling Is Tuned to Threat Severity and Can Act as a Colony Alarm Signal". In: *PLoS Biol* 14.3, e1002423. ISSN: 1545-7885 (Electronic) 1544-9173 (Print) 1544-9173 (Linking). DOI: 10.1371/journal.pbio.1002423. URL: https://www.ncbi.nlm.nih.gov/pubmed/27014876.

Tang, Binhua et al. (2019). "Recent Advances of Deep Learning in Bioinformatics and Computational Biology". In: *Front Genet* 10, p. 214. ISSN: 1664-8021 (Print) 1664-8021 (Electronic) 1664-8021 (Linking). DOI: 10.3389/fgene.2019.00214. URL: https://www.ncbi.nlm.nih.gov/pubmed/30972100.

Tautz, Jürgen (1996). "Honeybee waggle dance: Recruitment success depends on the dance floor". In: *The Journal of Experimental Biology* 199.6, pp. 1375–1381. ISSN: 0022-0949. DOI: 10.1242/jeb.199.6.1375.

— (2009). "The largest organ of the bee colony - Construction and function of the comb". In: *The buzz about bees: Biology of a superorganism*. Berlin/Heidelberg: Springer.

Tautz, Jürgen, J. Casas, and D. Sandeman (2001). "Phase reversal of vibratory signals in honeycomb may assist dancing honeybees to attract their audience". In: *The Journal of Experimental Biology* 204, pp. 3737–3746.

Tautz, Jürgen and Martin Lindauer (1997). "Honeybees establish specific sites on the comb for their waggle dances". In: *Journal of Comparative Physiology A* 180, 537–539.

Tautz, Jürgen, Kristin Rohrseitz, and David C. Sandeman (1996). "One-strided waggle dance in bees". In: *Nature* 382.6586, p. 1. ISSN: 0028-0836. DOI: 10.1038/382032a0.

Terenzi, Alessandro et al. (2019). *Features extraction applied to the analysis of the sounds emitted by honey bees in a beehive*. Conference Paper.

Terenzi, Alessandro et al. (2022). *Comparison of feature extraction methods for sound-based classification of honey bee activity*. Conference Paper.

The pandas development team (2020). *pandas-dev/pandas: Pandas*. Computer Program. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Thom, Corinna, David C. Gilley, and Jürgen Tautz (2003). "Worker piping in honey bees (Apis mellifera): the behavior of piping nectar foragers". In: *Behavioral Ecology and Sociobiology* 53.4, pp. 199–205. ISSN: 0340-5443 1432-0762. DOI: 10.1007/s00265-002-0567-y.

Thompson, Jim (2016). "Installing A Queen". In: *Bee Culture* 144.2, pp. 60 –64.

Thompson, Neil C. et al. (2021). "Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable". In: *IEEE Spectrum* 58.10, pp. 50–55. DOI: 10.1109/MSPEC.2021.9563954.

Thurm, Ulrich (1964). "Mechanoreceptors in the Cuticle of the Honey Bee: Fine Structure and Stimulus Mechanism". In: *Science* 145.3636, pp. 1063–1065. URL: https://www.jstor.org/stable/1714100.

Tlačbaba, Jaromír et al. (2014). "The Acoustic Emission in the Nest of the Honey Bee Depending on the Extreme Weather Conditions". In: *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis* 62.1, pp. 245–254. ISSN: 12118516 24648310. DOI: 10.11118/actaun201462010245.

Torres, D. J., U. M. Ricoy, and S. Roybal (2015). "Modeling Honey Bee Populations". In: *PLoS One* 10.7, e0130966. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0130966. URL: https://www.ncbi.nlm.nih.gov/pubmed/26148010.

Toth, Amy L. et al. (2005). "Nutritional status influences socially regulated foraging ontogeny in honey bees". In: *Journal of Experimental Biology* 208.Pt 24, pp. 4641–9.

ISSN: 0022-0949 (Print) 0022-0949 (Linking). DOI: 10.1242/jeb.01956. URL: https://www.ncbi.nlm.nih.gov/pubmed/16326945.

Towne, William F. and Wolfgang H. Kirchner (1989). "Hearing in Honey Bees: Detection of Air-Particle Oscillations". In: *Science* 244.

Tsujiuchi, S. et al. (2007). "Dynamic range compression in the honey bee auditory system toward waggle dance sounds". In: *PLoS One* 2.2, e234. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking). DOI: 10.1371/journal.pone.0000234. URL: https://www.ncbi.nlm.nih.gov/pubmed/17311102.

Utaipanon, Patsavee, Michael J. Holmes, and Benjamin P. Oldroyd (2019). "Queenless colonies contribute to the male breeding population at honey bee drone congregation areas". In: *Insectes Sociaux* 66.4, pp. 593–599. ISSN: 0020-1812 1420-9098. DOI: 10.1007/s00040-019-00720-0.

Van Rossum, Guido and Fred L. Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. ISBN: 1441412697.

Vance, J. T. et al. (2014). "Hovering flight in the honeybee Apis mellifera: kinematic mechanisms for varying aerodynamic forces". In: *Physiol Biochem Zool* 87.6, pp. 870–81. ISSN: 1537-5293 (Electronic) 1522-2152 (Linking). DOI: 10.1086/678955. URL: https://www.ncbi.nlm.nih.gov/pubmed/25461650.

Vinuesa, Ricardo et al. (2014). "Opportunities for machine learning in scientific discovery". In: *arXiv*, pp. 1–22.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2. URL: https://rdcu.be/b08Wh.

Wahl, Richard (2023). "Off the Wahl Beekeeping: Mechanics of a Hive Inspection". In: *Bee Culture* 151.8, pp. 52–54. URL: https://www.proquest.com/trade-journals/off-wahl-beekeeping-mechanics-hive-inspection/docview/2847079808/se-2.

Wang, Mei and Weihong Deng (2021). "Deep face recognition: A survey". In: *Neurocomputing* 429, pp. 215–244. ISSN: 09252312. DOI: 10.1016/j.neucom.2020.10.081.

Wang, Ying et al. (2016). "Starvation stress during larval development facilitates an adaptive response in adult worker honey bees (Apis mellifera L.)" In: *J Exp Biol*

219.Pt 7, pp. 949–59. ISSN: 1477-9145 (Electronic) 0022-0949 (Linking). DOI: 10.1242/jeb.130435. URL: https://www.ncbi.nlm.nih.gov/pubmed/27030775.

Waskom, Michael (2021). "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60. ISSN: 2475-9066. DOI: 10.21105/joss.03021.

Weiss, K. (1984). "Regulierung des Proteinhaushaltes im Bienenvolk (Apis mellifica L.) durch Brutkannibalismus". In: *Apidologie* 15.3, pp. 339–354.

Winston, M. L. (1979). "Events following queen removal in colonies of africanized honeybees in South America". In: *Insectes Soc* 26.4, pp. 373–381.

— (1987). *The Biology of the Honey Bee*. Harvard University Press. ISBN: 9780674074095.

Woyke, J. (1977). "Cannibalism and Brood-Rearing Efficiency in the Honeybee". In: *Journal of Apicultural Research* 16.2, pp. 84–94. DOI: https://doi.org/10.1080/00218839.1977.11099866.

Wray, Margaret K., Heather R. Mattila, and Thomas D. Seeley (2011). "Collective personalities in honeybee colonies are linked to colony fitness". In: *Animal Behaviour* 81.3, pp. 559–568. ISSN: 00033472. DOI: 10.1016/j.anbehav.2010.11.027.

Wright, Geraldine A., Susan W. Nicolson, and Sharoni Shafir (2018). "Nutritional Physiology and Ecology of Honey Bees". In: *Annual Review of Entomology* 63.1, pp. 327–344. DOI: 10.1146/annurev-ento-020117-043423.

Ying, Xue (2019). "An overview of overfitting and its solutions". In: *Journal of Physics: Conference series* 1168. DOI: 10.1088/1742-6596/1168/2/022022.

Zacepins, Aleksejs et al. (2016). "Remote detection of the swarming of honey bee colonies by single-point temperature monitoring". In: *Biosystems Engineering* 148, pp. 76–80. ISSN: 15375110. DOI: 10.1016/j.biosystemseng.2016.05.012.

Zaman, Asaduz and Alan Dorin (2023). "A framework for better sensor-based beehive health monitoring". In: *Computers and Electronics in Agriculture* 210. ISSN: 01681699. DOI: 10.1016/j.compag.2023.107906.

Zhu, Y. et al. (2024). "MSPB: a longitudinal multi-sensor dataset with phenotypic trait measurements from honey bees". In: *Sci Data* 11.1, p. 860. ISSN: 2052-4463 (Electronic) 2052-4463 (Linking). DOI: 10.1038/s41597-024-03695-1. URL: https://www.ncbi.nlm.nih.gov/pubmed/39122730.

4539 Łopuch, Sylvia and Adam Tofilski (2019). "Use of high-speed video recording to detect
4540 wing beating produced by honey bees". In: *Insectes Sociaux* 66.2, pp. 235–244. ISSN:
4541 0020-1812 1420-9098. DOI: 10.1007/s00040-018-00678-5.
4542 Łopuch, Sylwia and Adam Tofilski (2017). "Direct Visual Observation of Wing Move-
4543 ments during the Honey Bee Waggle Dance". In: *Journal of Insect Behavior* 30.2,
4544 pp. 199–210. ISSN: 0892-7553 1572-8889. DOI: 10.1007/s10905-017-9610-8.