University of Southampton

Faculty of Natural and Environmental Science

School of Ocean and Earth Sciences

Computer Vision and Explainable AI for Morphological Interpretation: Applications in Taxonomy, Cryptic Species, and Natural History Collections

by

Jack Daniel Hollister MRes BSc

ORCID ID 0000-0001-8697-5076

Thesis for the degree of Doctor of Philosophy

September 2025

"Sleep late, have fun, get wild, drink whisky and drive fast on empty streets with nothing in mind but falling in love and not getting arrested." ~ Hunter S. Thompson

University of Southampton

Abstract

Faculty of Natural and Earth Science

Doctor of Philosophy

Computer Vision and Explainable AI for Morphological Interpretation: Applications in Taxonomy, Cryptic Species, and Natural History Collections

by

Jack Daniel Hollister

Taxonomy provides the foundation for ecology, evolution, and conservation, but it continues to face significant challenges. Morphology-based approaches are hampered by issues such as cryptic similarity, phenotypic plasticity, and the global decline in taxonomic expertise. Molecular methods, while precise in detecting genetic divergence, often fail to identify the visible traits that separate lineages. This persistent disconnect between genotype and phenotype limits the integration of molecular and morphological evidence in biodiversity research. The rapid digitisation of natural history collections now provides unprecedented image datasets, but their scale demands automated, interpretable approaches for morphological analysis. This thesis investigates the application of computer vision and explainable artificial intelligence to the interpretation of morphology across species, populations, and collections. The central aim is to evaluate whether computer vision models, when coupled with heatmap-based interpretability methods, can both achieve high classification accuracy and reveal biologically meaningful traits in ways transparent to human experts.

Four empirical studies demonstrate this potential. Chapter 2 trained convolutional neural networks to classify cryptic limpets from the Baja California peninsula, achieving accuracies marginally higher than expert taxonomists. Heatmaps confirmed that models attended to diagnostic shell features. Chapter 3 developed a computer vision pipeline to detect mislabelled butterfly specimens in the digitised Lepidoptera collection of the Natural History Museum, London, revealing many potential errors where many were verified by experts and genetics. Chapter 4 applied computer vision to genetically divergent but morphologically similar limpet clades, showing that saliency maps and shape analyses could expose subtle, previously overlooked shell differences aligned with phylogeographic structure. Chapter 5 compared

human and machine attention in the classification of British butterflies, demonstrating substantial overlap between heatmap focus and diagnostic traits in identification guides, while also highlighting features not mentioned in the literature and attention to novel artefacts. Together, these studies show that explainable computer vision can complement and extend traditional taxonomy: accelerating large-scale identifications, flagging curation errors, and uncovering cryptic morphological divergence. More broadly, they position interpretable artificial intelligence as a practical instrument for biodiversity science, one that can bridge molecular and morphological evidence, strengthen the curation of natural history collections, and provide reproducible baselines for the study and conservation of biodiversity.

Table of Contents

Table of Tables

Table of Figures

Table of Figures

Table of Figures

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Jack Daniel Hollister

Title of thesis: Computer Vision and Explainable AI for Morphological Interpretation: Applications in Taxonomy, Cryptic Species, and Natural History Collections

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

HOLLISTER, J. D., CAI, X., HORTON, T., PRICE, B. W., ZARZYCZNY, K. M. & FENBERG, P. B. 2023. Using computer vision to identify limpets from their shells: a case study using four species from the Baja California peninsula. Frontiers in Marine Science, 10.

HOLLISTER, J. D., MARTIN, G., CAI, X., HORTON, T., POWELL, O., STERLING, M., TURNBULL, G., PRICE, B. W. & FENBERG, P. B. 2025. A Computer Vision Method for Finding Mislabelled Specimens Within Natural History Collections. Ecology and Evolution, 15, e71648.

Signature: ..... Date: 11.02.2026

Acknowledgements

First and foremost, I would like to thank my supervisory team: Dr Phillip Fenberg, Dr Tammy Horton, Dr Ben Price, and Dr Xiaohao Cai. Their mentorship, guidance, and constant support have been a continual source of inspiration throughout my PhD. Each of them, experts in their respective fields, and completely different to each other, has taught and guided me enormously over the past four years. Their input has enabled me to complete this multidisciplinary thesis. A particular and heartfelt thank you goes to my main supervisor, Phil, who has worked most closely with me. He has always been my first point of call whenever I encountered challenges in my work or elsewhere. Together we have travelled to several countries, shared many great meals and drank countless coffees, and I am grateful to call him not only my supervisor but also a good friend. I would also like to thank the University of Southampton and the NERC Doctoral Training Programme. Without this programme, my life would not be where it is today. It has given me opportunities to travel, to learn, to grow, and to secure employment. I am deeply proud of what I have achieved as a result. My thanks also go to the Natural History Museum, London. As part of my PhD programme, the Museum provided world-leading experience, and I was fortunate to work with so many people across different departments that it would be impossible to name them all. However, I would especially like to thank Geoff Martin for his time and dedication in supporting several of my chapters. I would next like to acknowledge DDI, and the AI and Innovation departments at the NHM. During my PhD, I was able to secure two job roles within these teams. Both were possible thanks to the skills and knowledge I developed during my doctoral research, and both helped me in turn to complete my thesis, offering professional experience that a PhD alone could never provide. I am proud that as I hand in this thesis, I do so while continuing a position in a full-time role, as I have for the past year and a half. Within these teams, I want to thank Dr Vince Smith and Ben Scott, my managers and heads of departments, who recognised my skills and welcomed me despite my competing PhD commitments. I also thank my colleagues, Dr Arriana Salli-James and Dr Sanson Poon, for their constant support throughout both roles and the final stages of my doctorate. I would like to thank my wife, Hettie, as well as my friends and family, for their unwavering support. Their encouragement, patience, and willingness to listen to me ramble about my work involving computers, insects, and many other creatures they probably never cared to know about have meant more than I can say. Finally, I would like to say a big thank you to my Dog Clemmy, who has kept me company whenever I worked at home, writing and researching away, and forcing me to go outside for some fresh air. Some of my best ideas and breakthroughs have come to me at times on walks around the neighbourhood and while sitting in the park with you.

Acknowledgements

There are lots of people I have missed of this list, but just because I haven't mentioned you, doesn't mean you haven't had a positive effect on my journey, and you all have my appreciation.

Thank you all,

Much Love,

Jack

Definitions and Abbreviations

[CV ..... Computer vision, a field of computer science that enables machines to interpret and analyse visual information from images or videos.

AI ..... Artificial Intelligence, is the use of computers and algorithms to perform tasks that normally require human intelligence, such as recognising patterns, making decisions, or learning from data.

XAI ..... Explainable Artificial Intelligence, is artificial intelligence that makes its decisions understandable to humans by showing how and why it reached a result.

NHC ..... Natural history collections, organised archives of preserved specimens, fossils, and associated data that document the diversity of life and Earth's history.

NHM ..... Natural history museums are institutions that preserve, research, and display specimens and artefacts from the natural world to advance science and share knowledge with the public.

DL ..... Deep learning, a type of machine learning that uses multi-layered neural networks to automatically learn patterns and representations from large amounts of data.

CNN ..... Convolutional Neural Networks**,** a kind of deep learning model designed to analyse images by detecting features such as edges, shapes, and textures through specialised convolutional layers.]

Chapter 1 Introduction

1.1 Biodiversity, Morphology, and Taxonomy

Biodiversity research relies on accurate and reproducible systems of organism-based classification (Kürzel et al., 2022). Taxonomy provides the framework through which species and higher-level groups are described, compared, classified, and understood, forming the foundation for ecology, evolution, and conservation biology (Dayrat, 2005). Its influence, however, extends far beyond these fields. Accurate taxonomic knowledge underpins environmental sciences, where species serve as indicators of ecosystem health and past environmental conditions (Bortolus, 2008, Hollister et al., 2022). It is fundamental to agriculture and food security, ensuring pests, pollinators, and pathogens are correctly identified (Lughadha, 2004). In public health, taxonomy guides the control of disease vectors and parasites (Wilcox and Ellis, 2006). Even cultural and socio-economic practices, such as traditional medicine, fisheries, and wildlife trade, are framed by the ability to correctly recognise and name species (Mace, 2004).

Traditionally, morphological traits, such as body shape, colour patterns, or structural features, have been the primary means of distinguishing between taxa. These characters remain invaluable because they are visible, measurable, and accessible in both contemporary and historical specimens. However, species can be highly variable across their geographic ranges, developmental stages, or ecological contexts, while in other cases outward appearance remains deceptively uniform despite underlying genetic divergence (Zarzyczny et al., 2024, Mesnick and Ralls, 2018, Hebert et al., 2004, Hending, 2025). Such complexities contribute to long-standing taxonomic uncertainty and can complicate biodiversity assessments, ecological monitoring, and conservation planning.

1.2 Morphological Challenges in Taxonomy

The diverse morphology of life on earth presents taxonomists with a wide range of challenges that can obscure true species boundaries or inflate apparent diversity. These include cryptic similarity, plasticity, sexual dimorphism, ontogenetic change, polymorphism, convergence, hybridisation, aberrations, and environmental artefacts. Each poses distinct difficulties for reliable classification.

- **Cryptic similarity -** Occurs when multiple genetically distinct lineages appear morphologically indistinguishable. Such cases are widespread across plants, animals, and fungi, where external characters provide little signal of underlying divergence (Bickford et al., 2007) which can result in many species being masked as a single species. For instance, a species of butterfly, *Asraptes fulgerator* (Walch, J.E.I., 1775), was originally described visually in 1775, however was found out to be a complex of 10 different species in 2004 (Hebert et al., 2004). One study estimated that for every species identified through morphology alone, there are an average three cryptic species hidden within it (Pfenninger and Schwenk, 2007). Such hidden diversity contributes to the Linnean shortfall which is the gap between the number of species formally described and the true number of species on Earth, thereby leading to systematic underestimation of global biodiversity (Brito, 2010, Walters et al., 2021).
- **Phenotypic plasticity** - Represents the inverse challenge to cryptic similarity, where individuals of a single species display highly variable morphologies. Classic examples include shell shape variation in intertidal molluscs exposed to differing wave regimes (Trussell, 1996) and leaf form differences in plants grown under contrasting light conditions (Sultan, 2000). Plasticity can blur the boundary between adaptive form and heritable character, risking over-splitting, causing a misinterpretation of diversity.
- **Sexual dimorphism** - Males and females of the same species may differ from each other. Sexual dimorphism occurs across many groups, from plumage in birds (Dale et al., 2015) to wing shape and colour in butterflies (Dharmaraaj and Kunte, 2025). Misinterpreting sexes as separate taxa has been shown to cause inflated species numbers (Isaac et al., 2004).
- **Ontogenetic change** - Morphological differences across developmental stages can also obscure taxonomy. Larval, juvenile, and adult forms of many insects and amphibians are so distinct that they have been described as separate taxa when life histories are unknown (Barton, 2024).
- **Polymorphism** - The occurrence of multiple stable, genetically determined forms within a species, and environmentally induced seasonal plasticity both complicate taxonomy. Colour morphs (Livraghi et al., 2025) and seasonal wing forms in butterflies triggered by photoperiod or temperature (Cabon et al., 2025) illustrate how one species can present as several morphologically distinct entities. Again, can overinflate diversity estimates.
- **Convergent evolution** - Generates morphological similarity between unrelated lineages occupying comparable ecological niches. Such convergence can easily mislead morphology-based classification by masking true evolutionary relationships. For

example, the mimicry rings of butterflies, where distantly related species can evolve near-identical wing colouration and patterning to share protection from predators (Mallet and Joron, 1999, Merrill et al., 2015) or the repeated evolution of limpet-shaped shells across unrelated gastropod lineages, driven by adaptation to wave-swept intertidal habitats (Vermeij, 2002, Urdy et al., 2010). In both instances, similar selective pressures result in analogous morphologies that can obscure phylogenetic boundaries and complicate taxonomy when assessments are based on form alone.

- **Hybridisation and introgression** – Can produce individuals with mosaic morphologies intermediate between parent species. These can be mistaken for novel taxa or obscure true boundaries, as documented in plants and many vertebrate groups (Mallet, 2005).
- **Aberrations** - Rare morphological anomalies within individuals that can further confuse taxonomy. Aberrant forms, often caused by developmental instability, disease, or injury, may not represent heritable characters but can be misinterpreted as diagnostic features (Nijhout, 1986).
- **Environmental wear and preservation artefacts** - Can also be significant in diagnoses. In molluscs, shell erosion can obscure key features (Akpan and Farrow, 1985), while in natural history collections (NHC), faded pigments or damaged can complicate species recognition (Hendry et al., 2016). Fossil specimens, likewise, are often altered by erosion that can erase or distort morphological traits (Behrensmeyer et al., 2000).

Taken together, these challenges illustrate why morphology, although central to taxonomy, can provide an incomplete or misleading picture of biodiversity.

1.3 Visual-Based Taxonomy

For much of its history, taxonomy has relied on visual examination of specimens. Identification through external morphology, supported by field guides, dichotomous keys, and expert comparison, remains central to taxonomic practice (Trail, 2021, Bik, 2017). This approach has the advantages of being non-destructive, immediately accessible, and directly tied to the diagnostic characters used in formal species descriptions. As a result, visual-based taxonomy continues to underpin ecological surveys, museum curation, and biodiversity monitoring and has done so for hundreds of years.

However, visual identification is also prone to error. Even experienced taxonomists may misidentify specimens when working with groups that are morphologically variable or cryptic (Hollister et al. 2023; Hollister et al. 2025). They are also prone to human-based errors such as

suffering from fatigue, lapses in concentration, cognitive bias, or reliance on incomplete prior knowledge (Wäldchen et al., 2022, Behrens et al., 2023). Studies have shown that accuracy can be far lower than often assumed. For example, controlled trials with insect identifications reported overall accuracies below 60% even for expert observers (Austen et al., 2016). Similarly, large-scale comparisons with molecular benchmarks revealed only about 71% agreement with visual identifications (Meyer and Paulay, 2005).

The effectiveness of visual taxonomy is also strongly influenced by practical factors such as the quality of available keys (Wäldchen et al., 2022), and the condition of reference collections (Vollmar et al., 2010). Specimen throughput can be another limiting factor: careful examination is time-intensive, particularly for large-scale digitisation projects or ecological surveys generating thousands of specimens (Nelson and Ellis, 2019). There are also funding issues, which inevitably impact staffing in many collections across the world (Paknia et al., 2015). Additionally, there is a decline in taxonomic expertise worldwide. For many groups, only a handful of active specialists remain (Agnarsson and Kuntner, 2007), expertise is unevenly distributed across taxa and regions, leaving large numbers of species without dedicated experts (Ebach et al., 2011, Troudet et al., 2017). At the same time, university training programmes and collection-based teaching have contracted, offering fewer opportunities for students to develop skills in morphology-based identification (Agnarsson and Kuntner, 2007). Yet the demand for accurate identifications continues to increase, driven by accelerating biodiversity surveys (Cerrejón et al., 2025), the rapid expansion of citizen science initiatives (Jansen et al., 2024), and the scaling-up of large museum digitisation projects worldwide (Yap et al., 2024).

The mismatch between rising demand and shrinking expertise increases the risk of misidentifications, slows the pace of research, and weakens the foundations upon which biodiversity science depends. These limitations underscore the need for complementary approaches that retain the interpretability of morphology while addressing the bottlenecks of human subjectivity and limited workforce capacity.

1.4 Molecular Approaches and Their Limitations

The advent of molecular methods reshaped taxonomy and biodiversity science. The introduction of DNA barcoding demonstrated that short, standardised gene regions could reliably distinguish species across diverse animal groups (Hebert et al., 2003), and subsequent work showed how barcoding complements traditional taxonomy, molecular phylogenetics, and population genetics (Hajibabaei et al., 2007). More recently, genome-wide SNP analyses have provided high-resolution insights into population structure and adaptive variation (Nielsen et al., 2009,

Seehausen et al., 2014, Luikart et al., 2018). These approaches offer an objective and replicable framework for classification, often revealing hidden diversity where morphology alone proves insufficient. Molecular data have been central in clarifying long-standing taxonomic uncertainties, confirming or rejecting species validity, and reshaping our understanding of evolutionary relationships across a wide range of taxa (Dayrat, 2005, Padial et al., 2010). The contributions of molecular approaches are substantial. They have uncovered cryptic species complexes (Hebert et al., 2004), resolved nomenclatural confusion (Puillandre et al., 2012), and provided robust insights into phylogenetic and population-level patterns (Avise, 2009, Nielsen et al., 2009). Increasingly, genetic data are also incorporated into biodiversity surveys, helping to standardise identifications and allowing researchers to detect diversity that would otherwise remain unrecognised through morphology alone (Taberlet et al., 2012, Cristescu, 2014).

Despite the impact, molecular approaches also face important limitations. Generating sequence data requires specialised equipment, technical expertise, and financial resources that are not equally available across regions or institutions (Will et al., 2005, Dayrat, 2005). Laboratory errors, contamination, and inconsistent standardisation can compromise results, while reference databases may themselves contain misidentified specimens which will all feed down to researchers (Cheng et al., 2023). Scaling these methods to large collections of specimens, typically housed in NHCs remains challenging, as processing is time-consuming and destructive sampling is not always feasible (Raxworthy and Smith, 2021).

Perhaps most significantly, molecular methods often expose a gap between genotype and phenotype. While they can demonstrate that taxa are genetically distinct, they provide limited insight into how they differ morphologically, ecologically, or functionally (Dayrat, 2005, Karbstein et al., 2024b). As a result, genetic clusters can appear abstract, disconnected from the visible traits that shape ecological interactions and adaptive potential (Deng et al., 2023, Baird et al., 2024). For biodiversity research, conservation, and taxonomy, it is not enough to know that species are distinct, it is also necessary to interpret the traits that distinguish them. There has been progress in explaining morphology through genetics, for example, in studies demonstrating heritable components of wing morphology in grasshoppers (*Pseudochorthippus parallelus*) (Cabon et al., 2025) and in mapping genetic loci underlying yellow–orange wing colour variation in sulphur butterflies (*Colias* spp.) (Hanly et al., 2023).

Molecular methods are therefore invaluable but incomplete as stand-alone tools. Their integration with morphological and ecological approaches offers the greatest potential: combining genetic evidence of divergence with phenotypic data to reveal how differences manifest in form and function (Padial et al., 2010). NHCs are central to this effort, as they provide the physical specimens that anchor molecular sequences to observable traits and ecological metadata (Raxworthy and Smith, 2021). By linking genetic clusters to voucher specimens, these

collections allow researchers to bridge the genotype–phenotype gap and build a more complete understanding of biodiversity.

1.5 Natural History Collections and Digitisation

NHCs represent one of the most significant archives of global biodiversity. Housing upwards of 3 billion specimens across museums, herbaria, and universities worldwide, these collections document the morphology, distribution, and diversity of life across space and time (Ariño, 2010, Paul et al., 2025). They underpin taxonomy by providing type specimens, but also serve as irreplaceable resources for ecological, evolutionary, and conservation research. By preserving material collected over centuries, they allow researchers to examine how species and ecosystems have changed in response to environmental pressures, human activity, and global change (Meineke et al., 2018). The scientific value of these collections is extensive. Specimens provide historical baselines against which contemporary biodiversity can be compared, revealing shifts in species distributions, abundance, and morphology (Shaffer et al., 1998, Lister, 2011). They enable the study of long-term evolutionary and ecological dynamics, including trait evolution, phenotypic plasticity, and the presence of cryptic diversity (Holmes et al., 2016, Rowe et al., 2011). Crucially, they also link past and present datasets, connecting field-collected material with genomic, ecological, and morphological data in ways that support integrative biodiversity research (Raxworthy and Smith, 2021, Cook et al., 2014).

The digitisation of NHCs typically follows a structured photographic workflow designed to maximise reproducibility, image quality, and downstream usability. Specimens are prepared and positioned in a consistent orientation relative to the camera, often using standardised backgrounds and colour charts to minimise visual noise and support later data correction (Brecko & Mathys 2020). Image capture is conducted under controlled lighting conditions, commonly employing diffuse, colour-balanced illumination to reduce shadows, glare, and specular reflections, while fixed camera settings (aperture, exposure, white balance and focus) are maintained across sessions to ensure consistency between specimens (Brecko & Mathys 2020). Calibration targets for scale and colour are frequently included to allow for post hoc correction and comparability across datasets. Images are captured at sufficiently high resolution to preserve diagnostically relevant features, and metadata describing capture conditions, equipment and specimen identifiers are recorded alongside the image files. Quality control steps are applied to identify blurred, poorly exposed, or inconsistently framed images prior to ingestion into collection management systems. Collectively, these practices aim to produce standardised digital surrogates that are suitable not only for archival and access purposes but also for

quantitative analyses, including computer vision (CV)–based approaches, where variation introduced during imaging can otherwise confound biological signal (Brecko & Mathys 2020; European Commission 2019; Nelson & Ellis 2019).

In recent decades, large-scale digitisation initiatives have begun to revolutionise access to these collections. Programmes such as iCollections in the UK, DiSSCo in Europe, and the NSF ADBC programme in the United States have produced millions of high-resolution images of pinned insects, mollusc shells, herbarium sheets, and microscopic slides (Blagoderov et al., 2012, Hardisty et al., 2020, Nelson and Ellis, 2019). These digital surrogates make specimens globally accessible and provide the raw material for computational approaches such as computer vision (CV) (Beaman and Cellinese, 2012). In parallel, online aggregators and citizen science platforms such as GBIF and iNaturalist have expanded availability still further, creating biodiversity datasets of unprecedented scale (Nugent, 2018, Heberling and Isaac, 2018). Despite these advances, challenges remain. Historical misidentifications are common, and errors embedded in collections can propagate through digital datasets (Goodwin et al., 2015). Metadata are often incomplete, and imaging standards vary between institutions (Nelson and Ellis, 2019). Many collections face backlogs, with vast numbers of specimens still yet to be digitised or unstudied (Hedrick et al., 2020).

Digitisation has unlocked extraordinary potential, but human-led identification and curation cannot keep pace with the scale of material available. This creates a pressing need for automated, scalable methods that can process, validate, and interpret morphological information at unprecedented speed and consistency (Lürig et al., 2021, Wäldchen and Mäder, 2018). In this sense, collections are no longer passive archives of biodiversity but dynamic, digitised datasets, where their full value can be facilitated through advanced AI-based CV tools.

1.6 Computer Vision in Ecology and Evolution

CV applications in ecology and evolutionary biology rely on images acquired under markedly different conditions, and these acquisition contexts strongly influence model performance, robustness, and interpretability. Field-based imagery, such as photographs collected during ecological surveys or contributed through citizen science platforms, is typically characterised by substantial heterogeneity in lighting, background complexity, viewing angle, scale, and specimen posture (Wäldchen and Mäder, 2018; Christin et al., 2019). While such images are invaluable for large-scale monitoring and distributional analyses, this uncontrolled variation can obscure fine-

scale morphological signal and complicate the extraction of consistent diagnostic features (Norouzzadeh et al., 2018; Lürig et al., 2021).

In contrast, images generated through the digitisation of natural history collections are typically acquired under highly standardised conditions designed to minimise non-biological variation and maximise reproducibility (Blagoderov et al., 2012; Brecko and Mathys, 2020; Nelson and Ellis, 2019). Such imaging regimes enable more reliable comparisons of morphological traits across specimens, species, and populations, and are therefore particularly well suited to quantitative analyses and interpretable computer vision approaches, where sensitivity to subtle visual differences can otherwise be confounded by artefacts of image acquisition (Wäldchen and Mäder, 2018; Ahmed et al., 2024).

Applications of CV in biodiversity science are becoming diverse. Models have been trained to classify species from photographs of insects, birds, plants, or nanofossils with high levels of accuracy (Carranza-Rojas et al., 2017, Norouzzadeh et al., 2018, Wäldchen and Mäder, 2018, Poon et al., 2024) and other studies have demonstrated its utility in classifying cryptic taxa (Pinho et al., 2023). Beyond simple classification, CV can also extract phenotypic traits such as wing venation, body size, and colour patterning, allowing large-scale quantitative analyses (Wilson et al., 2023, Eshghi et al., 2024). CV has also been applied in ecological monitoring, for instance in the automated processing of camera trap images (Norouzzadeh et al., 2018) and remote sensing data (Ballesteros et al., 2020), and even in palaeontology, where it has been used to classify fossil morphology (Yaqoob et al., 2025). The advantages of CV are considerable where these systems can not only achieve levels of accuracy on par with human experts but can do so at speeds far greater than any human could (Smith et al., 2024). Unlike human-led identification, CV outputs are reproducible and less subject to individual variation or fatigue, making them particularly valuable for large-scale biodiversity datasets (Ahmed et al., 2024). Importantly, CV approaches are non-destructive and naturally compatible with digitised collections, allowing researchers to analyse morphology directly from high-resolution specimen images or even something more nuanced like hyperspectral datasets (Liu et al., 2022).

Nevertheless, challenges remain. Model performance can vary across taxa, and accuracy often declines when classifiers are applied outside their original training domain (Zhou et al., 2021). The reliability of CV methods is heavily dependent on the quality and reproducibility of labelled datasets, with many biodiversity studies lacking transparency in data partitioning and code availability (Ahmed et al., 2024). Moreover, DL models are frequently criticised as “black boxes,” offering little transparency into the features used to make their predictions (Von Eschenbach, 2021, Rudin, 2019). At the same time, other studies adopt these outputs with

limited scrutiny, often using results without interrogating the underlying causation or model behaviour (Lipton, 2018, Doshi-Velez and Kim, 2017).

Despite these challenges, CV presents itself as a new set of tools for biodiversity science. It is not intended to replace traditional or molecular approaches, but to complement them by providing scalable, reproducible, and efficient analyses of morphological data. By alleviating the bottlenecks caused by declining taxonomic expertise and the vast scale of digitised collections, CV opens new possibilities for taxonomy, ecology, and evolutionary biology (Ahmed et al., 2024, Karbstein et al., 2024).

1.7 Explainable AI for Morphological Interpretation

A range of computational approaches have been developed to support morphological analysis and species discrimination beyond traditional visual inspection. Prior to the widespread adoption of deep learning, many studies relied on explicit feature extraction, in which predefined measurements or descriptors were manually selected and quantified. Geometric morphometrics, for example, has been widely applied to capture variation in shape using landmark- or outline-based methods, providing powerful tools for analysing morphological differences across taxa, populations, and environments (Rohlf and Marcus, 1993; Adams et al., 2013). Such approaches offer strong interpretability and direct links to biological form, but they often require substantial expert input, careful landmark placement, and assumptions about homology that may not hold across highly variable or complex structures.

Other approaches focus on engineered or semi-automated feature extraction, where image descriptors such as texture, colour histograms, or edge-based features are used as inputs to statistical or machine learning classifiers (Christin et al., 2019; Lürig et al., 2021). While these techniques can scale more readily than fully manual approaches, they still depend on a priori decisions about which features are biologically meaningful and may struggle to capture subtle or unexpected patterns of variation. More recently, deep learning–based computer vision models have reduced the need for explicit feature specification by learning hierarchical representations directly from image data, often achieving superior classification performance across a wide range of biological applications (Wäldchen and Mäder, 2018; Ahmed et al., 2024). However, these gains in performance have been accompanied by concerns over interpretability, as model decisions are not inherently transparent.

In response to this limitation, a growing body of work has focused on explainable artificial intelligence (XAI), which aims to make the decision-making processes of complex models more interpretable to human users (Doshi-Velez and Kim, 2017; Rudin, 2019). Within biodiversity science, XAI techniques provide a means of linking model predictions back to visible morphological features, allowing computational outputs to be evaluated in the context of established taxonomic knowledge (Murdoch et al., 2019; Pichler and Hartig, 2023). Alternative interpretability frameworks also exist, including model-agnostic explanation methods such as LIME and SHAP, and concept-based approaches that quantify the influence of user-defined traits (Ribeiro et al., 2016; Lundberg and Lee, 2017; Linardatos et al., 2020). Each offers distinct advantages, but many are less naturally suited to high-resolution image data or to the direct visual interrogation of specimen morphology.

Against this backdrop, heatmap-based XAI methods have emerged as a particularly practical approach for image-based morphological studies. By highlighting regions of an image that contribute most strongly to a model's prediction, these techniques allow researchers to assess whether classifiers attend to biologically plausible features, identify potential artefacts, and generate hypotheses about diagnostic characters. While such methods do not provide a complete explanation of model internals, they offer an interpretable bridge between automated classification and traditional, character-based reasoning, making them especially relevant for applications in taxonomy, natural history collections, and morphological research.

Heatmaps are a broad class of visualisations used to represent the relative importance of image regions for a given model output, typically encoded through colour intensity. Saliency maps represent one of the earliest and most direct approaches to visual interpretability in deep learning, generated by computing the gradient of a model's output with respect to the input image to produce a pixel-level sensitivity map highlighting regions where small changes most strongly affect predictions (Simonyan et al., 2013). Gradient-weighted Class Activation Mapping (Grad-CAM) is an activation-based interpretability method that uses gradients flowing into the final convolutional layers of a convolutional neural network to generate class-specific localisation maps (Selvaraju et al., 2017). Unlike saliency maps, Grad-CAM produces coarser, region-level heatmaps that reflect broader areas of attention associated with a particular class decision. Although these methods differ in their computational basis and spatial resolution, both fall under the broader category of heatmap-based XAI techniques and provide complementary perspectives on model behaviour when interpreting automated classifications in terms of visible morphological features.

Here, then, the central problem and opportunity come into focus. There is a need for approaches that retain the transparency of character-based taxonomy while scaling to the

volume and complexity of modern, digitised collections. Computer vision offers the necessary speed and consistency, and XAI supplies the missing interpretability, allowing model decisions to be traced to visible image features. The next section outlines the position of this work within biodiversity science and explains how the following data chapters build from these motivations to evaluate when, and how, computer vision and its explainable toolsets can strengthen taxonomic practice, natural history curation, and the study of morphological variation within and between species.

1.8 Positioning of this Thesis

The study of biodiversity requires tools that reconcile molecular evidence of divergence with the morphological traits that remain central to taxonomy, ecology, and conservation. Traditional morphological approaches provide accessible and interpretable characters but often fail to capture cryptic or subtle variation. Molecular methods, meanwhile, have reshaped species delimitation by revealing genetically distinct lineages, yet frequently leave unanswered questions about how such divergence manifests in visible form, function, or ecology. This disconnect is commonly referred to as the genotype–phenotype gap. Computer vision and explainable artificial intelligence offer a means of addressing this challenge by enabling scalable, reproducible, and interpretable analysis of morphology from image data.

Interpretable computer vision approaches make it possible to move beyond classification alone, allowing image regions and features contributing to model predictions to be visualised and quantified. This creates opportunities to generate testable hypotheses linking genetic structure to candidate morphological traits, to prioritise specimens or characters for further investigation, and to evaluate the consistency of morphological signal across populations and taxa. In this context, explainable computer vision is not intended to replace traditional taxonomic or molecular approaches, but to complement them by providing systematic and transparent tools for analysing morphology at scale. Such integration has the potential to enhance taxonomic workflows, support the curation of large digitised natural history collections, and reveal previously overlooked patterns of morphological variation within and between species.

The thesis is structured around four data chapters:

Chapter 2 applies a CV pipeline to classify several limpet species with cryptic and variable shell morphology. Results are compared with expert predictions, and heatmaps are used to explore differences in classifier attention between species.

Chapter 3 investigates the use of CV to detect mislabelled specimens in large butterfly image datasets, addressing a persistent challenge in digitised NHCs.

Chapter 4 applies CV to genetically divergent limpet populations, testing whether XAI can be used to reveal morphological differences between populations on opposite sides of phylogeographic breaks.

Chapter 5 compares human and machine attention in the classification of closely related British butterflies, examining whether convolutional neural networks (CNN) focus on the same morphological features used by taxonomists.

Together, these chapters demonstrate the potential of XAI not only to support traditional taxonomy but also to expand its scope by uncovering cryptic or overlooked features, standardising large datasets, and helping to bridge the gap between genotype and phenotype. The broader contribution of this thesis is to position CV and explainability as integral components of modern biodiversity science, offering scalable and interpretable tools for studying morphology across species, populations, and NHCs.

1.9 Publications and Author Contributions

Chapter 2: *Using computer vision to identify limpets from their shells: a case study using four species from the Baja California peninsula*.

Published in *Frontiers in Marine Science* on 27 July 2023.
DOI: https://doi.org/10.3389/fmars.2023.1167818

Authors: Jack D. Hollister, Xiaohao Cai, Tammy Horton, Benjamin W. Price, Karolina M. Zarzyczny & Phillip B. Fenberg.

Author contributions: JDH collected images, developed and ran CV models, generated heatmaps, and led manuscript writing. PF collected field samples, contributed substantially to writing, and provided expert input on morphology. KZ collected and identified specimens using molecular methods, contributed morphological insights, and edited drafts. BP, TH, and XC provided feedback on CV methods and edited drafts. All authors approved the final manuscript.

Chapter 3: *A computer vision method for finding mislabelled specimens within natural history collections*.

Published in *Ecology and Evolution* on 13 July 2025.
DOI: https://doi.org/10.1002/ece3.71648

Authors: Jack D. Hollister, Geoff Martin, Xiaohao Cai, Tammy Horton, Owain Powell, Mark Sterling, Glory Turnbull, Ben W. Price & Phillip B. Fenberg.

Author contributions: JDH conceived the study, curated data, developed methodology, ran analyses, and led writing. GM contributed to data curation and analysis. OP, MS, and GT supported data analysis. BWP supported data curation, analysis, and editing. XC and TH supervised and edited drafts. PBF supervised the project, contributed to methodology and analysis, and co-wrote and edited the manuscript. All authors approved the final manuscript.

Chapter 4: *Genes, shells, and AI: Using computer vision to detect cryptic morphological divergence between genetically distinct populations of limpets*.

Published in *Scientific Reports* on 12 December 2025.

DOI: 10.1038/s41598-025-30613-1

Authors: Jack D. Hollister, David A. Paz-García, Rodrigo Beas-Luna, Tammy Horton, Xiaohao Cai & Phillip B. Fenberg.

Author contributions: JDH conceived the project, collected and imaged specimens, organised datasets, developed code, conducted analyses, and drafted the manuscript. PBF co-collected specimens, curated datasets, and edited drafts. DAPG and RBL assisted with fieldwork and manuscript editing. TH and XC provided supervision and contributed to manuscript revision.

Chapter 5: *Do You See What I See? Comparing human and convolutional neural network attention to butterfly morphological features*.

In preparation for submission to a relevant journal.

Authors: Jack D. Hollister, Geoff Martin, Tammy Horton, Xiaohao Cai, Ben W. Price & Phillip B. Fenberg.

Author contributions: JDH curated datasets, developed code, conducted analyses, and drafted the manuscript. GM and BWP advised on butterfly morphology and contributed to editing. PBF assisted with analysis and manuscript development. XC and TH supervised and contributed to manuscript revision.

Chapter 2 Using computer vision to identify limpets from their shells: A case study using four species from the Baja California peninsula

Jack D. Hollister [1,2,3*], Xiaohao Cai [1], Tammy Horton [2,3], Benjamin W. Price [3], Karolina M. Zarzyczny [1,3], Phillip B. Fenberg [1,3]

1. School of Ocean and Earth Sciences, University of Southampton, Southampton, United Kingdom
2. National Oceanography Centre, Southampton, United Kingdom
3. Natural History Museum, London, United Kingdom

2.1 Abstract

The shell morphology of limpets can be cryptic and highly variable, within and between species. Therefore, the visual identification of species can be troublesome even for experts. Here, we demonstrate the capability of CV models as a new method to assist with identifications. We investigate the ability of computers to distinguish between four species and two genera of limpets from the Baja California peninsula (Mexico) from digital images of shells from both dorsal and ventral orientations. Overall, the models performed marginally better (97.9%) than experts (97.5%) when predicting the same set of images and did so 240x faster. Moreover, we utilised a heatmap system to both verify that models are focussing on the specimens and to view which features on the specimens the models used to distinguish between species and genera. We then enlisted the expertise of limpet ecologists specialised in identification of species from the Baja peninsula to comment on whether the heatmaps are indeed focusing on specific morphological features per species/genus. They confirm that in their opinion, the majority of the heatmaps appear to be highlighting areas and features of morphological importance for distinguishing between groups. We then asked them to comment on why the models and themselves may have got a few predictions wrong. Our findings reveal that the cutting-edge technology of CV holds tremendous potential in enhancing species identification techniques used by taxonomists and

ecologists. Not only does it provide a complementary approach to traditional methods, but it also opens new avenues for exploring the biology and ecology of limpets in greater detail.

2.2 Introduction

Limpets are abundant, diverse, and ecologically important members of rocky shore communities (Kordas et al., 2017; Firth, 2021). In addition, some limpet species are important culturally and as food sources for modern and pre-historic human societies (Fenberg and Roy, 2008; 2012; Firth, 2021; Weisler and Rogers, 2021). Yet, despite their ubiquity, limpet species can sometimes be difficult to tell apart in the field (Simison and Lindberg, 2003; Burdi, 2015), at archaeological sites (Rogers and Weisler, 2020a) and in museum collections (Kuo and Sanford, 2013) owing to their highly variable shell morphologies and colour patterns (Nakano and Spencer, 2007). Even within species, shell features can vary according to substrate, size (age), population, and geographic region, sometimes resulting in distinct shell morphologies (Williams, 2017) and shapes (Rogers and Weisler, 2020b) To further complicate matters, shell erosion and encrusting symbionts can also impede visual identification. As a result, taxonomists frequently rely on using internal anatomical features such as radular structure, as distinguishing characters (Simison and Lindberg, 1999). In more recent decades, molecular methods have revealed new limpet species, confirmed/rejected species validity, and clarified nomenclatural confusion among morphologically similar species (Simison and Lindberg, 2003; Crummett and Eernisse, 2007). Nevertheless, the use of internal anatomical or molecular characters for distinguishing similar looking and highly variable species can be time consuming and resource limiting, while offering little advance in species-level identifications using the most easily accessible external features – their shells.

Recent developments in computer-based image recognition and detection may be harnessed to develop accurate, fast, and cost-effective means to distinguish between limpet species from their shells. In addition, these emerging technologies can also provide insight into the morphological characteristics that can be used to distinguish between similar looking species (Pinho et al., 2022). The aim of this paper is to evaluate the feasibility of these new computer-based methods for distinguishing between limpet species and genera using digital images of their shells.

CV is currently pushed forward by DL and artificial intelligence (AI) and focuses on the development of algorithms and techniques for computers to process, understand and analyse visual data inputs. This can involve tasks such as image and video recognition (the recognition of

specified subjects within images and video), object detection (the recognition and location of subjects within an image and video) and scene understanding (the recognition of a subject within a 3D environment with respect to its relationship to other subjects). CV involves the understanding of pixel patterns and their respective colour values. Furthermore, CV systems have the capability to operate for prolonged periods, handle very large datasets, and produce results at very fast speeds (Wilson et al., 2022), which are unachievable and/or unfeasible for humans.

Recently, CV has been adopted by the life sciences as a method to visually identify and group organisms together based on their morphology (Wäldchen and Mäder, 2018; Greeff et al., 2022; Hollister et al., 2022), and has been recognised as an emerging tool for ecology, evolution, and taxonomic research (Høye et al., 2021; Lürig, 2022). The accelerated use of CV in the natural sciences has coincided with the massive digitisation efforts of natural history museums (NHM) (Popov et al., 2021; Wilson et al., 2022), where tens of millions of digital images of specimens and collection data are now available for researchers worldwide. For example, Wilson et al., (2022) applied CV models to >180,000 specimens of digitised natural history specimens of butterflies, resulting in highly accurate sex identifications and body size measurements over a short timescale (one week), showcasing the emerging power of CV for the natural sciences.

The evaluation of CV methods for identifying similar looking species has not been well studied to date with mixed results from the few studies that have. For example, CV models achieved accuracy scores of ~ 50% for identifying species of British carabid beetles (Hansen et al., 2020). But more recently, some researchers have achieved highly accurate results (upwards of 97%) for identifying species of cryptic lizards (Pinho et al., 2022), suggesting that CV models are either getting more accurate and/or that the results can be taxon specific. Regardless, even if highly accurate CV models are achieved, on their own, they do not give researchers any information about how specimens of different species can be distinguished from each other. Similarly, while DNA barcoding can allow for the species-level identification of specimens, traditional morphological taxonomy is required to find distinguishing features between species (Tautz et al., 2003). For CV to be practically useful for identification purposes, they must not only be trained on specimens with known species-level identification (which can be achieved through DNA barcoding and/or expert identification), but newly developed methods need to be integrated to the workflow to provide insights to the decisions made by CV models. In other words, we need to overcome the “black box” problem (Savage, 2022).

DL based systems are often viewed as "black boxes” with internal processes too complicated for comprehension, which can lead to the development of biased models that generate incorrect or biased results, leading to distrust in their results (Sham et al., 2022). To

address these issues, a significant number of researchers are working to improve various aspects of AI. Fortunately, CV has made significant strides in this area, as evidenced by the development of XAI in the form of heatmaps. Heatmaps come in many forms and can be used with a variety of applications. Within CNNs heatmaps are often used as a visualisation tool that can be generated to show which features are learned during the training processes and which parts of an input image were used to make predictions (Selvaraju et al., 2017).

Ecologists and taxonomists are now beginning to realise the potential of integrating heatmaps into their CV models to help classify morphologically similar or cryptic species and to highlight morphologically important characters. Recently, researchers applied CV models and heatmaps for species identification problems of a cryptic group of lizards (Pinho et al., 2022). The researchers found that the heatmaps from their CV models were focussing on areas of the body that were morphologically variable between species (while also noting that future research should focus on the interpretation of heatmap results). Although still in its infancy, we believe that the use of CV and heatmaps will provide insightful, cost effective, and rapid means for the identification of limpet species using shell features. If found to be robust, similar techniques could be used to tell apart cryptic species, populations, and perhaps even shell differences caused by microhabitat or phylogeographic factors.

In this paper, we apply CV models and heatmaps to four limpet species from the rocky intertidal of the Baja California peninsula, Mexico: *Lottia strigatella*, (Carpenter, 1864), *Lottia conus* (Test, 1945), *Fissurella volcano* (Reeve, 1849), and *Fissurella rubropicta* (Pilsbry, 1890). Each species overlaps in range and occupies the rocky shore habitat in the high to mid-intertidal zone. We focus on these species because of their diverse shell morphologies and colour patterns on their dorsal and ventral sides. For example, *Lottia conus* has a variety of dorsal shell patterns that can be described as “wavy”, “ribbed”, “speckled”, or “mixed” (Burdi, 2015; Ross, 2022).

The “true limpets”, which include the *Lottia* species, are in the subclass as Pattelogastropoda, whereas the *Fissurella* species (keyhole limpets) are members of the distantly related subclass, Vetigastropoda. We include *Fissurella* in this analysis because they are ecologically and functionally similar to the true limpets and they live in the same rocky shore habitat. But importantly, the shells of the *Fissurella* species are easily distinguishable by eye from the *Lottia* species due to the distinctive keyhole found only in the Fissurelidae family. Therefore, we expect the heatmaps to also focus on this shell difference when making predictions on which genus a specimen belongs (*Lottia* versus *Fissurella*) and have high accuracy scores. If true, it will give us confidence that the models are focussing on important morphological differences for distinguishing between taxa.

Species level identifications within both genera are more difficult, and therefore, more challenging for both human and CV-based methods of identification. For example, authors have observed multiple cases of misidentifications of *F. volcano* with *F. rubropicta* (and vice versa) in museum collections and *L. strigatella* and *L. conus* each have their own history of taxonomic confusion (Simison and Lindberg, 2003; Burdi, 2015). The *Lottia* species can sometimes be difficult to tell apart as they are both relatively small, have highly variable shell patterns, and live within the same microhabitat (on top of rocks or as epibionts on other shells in the high to mid-intertidal). By applying CV and heatmaps to digital images of the shells of these species, our broader aim is to help solve these classification problems while also identifying shell characteristics that researchers can use to distinguish species, both in the field and in museum collections. To this end, we have trained CV models on specimens with confirmed species identifications (using DNA barcoding) and calculated model accuracies for making correct predictions. We compare the results of the CV models with expert identifications of the same specimens (without prior knowledge of the model results). We then used expert opinion to determine if the heatmaps focused on important or unique morphological features that may be useful for identification purposes. Finally, we asked the experts to view incorrect predictions and provide interpretations as to why these were made.

2.3 Methodology

2.3.1 Field Sampling and DNA Barcoding

Four species were selected for this investigation: *Lottia conus*, *Lottia strigatella*, *Fissurella volcano* and *Fissurella rubropicta*. These species are co-distributed in the mid to high rocky intertidal zone along the Pacific coast of the Baja California peninsula (Mexico). Specimens were sampled from the field at sites spanning the peninsula, from ~23-30ºN. Limpet specimens were fixed in 70% ethanol in the field and transferred to absolute ethanol in the laboratory. To confirm species identification, DNA was extracted from foot tissue using the DNeasy Blood and Tissue Kit following the manufactures instructions (Qiagen). For all species, we amplified a ~630bp fragment of a section of mitochondrial Cytochrome Oxidase Subunit I (COI) gene and sequenced on an ABI 3730 DNA Analyser at the Natural History Museum, London (UK). Total specimen numbers: *L. strigatella = 158*, *L. conus = 120*, *F. volcano = 82*, and *F. rubropicta* = 70. Pairwise sequence distances within each group were calculated and a neighbour joining tree was performed in MEGA (Tamura et al., 2021) to confirm the monophyly of each species. Pairwise distances within each group are small and range from 0.05 (*Lottia conus*) to 0.00 (*Fissurella*

rubropicta) and monophyly of each species was confirmed. Further, we used BLAST searches to match sequences to species on the NCBI database. *Fissurella volcano* and *Lottia strigatella* are on the NCBI database and our sequences matched with a percent identity of >97%. *Lottia conus* sequences were matched (>95%) to sequences obtained from Dawson et al. 2014. There are no published COI sequences of *Fissurella rubropicta*, but the very low pairwise distances between specimens within this group (see above) and its clear divergence from *F. volcano* sequences (77%) gives us high confidence of the identity of this species for our models. Correctly labelled data are essential for creating accurate and un-biased training datasets and to assess the accuracy of model results (Rädsch et al., 2023). We use DNA barcoding, but if available, researchers could also use expert identification from taxonomists to confirm species identity (or a combined approach). Further details of the molecular methods of each species and GenBank accession numbers are in Zarzyczny et al. (under review).

2.3.2 Dataset construction

High-resolution images of the dorsal and ventral sides of the shells of each specimen were captured using an Olympus SZX10 microscope. To optimize image quality, a focal step function was implemented, and a black velvet backdrop was used to minimize background interference. Images were taken in a room with controlled lighting to allow for uniformity. In total, six image sets were created. Four models examined species vs species differences and two examined genus vs genus differences. These were as follows: Dorsal *L. conus* vs *L. strigatella*; ventral *L. conus* vs *L. strigatella*; dorsal *F. rubropicta* vs *F. volcano*; ventral *F. rubropicta* vs *F.volcano*; dorsal *Lottia* vs *Fissurella*; and ventral *Lottia* vs *Fissurella*. Therefore, each model was made up of two classes. The models were designed to learn features from visual data inputs (the images) through a computational training process, resulting in predictions based on the learned features.

The images in each class were divided into three groups: training, validation, and test. The training images were used to train the model, the validation images were used for self-verifying and updating model weights during the training process, and the test images were reserved for the final evaluation of the model performances. Due to the limited number of specimens available, models may struggle to train effectively due to a lack of data to identify unique features. To address this issue, we employed image augmentation, a technique that generates artificial images based on the original stock. This has been shown to improve model performance when faced with such situations by creating a larger stock of images, but where the desired features remain unique and non-repeated (Perez and Wang, 2017; Xu et al., 2023). To preserve the integrity

of the specimens' morphology, we chose augmentations that did not alter their colour or shape. We utilized a range of random flips (vertical and horizontal), two rotations functions (a fixed 90^0 clockwise or anticlockwise and a separate clockwise or anticlockwise rotation, up to a maximum of 89^0), and a zoom out (decreases the size by a maximum of 10%). Each of these were set with an 80% probability of being selected and programmed to not create duplications. Before any augmentation was applied, 20 images from each class (i.e., 40 images in total per model) were randomly selected from each image set and set aside as the test set. The test set must remain neutral, un-augmented and unseen by the model. 20 images from each class were randomly chosen and used as the validation set. These 20 validation images were augmented to a combined total of 400 images. The remaining images in each class were used for training and were augmented to a combined total of 3000 images. Overall, each model would contain 6000 training images, 800 validation images and 40 test images.

2.3.3 Computer vision model

We used a high-specification workstation equipped with an NVIDIA GPU with TensorFlow and Python programming. The image classification technique, which consisted of a CNN, was deemed the most appropriate for this scenario. The VGG16 CNN algorithm (Simonyan and Zisserman, 2014) with custom top layers and transfer learning using the ImageNet dataset, was employed. Models were tuned using KerasTuner to find optimum hyperparameter and learning rate (Joshi et al., 2021). They were initially trained for three epochs. Afterwards, all models were fine-tuned by unlocking previously trained layers starting from layer 11, and each model had a final learning rate set lower than its original (initial learning rate/10) and continuously trained until the model validation accuracy plateaued. The time taken for the model development (training and validation) and testing phases for each model were noted for comparison with the expert classification (see below).

2.3.4 Model and expert identifications: evaluation and comparisons

The experts visually identified the same test sets of images for each model but with the species labels removed and with no prior knowledge of model results. They also kept track of how long it took to go through the dataset. Expert accuracy scores were then compared to the confirmed species identifications based on the barcoding results. They were subsequently compared to the prediction accuracy scores for each CV model. Further, the incorrect predictions for specimens

for each method (expert versus CV model identification) were compared to look for any congruent patterns (e.g., do both methods misidentify the same specimens?). The accuracies for both methods of identification (model versus expert) were calculated as the proportion of the correct predictions out of the total number of possible predictions. Accuracies are therefore scored between 0 – 100%, with 100% being a perfect score. The model and expert predictions were further evaluated using a bootstrap analysis to create a 95% confidence interval on the accuracy scores (resampling single specimen predictions with replacement 10,000 times). Overlap in 95% CI was used to judge if there were significant differences between expert and model predictions. Differences in the time taken to make predictions between the expert and the models were also noted. In addition, for each specimen that was incorrectly identified, the experts made a post-hoc judgement as to why they thought an incorrect identification was made and whether the models and experts made the same mistakes.

2.3.5 Heatmap evaluation

The Gradient-weighted Class Activation Mapping (GradCam) system (Selvaraju et al., 2016) was selected to create heatmaps for each specimen image in the test datasets. GradCam is a technique used in CV to understand which parts of an image influenced a DL model decision. It works by analysing DL model activations and gradients to create a heatmap that highlights the important regions in the image. This heatmap helps us see what the model focused on when making its prediction. The heatmap images of each specimen in the test datasets were then shown to the experts to help evaluate which features of the shell, if any, were used to make predictions. Both experts are limpet ecologists (PBF and KMZ) that use visual cues to determine species identification of Baja Peninsula limpets, often using digital images of shells taken in the field.

To further evaluate the use of heatmaps to distinguish between classes, we compared the intensity values between each class per model. When the heatmaps are produced, a value is assigned to each pixel depending on how strongly a particular pixel contributes to the classification decision made by the model. The higher the intensity score of a pixel, the more significant its contribution to the predicted class. These values are then summed up to produce the overall heatmap intensity value per specimen. Although the value itself cannot tell us what part of the shell is being used for prediction, significant differences in overall heatmap intensity values between classes might be evidence that the models are using different features of each class to make predictions. Comparisons in the mean difference of heatmap intensity values

between each class per model were evaluated using two sample Wilcoxon tests (due to violations of normality for some models).

2.3.6 Pairwise Model justification

Pairwise (one-versus-one) classification models were used rather than a single multiclass model to support clearer interpretation of classifier behaviour. Decomposing classification problems into focused pairwise contrasts has been shown to simplify decision boundaries and reduce ambiguity when analysing model outputs, compared with aggregated multiclass formulations (Hand and Yu, 2001). From an interpretability perspective, explainable artificial intelligence approaches emphasise the value of local, class-specific explanations, which are more readily obtained when models are trained to discriminate between two focal classes rather than across a larger pooled label space (Ribeiro et al., 2016). The use of pairwise models therefore provides a more interpretable framework for examining which morphological regions contribute to discrimination between specific species, while retaining comparable classification performance to multiclass alternatives.

2.4 Results

2.4.1 Final model and expert accuracies

Accuracy scores of all trained models with 95% confidence intervals in brackets		
Model	**Expert Accuracy**	**Computer model Accuracy**
Model 1: Dorsal *Lottia* vs *Fissurella*	100% [100, 100]	100% [100, 100]
Model 2: Ventral *Lottia* vs *Fissurella*	100% [100, 100]	100% [100, 100]
Model 3: Dorsal *L. conus* vs *L. strigatella*	95% [87.5, 100]	95% [87.5, 100]
Model 4: Ventral *L. conus* vs *L. strigatella*	92.5% [82.5, 100]	97.5% [92.5, 100]
Model 5: Dorsal *F. rubropicta* vs *F. volcano*	97.5% [92.5, 100]	95% [87.5, 100]
Model 6: Ventral *F. rubropicta* vs *F. volcano*	100% [100, 100]	100% [100, 100]

Table 1. Accuracy scores of all trained models with 95% confidence intervals in brackets

The models and experts produced highly accurate results (Table 1). Overall, the models only incorrectly predicted five images (out of 240), for an overall accuracy score of 97.9%. The experts

also performed well overall, with only six images incorrectly predicted (out of the same 240 images), for an overall accuracy score of 97.5%. Both produced a 100% correct prediction rate using the test sets from models 1, 2 and 6. The experts' worst performance was with the test set from model 4 with an accuracy score of 92.5%. The models' worst performance were models 3 and 5, with an accuracy score of 95%. The 95% confidence intervals overlap for all models, suggesting a non-significant difference between model and expert identification of limpet shells. The experts performed the predictions on all the test images in 59 minutes while the models predicted their respective test images sets in less than 6 seconds (~25s in total).

2.4.2 Heatmaps and expert interpretation

After the heatmaps were shown to the experts, they confirmed the following: Across all six models, all the heatmaps were focussed on the specimens (except for one image within model 1 within the *Lottia* class). Across all six models, all heatmaps appeared to be focussed on specific areas of the shells (except for the same one image in model 1). Heatmaps often focused on a single area of the shells while others focused on multiple features. These features were often common across all images within each respective class (e.g., for the *Fissurella* class in the genus models 1&2, the focus was always on the keyhole). To review which features were highlighted most frequently, we tallied the responses within the comments made by the experts. For example, if a shell feature/area was focused on in all images from a single class, it would equal 20/20.

2.4.2.1 Expert opinion: Model 1

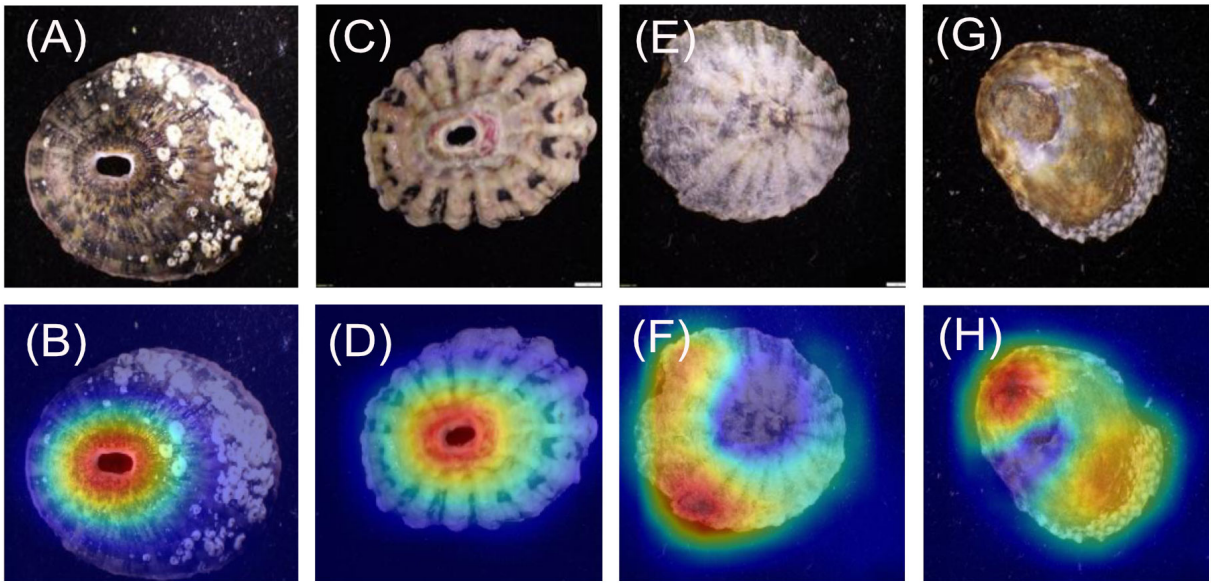


Figure 1. Model 1: Dorsal Lottia vs Fissurella.

For the *Fissurella* images (Fig. 1A – D), 20/20 heatmaps focused on the keyhole. For the *Lottia* images (Fig. 1E - H), 19/20 heatmaps focused on patterns around the shell margin. One heatmap focused its attention around the outside of the shell rather than on it but was still correctly predicted as *Lottia*. It was noted that the specimens within the *Lottia* class had a high degree of variable shell patterns and morphology.

2.4.2.2 Expert opinion: Model 2

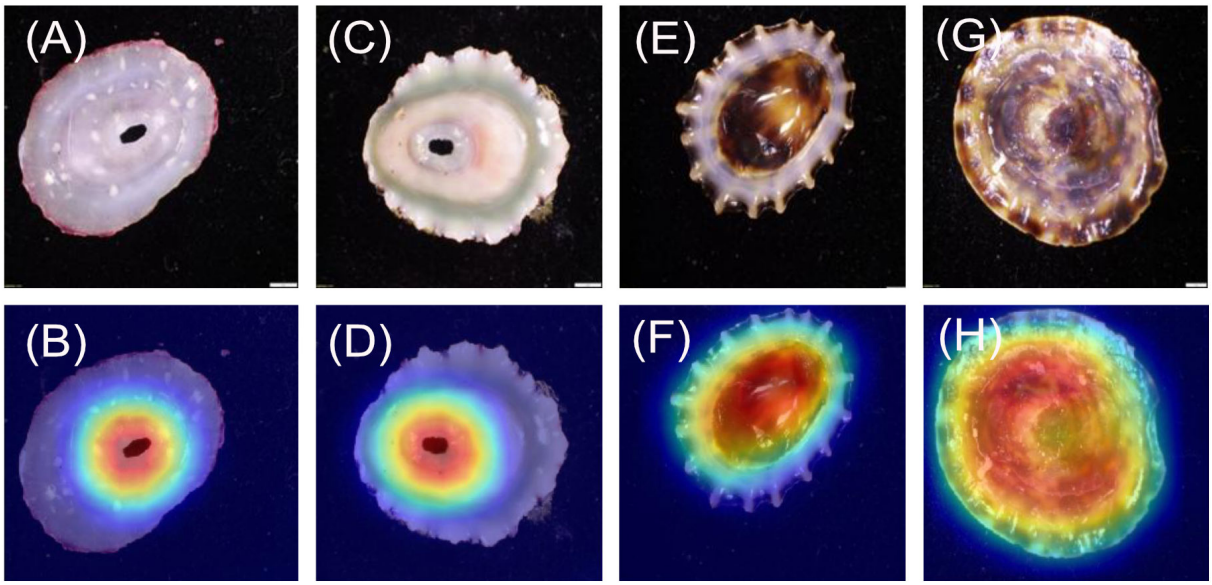


Figure 2. Model 2: Ventral Lottia vs Fissurella.

For the *Fissurella* images (Fig. 2A – D), 20/20 heatmaps focus on the keyhole. For the *Lottia* images (Fig. 2E – H) 19/20 focused on the areas within the muscle scar and not on the shell margins, while 1/20 focused on the muscle scar to the shell margin.

2.4.2.3 Expert opinion: Model 3

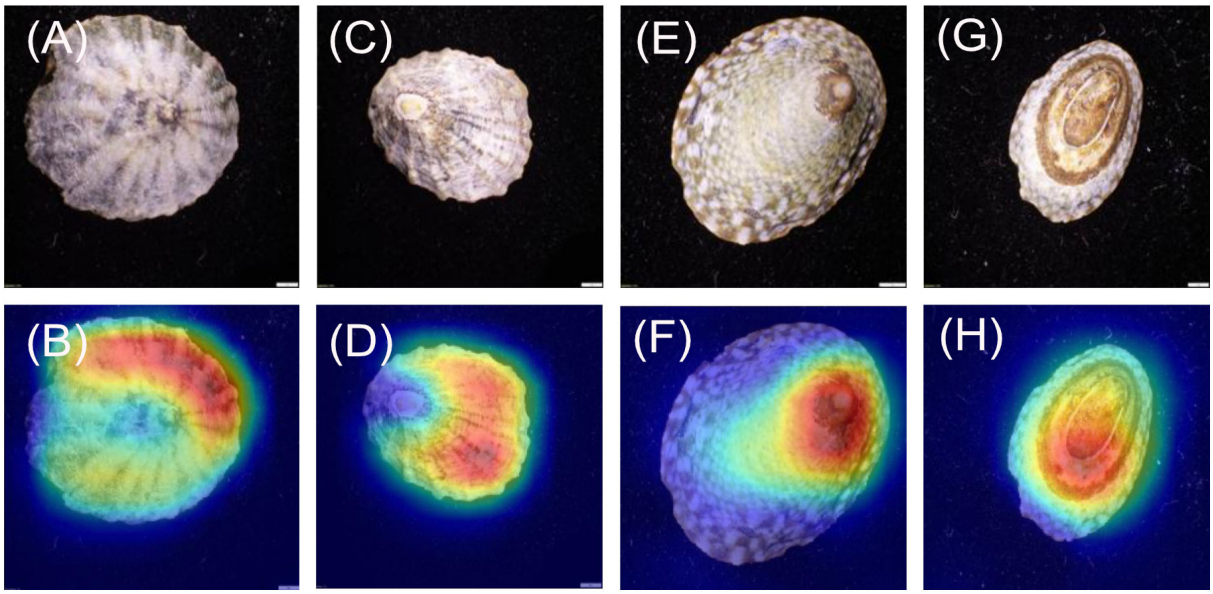


Figure 3. Model 3: Dorsal Lottia conus vs Lottia strigatella.

For the *L. conus* images (Fig. 3A – D), 20/20 heatmaps focus on the ribbing pattern on the shell, but not on the apex. For the *L. strigatella* images (Fig. 3E – H) 17/20 heatmaps focused on the apex, while 3/10 focussed on patterns around the apex.

2.4.2.4 Expert opinion: Model 4

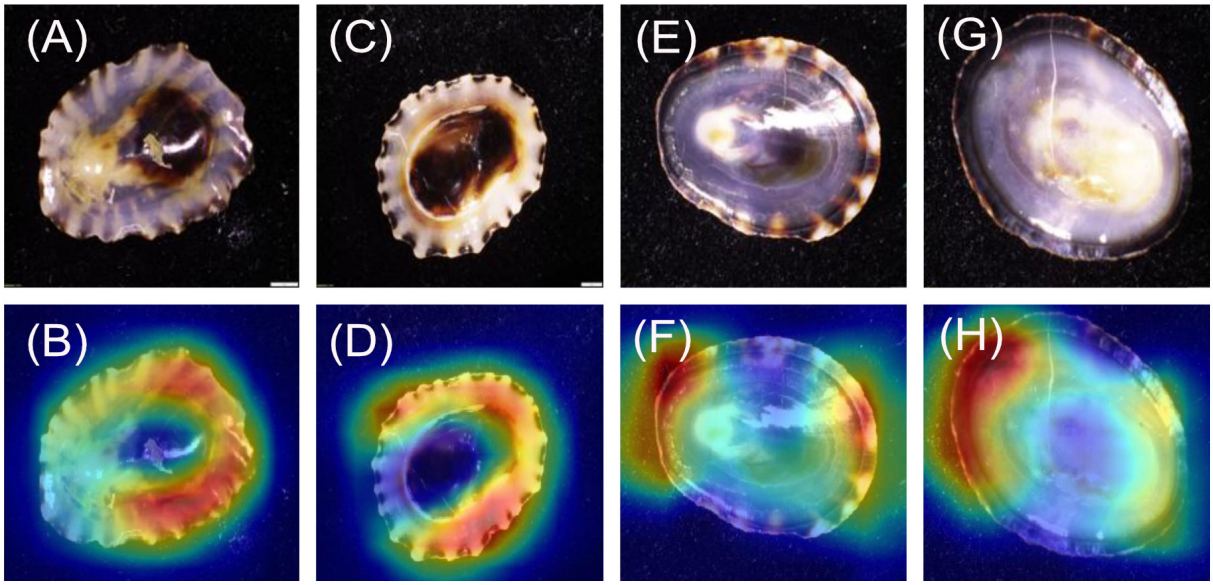


Figure 4. Model 4: Ventral Lottia conus vs Lottia strigatella.

For the *L. conus* images (Fig. 4A – D), 19/20 heatmaps focus on the area between the muscle scar and the shell margin. 1/20 focused on a very small portion of the shell margin, however, this shell was noted as containing no pattern and was predicted incorrectly (Fig. 7). For the *L. strigatella* images (Fig. 4E – H), 20/20 heatmaps focused on areas of the shell margin which is often bordered by a dark or mottled band. Additionally, 2/20 also focused on the centre of the interior portion of the shell within the muscle scar.

2.4.2.5 Expert opinion: Model 5

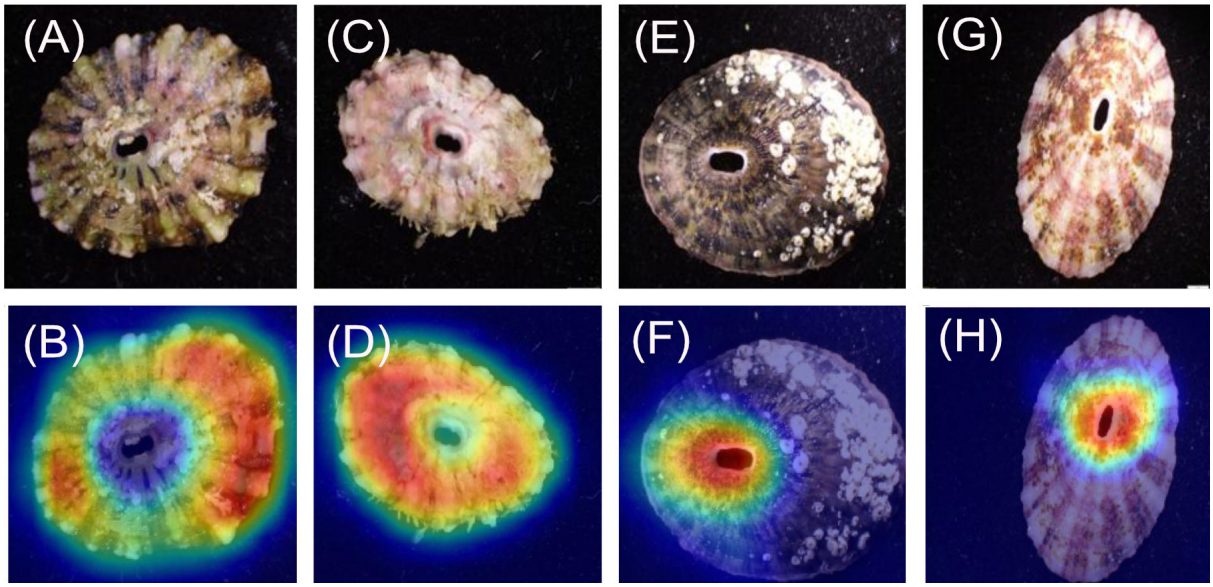


Figure 5. Model 5: Dorsal Fissurella rubropicta vs Fissurella volcano.

For the *F. rubropicta* images (Fig. 5A – D), 20/20 heatmaps focus on the ribbing pattern on the shell, but not on the keyhole. It was noted that some of the shells were highly eroded but the heatmap still focused on any remaining ribbing patterns. For the *F. volcano* images (Fig. 5E – H) 20/20 heatmaps focused directly on the keyhole. It was noted that the keyhole shape between the two species is different.; *F.* rubropicta is more lemniscate while *F.volcano* is ellipsed.

2.4.2.6 Expert opinion: Model 6

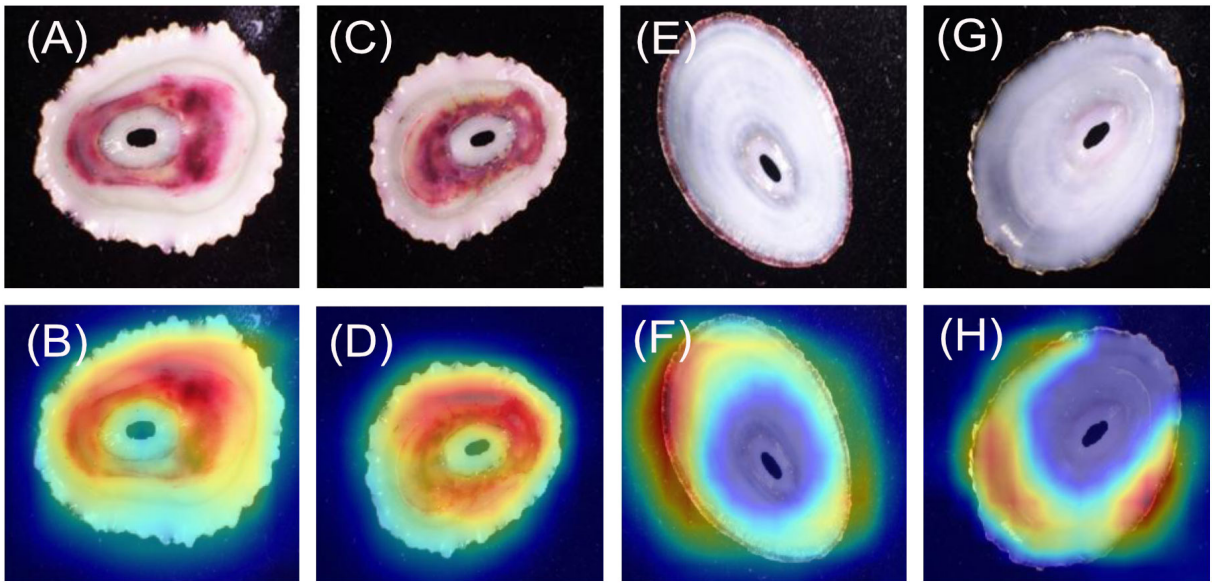


Figure 6. Model 6: Ventral Fissurella rubropicta vs Fissurella volcano.

For the *F. rubropicta* images (Fig. 6A – D), 20/20 heatmaps focus on the area between the muscle scar and callus (which usually contains a deep red colour) but not on the shell margin. For the *F. volcano* images (Fig. 6E – H) 18/20 heatmaps focused on the margin. 2/20 focused on the margin and on the interior of the shell.

2.4.2.7 Incorrect model predictions and expert interpretation

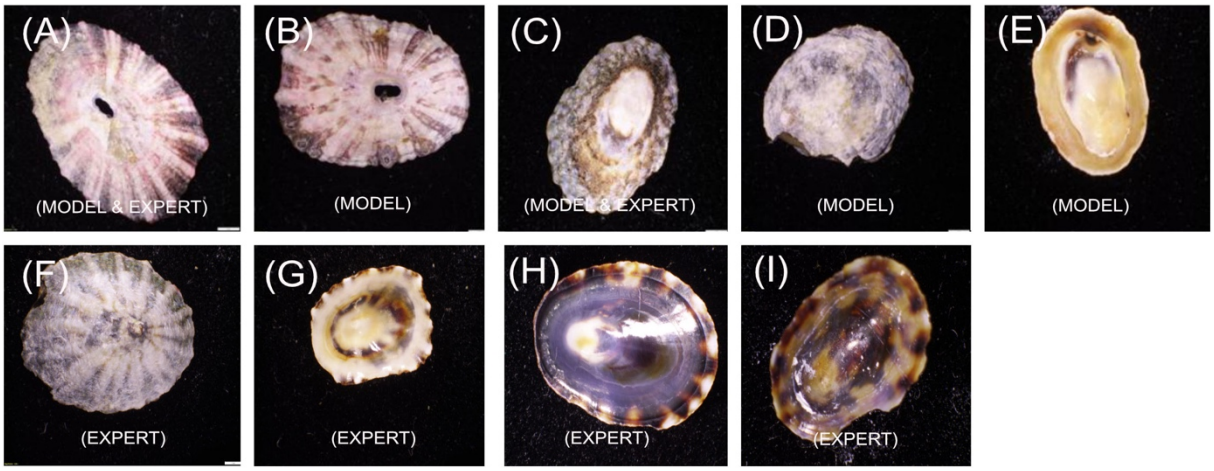


Figure 7. All incorrect model and expert image predictions.

All incorrect model predictions and all incorrect expert predictions (Fig. 7) were shown to the experts who were asked to provide an opinion on what morphological features may have caused the misidentification.

2.4.2.7.1 Expert opinion: Incorrect model predictions

The *F. volcano* specimen in Figure 7A was incorrectly predicted by both the model (model 5) and the experts (both incorrectly predicted it as *F. rubropicta*). Experts determined that this specimen has ribbing patterns normally associated with *F. rubropicta*. Experts were convinced they were correct but after visual inspection of the ventral side and a correct prediction by the ventral model (6), they concluded that this may just be an outlier individual with dorsal characteristics of both species of *Fissurella*. In image B, the *F. rubropicta* specimen was incorrectly predicted by model 5 as *F. volcano*. Experts determined that this specimen displayed features that they would expect from *F. volcano* as it has less defined ridging. Image C, an *L. conus* specimen was incorrectly predicted by model 3 as *L. strigatella*. This same specimen was also incorrectly predicted by the

experts. Upon subsequent inspection, the experts determined that the morphological features of this specimen are not typically associated with *L. conus* such as not having a banding pattern and the shell pattern is more stippled, which they often attribute to *L. strigatella*. Image D, an *L. strigatella* specimen was incorrectly predicted by model 3 as *L. conus*. The experts determined that the shell is highly eroded and very little morphological information can be used to make a prediction. Image E, a *L. conus* specimen was incorrectly predicted by model 4 as *L. strigatella*. Experts determined that it also has very little pattern and largely monochromatic, making it difficult to identify.

2.4.2.7.2 Expert opinion: Incorrect expert predictions

Images A and C were incorrectly predicted by both the models and the experts, with reasonings outlined above. Image F is a dorsal view of an *L. conus* specimen that was incorrectly predicted by the experts as *L. strigatella*. On reflection, experts commented that they can see some clear *L. conus* morphological features (clear banding pattern) and were unsure how they incorrectly predicted the specimen initially. Image G is a ventral view of an *L. conus* specimen that was incorrectly predicted by the experts as *L. strigatella*. Again, on reflection, experts determined that they could see *L. conus* features (ribbed margin) and were unsure how they incorrectly predicted the specimen. Image H is a ventral view of an *L. strigatella* specimen that was incorrectly predicted by the experts as *L. conus*. On reflection experts determined that the banding pattern around the margin is a feature they would usually associate with *L. conus*, making this specimen a difficult one to predict (but was correctly predicted by the model). Image I is a ventral view of a *L. strigatella* specimen and was incorrectly predicted by the experts as *L. conus*. Experts determined that the pattern on this specimen is unusual and is displaying a tortoiseshell pattern that they could attribute to both *Lottia* species (the model predicted this specimen correctly).

2.4.3 Heatmap intensity values

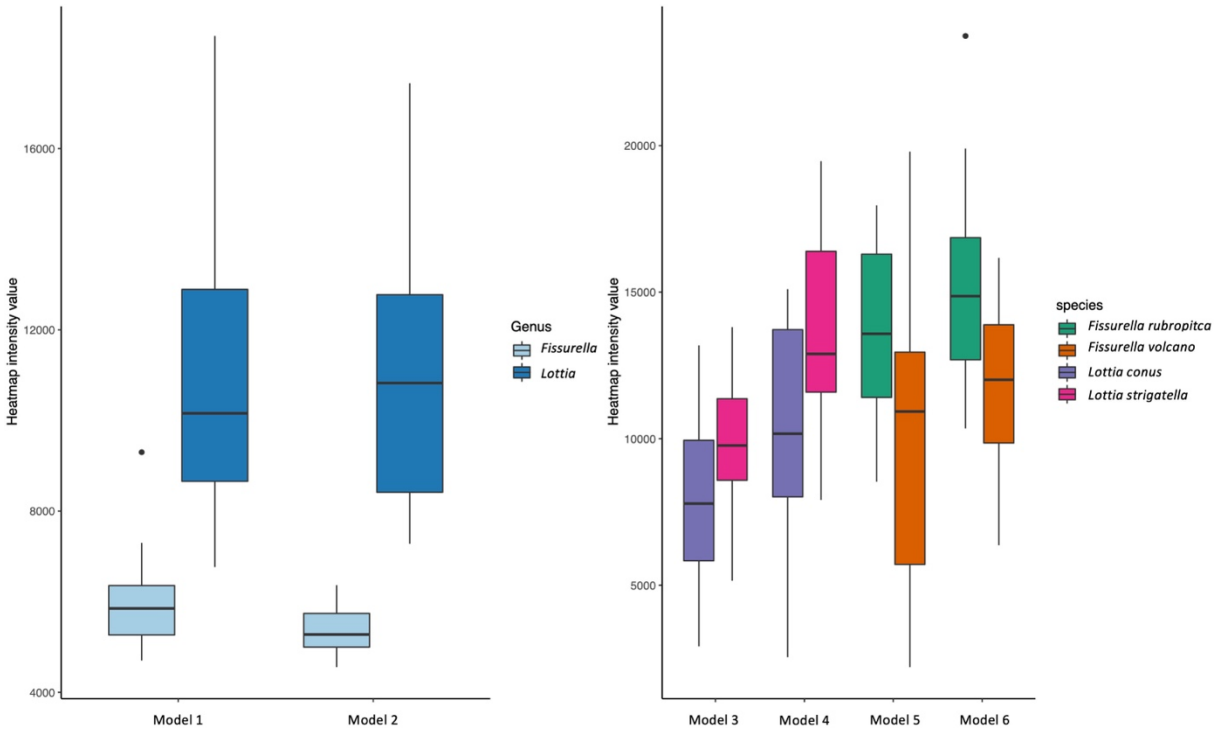


Figure 8. Boxplots showing heatmap intensity values for all models.

For the heatmap intensity values, all models showed a significant difference ($P<0.05$; two sample Wilcoxon tests) in the mean values between each class (Fig. 8). For both genus models (1&2), the *Fissurella* values are much lower and with a smaller range of values than the *Lottia* values. For the species comparisons (Fig. 8), both the dorsal and ventral views for the *Lottia* models (3&4) are higher, on average for *L. strigatella* than *L. conus*. Likewise, the values for *F. rubropicta* are on average higher than *F. volcano*.

2.5 Discussion

2.5.1 Computer vision-based limpet identification

The use of CV to help distinguish between species is starting to gain traction amongst ecologists and taxonomists (Wäldchen and Mäder, 2018;Greeff et al., 2022;Hollister et al., 2022). However, few have attempted to pair CV models with heatmaps to help visually distinguish between species with high morphological variability. Limpets, including those species used in this study, can have multiple colour morphs and shell patterns due to several different ecological and life history factors, including substrate type, age, and patterns of shell erosion (Bird 2011; Williams,

2017). It is therefore not uncommon for field ecologists and museum curators/taxonomists to make mistakes in species identification. To help assist identification, our CV models performed very well and the heatmaps largely focus on shell areas that are putatively morphologically informative between genera and species. These interpretations should be treated as hypothesis-generating, as heatmaps visualise regions of model attention rather than providing direct evidence of causal morphological mechanisms.

When considering the genera, the models achieved 100% predicted accuracy for the dorsal and ventral orientations (models 1 and 2 respectively). Previous research has shown that higher taxonomic levels tend to score greater than lower levels with computer-based classification problems (Hansen et al., 2020). This is most likely due to having more unique images and having a larger selection of features to associate to each respective class, both of which are shown to improve the performance of CV models (Shorten and Khoshgoftaar, 2019). This follows general taxonomic identification procedures, where higher taxonomic levels are more easily distinguished (Hennig, 1966). It is important to recognise that all *Fissurella* species have a distinctive keyhole in their shells, whereas true limpets (including *Lottia*) do not. This is a very clear method of distinguishing the two by eye and the high genus level accuracies evidence this through perfect model performance. This is further supported by the heatmap analysis which clearly shows that the models are focussing on the keyhole of all *Fissurella* specimens within both models. When viewing the *Lottia* specimens, the heatmaps are looking at different areas of the specimens, which is reflective of the varied morphology of *Lottia*. When viewing the heatmap intensity values (Fig. 8), the *Fissurella* class have a much lower mean and spread of values, while the *Lottia* class has a much higher mean and spread of values. This shows that the models utilise much less visual information to determine the *Fissurella* class while requiring a lot more information to determine the *Lottia* class. The experts commented that the keyhole, or a lack of, would be their defining feature to classify either class.

The species vs species models achieved more variable, but still highly accurate results. The ventral oriented models performed better than the dorsal oriented models across both species' groups. The *Lottia* ventral model (model 4) performed slightly worse (achieving a prediction accuracy of 97.5%) than the *Fissurella* ventral model (model 6), which achieved a prediction accuracy score of 100%. However, the *Fissurella* dorsal model (model 5) and the *Lottia* dorsal model (model 3) performed equally well (95%). We believe this slight difference in performance between the ventral and dorsal orientations lies in the fact that the dorsal sides will incorporate many factors that can alter appearance, such as erosion and encrusting symbionts that can cover the shell, all of which would hinder the accuracy of CV models. However, the ventral side remains hidden and protected from physical elements. Thus, the ventral side may provide a clearer picture of the differences between species and therefore provide maximum identification

opportunities for the CV models. Although this option is not preferable for field identification as the body tissue would need to be removed from shells. Dry shell collections of museum specimens or those collected for other purposes (e.g., population genetics) however, could benefit from the use CV on the ventral and dorsal shell for identification. Again, the heatmap intensity values for the species-based models showed significant differences between each class per model. This suggests that the models found morphological features or areas of similar importance within each class when making their respective predictions. We believe this type of assessment could help reinforce the decision made by CV system. For instance, if a prediction does not fit into a known boundary of heatmap intensities for a given class, then it could either be ruled as incorrect or could, at the very least, warrant further investigation, either by revisiting visually by an expert or by molecular means. The heatmap intensity values for the incorrectly predicted specimens (n=5) tend to be lower than the values for the correctly predicted specimens (n=235).

2.5.2 Expert identification and comparison to model performance

Experts performed marginally worse (by one specimen) than the model predictions when considering all images used for the test datasets (n=240). The experts achieved 100% on the genus-based models (models 1 and 2) and the ventral *F. rubropicta* vs *F. volcano* (model 6) which is equal to the model performance. The experts performed marginally better on model 5 by 2.5%, performed equally on model 3 and performed worse on model 4. These small differences however are not significant (Table 1). What is striking is the difference in time it takes for the experts and models to make their predictions. It took the experts on average 10 minutes to identify each test set (59 minutes in total) while each model could process their respective test images in less than 6 seconds (25 seconds in total). The experts combined years of limpet-based experience in the study region totals over 22 years, having viewed countless specimens to achieve their personal knowledge base. In contrast, each model used no more than 158 unique images and was created in less than 5 minutes of training time. Therefore, the sheer speed at which CV models can make accurate predictions is one of its primary advantages. This comparison reflects inference time only and does not include the time required to acquire specimens, generate and label training data, or develop and train the models. It therefore illustrates relative identification speed once systems are operational, rather than the total investment required to deploy a computer vision workflow.

The more unique images that are available for training, then the better the performance of the finished model (Shorten and Khoshgoftaar, 2019). However, at the time of the project, a

limited number of specimens were available to create the models, so it is highly likely that if more unique images were available (e.g., from confirmed museum specimens) then we believe that subsequent models could perform even better than those achieved within this project. Interestingly, when viewing the incorrect predictions by themselves, the experts felt that some of their incorrect predictions were a result of human error. A typical downside of the human condition is that performance can decrease due to fatigue or many other cognitive and physical conditions (Mallis et al., 2004), which computers do not suffer. Thus, in the future, we envision that many thousands of specimens can be fed into similar models for identification purposes (e.g., for bulk field collected specimens or un-catalogued museum accessions), alongside confirmation and quality control from expert taxonomists and molecular ecologists.

2.5.3 Heatmap production and expert interpretation

The heatmaps were found to almost always focus upon the specimens, regardless of the model or class. This is a good indicator that the models trained effectively despite the relatively low number of unique images. After the heatmaps were shown to the experts, it was agreed that almost all were focussing on parts of the shell considered to be morphologically important. Occasionally, models focussed on a singular feature, whilst other times they would focus on multiple features. Neither outcome could be considered incorrect. When visually identifying specimens, a human would use a variety of features to make a final decision. However, with cryptic or highly variable species (e.g., some limpet species), the number of defining features is likely to be limited and/or variable among specimens. For instance, the dorsal orientation of *L. conus* vs *L. strigatella* can appear similar (model 3), with only a couple of shell characteristics that can be used to distinguish them by eye. In addition, the dorsal side of both shells can be highly eroded making species identification more difficult when only viewed dorsally (e.g., as they are in situ). Regardless, the heatmaps for model 3 appeared to find consistent morphological differences between the species. The *L. conus* heatmaps mainly focussed on the shell patterns around the apex and looking at shell patterns, while the *L. strigatella* heatmaps mainly focussed on apex itself. The apex on the dorsal shell of *L. strigatella* is often highly eroded (Keen, 1971), more so than on *L. conus*. The apex is the oldest part of the limpet shell, and therefore it is often the most eroded. It is therefore possible that the pattern of shell erosion on the apex is different between the two *Lottia* species, which may reveal differences in their internal shell structures or microhabitats (Day et al., 2000), but this has not yet been studied in these species.

Again, there are consistent differences between the *Lottia* species on the ventral sides of their shells. The heatmaps mainly focussed on the area between the muscle scar and the margin

areas of the ventral sides of *L. conus* shells. Whereas on *L. strigatella*, focussed on the margin perimeter which often contains a dark band. The dorsal orientations of *F. rubropicta* vs *F. volcano* (model 5) heatmaps displayed consistent differences. The *F. rubropicta* heatmaps consistently focussed on the area around the keyhole/callus, but not on it, while the *F. volcano* images consistently focussed on the keyhole. The *F. rubropicta* specimens have more pronounced ribbing on their shells, which the heatmaps appear to focus on. Whereas *F. volcano* shells are smooth with black/reddish rays. These shell differences may be related to their microhabitat differences: *F. volcano* are usually found underneath rocks and in sheltered crevices (Morris et al., 1980) while *F. rubropicta* are exposed and found on top of rocks (PBF and KMZ personal observations). The smooth shells of *F. volcano* are more suited to life underneath rocks and in crevices, whereas the heavy ribbing of *F. rubropicta* likely helps reduce water loss (due to higher surface area) during long periods of aerial exposure. Again, the ventral orientation of *F. rubropicta* vs *F. volcano* (model 6) heatmaps displayed consistent differences. The *F. rubropicta* consistently focused on the area within the muscle scar and around the callus, while the *F. volcano* consistently focused on the margin which usually contains a dark band.

2.5.4 Future considerations

For CV models to be robust, images of accurately identified specimens are required for training purposes. To do this, we relied on DNA barcoding to confirm the species level identifications of the training dataset and to evaluate the accuracy of the test dataset identifications from both the CV models and experts. All specimens were therefore already identified to species level prior to developing the CV models. However, molecular work can be expensive and time consuming. To reduce costs and time, the workflow could be adjusted where only the training dataset are barcoded, and then a smaller sub-sample of specimens in the test dataset could be barcoded to statistically assess the accuracy of the models. Ultimately however, the more specimens that are available for training purposes, the more accurate the model results. If large datasets of confirmed and standardized training images are made publicly available for the known species in a study region, then future researchers could use them to supplement their own training datasets. In particular, we need more training images of the dorsal side of limpet shells, as they are primarily used for field identifications.

More research is also needed to help interpret the utility of heatmaps for understanding ecological questions related to limpet shell morphology (Bird, 2011; Hamilton et al., 2020). With more robust training datasets per species from multiple populations, age/size ranges, and habitat types, we may be able use heatmaps to help decipher if and how shell morphology varies

intra-specifically over local to regional scales. For example, the intensity and location of heatmaps may differ based on factors such as: microhabitat, population, size/age, and region. We can then use this information to shed new light on how and why limpet shells have such high morphological variability (Giesel, 1970).

2.6 Conclusion

This project demonstrates the effectiveness of using CV in identifying limpets based on images of their shells. Despite the variable shell morphologies and colour patterns within and between species, the CV models were able to classify them to genus (100%) and species level (95% - 100%) with high accuracy and quickly, even with small datasets. The use of heatmaps confirmed that the models were focusing on the limpet shells, and when reviewed by expert taxonomists, they agreed that the heatmaps highlighted significant and unique morphological features for each genus and species.

Typically, DL models are considered as 'black box' systems due to their complex decision-making processes and the 'impossibility' of truly understanding how these types of systems come to their final conclusions. However, the use of heatmaps offers a means to understand how CV makes its decisions. The results show that the models can differentiate between visually similar species or those with high morphological variability, and that they utilize unique morphological features to distinguish them. In the future, we envision this type of system being used by taxonomists as a tool to assist them in identifying important or new morphological features to help distinguish between visually similar and cryptic species. Additionally, similar methods could assist with field identification of limpets and potentially replace the need to collect numerous specimens purely for identification purposes. Computer models, once trained, require far less computation power to perform identifications, and most can be uploaded and used from a modern mobile phone.

It is important to consider the strengths and limitations of CV models for identification purposes. No single method is perfect, but combining the strengths of CV, molecular methods, and human expertise will allow us to gain new insights for taxonomy and ecology. Not only for limpets, but for all of biodiversity.

Chapter 3 A computer vision method for finding mislabelled specimens within natural history collections

Jack D. Hollister[1,2,3], Geoff Martin[1], Xiaohao Cai[4], Tammy Horton[3], Owain Powell[1], Mark Sterling[1], Glory Turnbull[1], Ben W. Price[1], Phillip B. Fenberg[1,2]

1. Natural History Museum, London, Cromwell Road, South Kensington, London SW7 5BD
2. School of Ocean and Earth Science, National Oceanography Centre, University of Southampton, Waterfront Campus, European Way, Southampton SO14 3ZH
3. National Oceanography Centre, European Way, Southampton SO14 3ZH
4. School of Electronics and Computer Science, University of Southampton, University Road, Southampton SO17 1BJ

3.1 Abstract

NHCs are essential for biodiversity and evolution research and for studying biotic responses to global change. However, the numbers of specimens within NHCs pose management challenges. Reduced funds, declining taxonomic training, and expanding collections can lead to mislabelled or missing specimens. This highlights the need for innovative and non-destructive methods of taxonomic verification for specimens in large collections. While genetic analyses offer precise verification, they are resource-intensive and less effective on degraded DNA from older specimens, with risks of damage to smaller specimens. CV can automate tasks such as species-level verification and morphological examination, though these techniques have yet to be incorporated and utilised by NHCs for such management tasks. Digitisation initiatives, such as those at the NHM, London, have gained momentum in recent years, converting specimens to digital formats and enhancing global accessibility. Here, we describe a CV pipeline applied to the digitised British and Irish Lepidoptera collection at the NHM. Specifically, our pipeline identifies specimens that do not match their labelled species status. The pipeline was executed for 100 runs for the Butterfly and Moth datasets, resulting in 99,350 out of 350,208 specimens (28.37 %)

being flagged at least once. We attribute a portion of these as pipeline errors, given the likelihood of some mislabelled specimens within training datasets. However, specimens flagged consistently across > 80 % of pipeline runs are likely mislabelled within the collections. Taxonomic experts visually examined 210 such specimens, finding 145 to be incorrectly labelled in the collection or the NHM data portal. Additionally, 30 specimens were sent for genetic verification to confirm species-level identification. This synergy of CV and genetic-based species identification enhances the accuracy and efficiency of managing NHCs, preserving their value for future generations.

3.2 Introduction

NHCs are essential datasets for much of modern-day ecology and evolution research (Popov et al., 2021), including as baseline data for documenting biotic response to global change (Wilson et al., 2023). With the recent push towards massive digitisation efforts by NHMs, NHCs have become ever more accessible to researchers, educators, and the general public. However, with more researchers accessing large digitised NHCs (Hardy et al., 2023), it becomes increasingly important to ensure that specimen label information is accurate (e.g., species name). But finding and correcting specimen label errors within large NHCs is resource-intensive and time-consuming.

The curation, upkeep, and maintenance of access to NHCs are major challenges for museums. For example, funding and staffing have not kept pace with the expansion of collections, leading to shortcomings in management and care of these critical resources (Paknia et al., 2015). Adding to these challenges is the decreasing reliance on traditional morphological identification methods, due to a decline in the number of taxonomic specialists, resulting from an ageing expert base and a lack of incoming specialists. As expertise in visual morphological identification decreases, maintaining the accuracy and integrity of these extensive collections becomes increasingly challenging (Godfray, 2002, Bik, 2017). One of the unavoidable consequences of these challenges is the general reduction in time and expertise dedicated to the maintenance of collections and specimen label information; including the time needed to properly curate an increasing number of new specimens deposited at museums. This can result in out-of-date taxonomic information, missing or illegible labels, incorrect species identification, and/or errors in database entry.

The exact number of mislabelled specimens or other label errors is hard to define and will be collection dependent. Some groups have been well studied and kept up to date, with rich

histories and knowledge associated with them (Salmon, 2000), while others can be severely lacking in knowledge and expertise. For instance, a recent study found that 58% of tropical plant specimens they reviewed were misidentified and estimated that 50% of all tropical plant specimens are likely to be mislabelled within NHCs (Goodwin et al., 2015). The authors indicate that this is due to the large influx of specimens deposited since 1970 and the lack of taxonomic experts with the knowledge base required to classify them. Regions in the tropics and developing countries, characterised by high biodiversity and complex environments have historically been under sampled, leading to a lower knowledge base associated with them compared to other areas (Moura and Jetz, 2021). NHCs also hold the exciting possibility of containing undiscovered species (Parsons et al., 2022). These species may be hidden under incorrect labels or overlooked because of their scarcity and strong morphological resemblance to known species. The minor differences distinguishing these species can be difficult to detect through standard examination, especially when they are closely related (i.e., cryptic species).

All the above underscores the need for accurate identification methods in collections, whether for curatorial purposes or biodiversity discovery. Modern methods for species identification, like genetic analysis, offer accuracy but come with high resource demands (Shendure et al., 2017). Applying genetic analysis to entire larger collections could lead to astronomical expenses and extensive time requirements. Moreover, the DNA in historical or dried specimens is often degraded thus providing less information than that of fresh or well-preserved samples and requires more robust genetic-based examinations (Marinček et al., 2022, Molbert et al., 2023, Rayo et al., 2024). Furthermore, many historical specimens are deemed to be too important for destructive sampling. As such, extracting DNA from these specimens is not always viable. In addition, many museums have embarked on the mass digitisation of their collections, a step that serves multiple purposes (Hardy et al., 2023). Digitisation not only preserves the physical integrity of specimens but allows them to become readily available for researchers across the globe, fostering wider collaboration and analysis, significantly enriching our understanding of biodiversity and natural history.

In parallel to the mass digitisation of collections is the major advancement of AI which has the potential to revolutionise the way collections are analysed and utilised (Groom et al., 2023). In particular, CV methods can be used for rapid species identification (Hollister et al., 2022), pattern recognition, and morphological analyses (Hollister et al., 2023). The careful coupling of CV with digitised NHCs can bring unprecedented efficiency, accuracy and speed to species identification, which is a core component of collections management and museum-based research.

Beyond verification, CV opens a myriad of possibilities for diverse research projects, ranging from tracking phenotypic changes with temperature (Wilson et al., 2023) to understanding complex ecological interactions (Johannes et al., 2024). The integration of CV into natural history research could not only streamline labour-intensive processes of verifying the integrity of the organisation of collections but also paves the way for innovative methods of exploring and interpreting the vast datasets these collections represent. As AI continues to evolve, it promises to unlock new dimensions of knowledge and collaboration in the study of biodiversity (Karbstein et al., 2024, Borowiec et al., 2022, Seeland et al., 2019, Wäldchen and Mäder, 2018). A CV-based system or assistive tool could help alleviate some of the burden of managing large NHCs by scanning large collections of digitised specimens at high speeds, highlighting discrepancies leading to a streamlined and more accurate verification processes.

One of the first massive digitisation projects was the 'iCollections', a programme undertaken by the NHM, London to digitise its collections of British and Irish butterflies (Paterson et al., 2016). The data captured includes species name, georeferenced location, collector, and collection date, along with a digital image of each specimen and a scale for size reference. This initiative is part of a broader NHM programme to digitise its vast collections, comprising approximately 80 million specimens and objects. The iCollections data have been used to address various scientific questions, such as how climate warming might affect species distribution, phenology, and body size (Wilson et al., 2023, Fenberg et al., 2016, Garner et al., 2024, Blagoderov et al., 2017). The digitised data has been made publicly accessible through the NHM data portal, offering valuable resources for researchers, conservationists, and the public.

Our research is focused on developing an advanced image classification pipeline specifically engineered to identify incorrectly labelled specimens at the species level within the iCollections. Utilising our pipeline, we can detect instances where specimens, presently labelled as one species, are consistently predicted by the system to belong to a different species. These flagged specimens are then organised and presented for a streamlined visual verification process by collection staff. In scenarios where a definitive determination remains inconclusive, we integrate more traditional methods such as reviewing ecological data associated with specimens (sample location, collector, and/or the geographic range of specimen) and when a conclusive answer is unable to be obtained, we utilised molecular methods to ascertain final verification. This blend of AI-driven analysis and more traditional techniques not only streamlines the verification process but also significantly contributes to the integrity and reliability of NHCs in the ever-evolving landscape of biodiversity.

3.3 Methodology

3.3.1 Data set creation and image preprocessing

The iCollections dataset comprises the British and Irish Lepidoptera (Lepidoptera Linnaeus, 1758) collections housed at the NHM. We split the collection into the butterflies and moths. Both groups were filtered to only include species where the total number of specimens was equal or greater than 400 per species, allowing for a sufficient number to train (250 images), validate (50 images) and run inference with the remaining images (≥100). Low numbers of training specimens have been shown to result in poor CV performance (Xu et al., 2023, Buslaev et al., 2020, Shorten and Khoshgoftaar, 2019). This threshold was not intended to be restrictive, but rather to ensure sufficient sample size for stable model training and evaluation. The majority of taxa in the dataset exceeded this threshold, and no taxa were excluded on biological grounds. The filtered butterfly dataset comprised 59 species and a total of 127,671 individual specimens while the moth dataset comprised 283 species, with a total of 222,537 individual specimens. Both training and validation images were synthetically augmented four times by the application of rotations, zooms and slight brightening, thereby generating varied synthetic images; augmenting datasets in this manner has been shown to enhance CV performance (Shorten & Khoshgoftaar 2019; Khalifa et al. 2022).

3.3.2 Model architecture and training procedure

We utilised a VGG16 (Simonyan and Zisserman, 2014) base with a custom selection of top layers, totalling 26 layers. This model used the ImageNet weights for the initial foundational learning, leveraging the pre-existing knowledge embedded within the base model. In the initial phase of training, the VGG16 base was maintained in a locked state, focusing the learning process on the custom top layers for a duration of five training runs. Then for the fine-tuning phase, the remaining layers were unlocked except for the bottom 8 layers. This was allowed to run indefinitely but had a strict ‘early stopping’ protocol that would cease training after 1 decrease in the validation accuracy score and would save the best weights once finished. Furthermore, the hyperparameters of the custom top layers of the model were optimised using the 'TF-keras-tuner' library. The resultant optimum values obtained from this process were consistently applied across all runs and across both moths and butterflies, ensuring uniformity and precision in our approach. Additionally, all model runs were seeded with the same value to ensure reproducibility and to initialise each model with identical starting parameters and neural network weights. This would also mean that when a respective trained model is used for inference, it will always give the same prediction results.

3.3.3 Dataset cropping

Initial trials of the dataset and model architecture employed a heat-map–based class activation mapping (CAM) system to verify that the neural network within the trained model utilised features upon the specimens rather than to irrelevant background noise. The 'Grad-CAM' system was selected for this purpose because it can visualise the pixels and regions that contribute most strongly to the prediction of the model by scoring pixels and overlaying a heat-map colour system based on this score (Selvaraju et al., 2020). Hollister et al. (2023) showed that properly trained CV models combined with heat-maps can highlight the morphological features that distinguish closely related species.

During preliminary tests, many heat-maps concentrated on the specimen labels instead of the insects themselves (Fig. 9A). To mitigate this, we implemented a separate preprocessing pipeline using the YOLOv8 object detection algorithm trained specifically to detect Lepidoptera specimens (Sohan et al., 2024). The pipeline crops each image using the bounding boxes returned during inference, thereby excluding most irrelevant background. Subsequent heat-map analysis of these cropped images showed a model's attention was now appropriately focused on the specimens rather than on the labels (Fig. 9B).

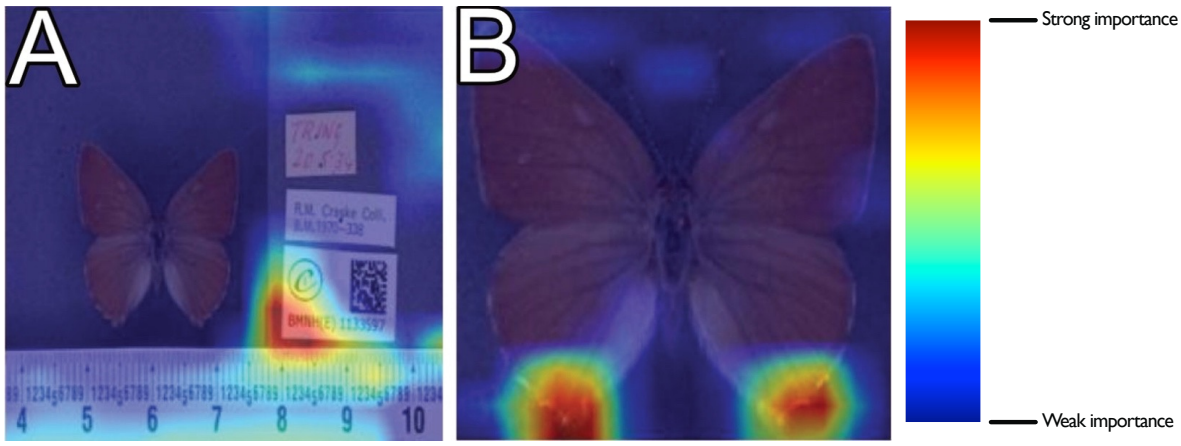


Figure 9. Example of heat-map attention on labels (A) vs directly on the specimen (B).

3.3.4 Pipeline development

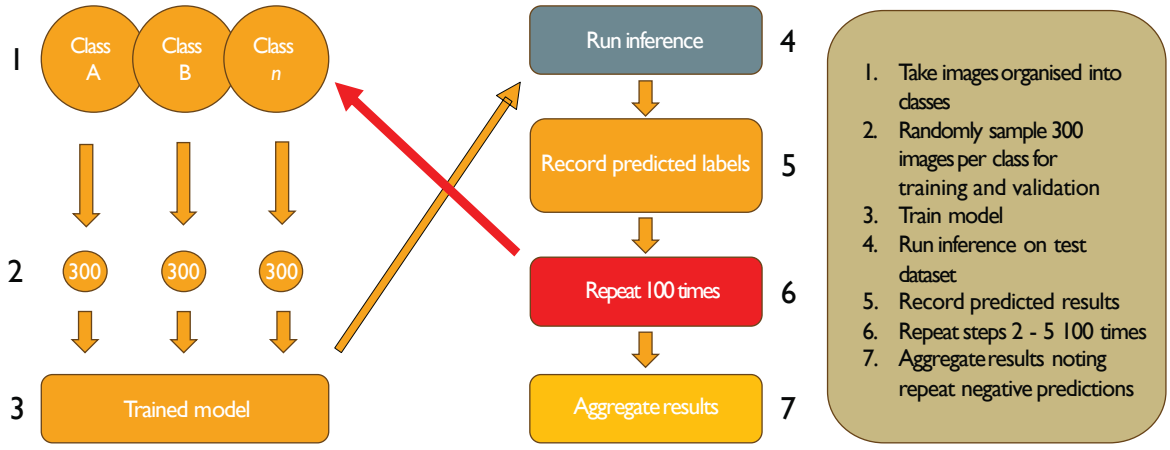


Figure 10. Flow diagram showing the pipeline process.

Our pipeline identifying specimens that do not match their labelled species status is shown in Fig. 10. The butterfly dataset comprised 59 species and the moth dataset 283 species, where each species is one class (step 1). For every run of the pipeline, 300 images were sampled at random from the full set of images for each class, of which 250 were reserved for training and 50 for validation (step 2). Then the images were augmented, and the model was trained and validated (step 3). All remaining images (109,971 butterfly and 137,637 moth) create the test set. The trained model then performed inference on the entire test dataset using TensorFlow's *Evaluation* protocol. This assigned each test image the label with the highest confidence score and compares it with the species label it is currently assigned to determine whether the prediction was correct (step 4) and was recorded (step 5).

Steps 2 to 5 were repeated 100 times, each repetition sampling a fresh training and validation subset (step 6). Because the test pool vastly exceeded the training and validation pools required for an individual run, there is a probability that images appeared in the test set several times. Across the 100 runs, the number of times the pipeline classified a specimen image as a different species label from its current species label was counted. When this misclassification was found to be designated as the same label on each of the trained models, this value was noted and was designated the image's 'Reoccurring Prediction Value' (RPV) (step 7). For example, if the pipeline classifies a specimen as species A for each of the 100 pipeline runs, but its current species label identifies it as species B, then it is assigned a RPV of 100.

3.3.5 Human interrogation

Taxonomists specialising in morphological identification of Lepidoptera from the NHM, with a combined expertise spanning over 50 years, were enlisted to help inspect specimens flagged by the model. Specifically, they were tasked with looking at specimens that the model identified as belonging to a species that is different from its current NHM species record. They were tasked with visually inspecting specimens that were flagged by the pipeline from within the NHM collections. They were told to verify specimens according to four options;

1. Labelled Wrong: The specimen was incorrectly labelled in the collection.
2. Pipeline Wrong: The pipeline made a mistake and incorrectly predicted a specimen as different species to that which it was labelled as in the collection.
3. Portal Wrong: The specimen was correctly labelled in the collection, however, it was incorrectly labelled as the wrong species (or not present) upon the NHM data portal
4. Unknown: The experts were unable to verify what the specimen was or that it was currently inaccessible.

They also added notes to each specimen examined, noting what could have resulted in either of the four choices. To visually inspect every specimen across the two groups would have taken a very long time for the small team of experts. Therefore, it was decided to go through a sample of the specimens with RPVs >80, allowing for a review of the most likely mislabelled specimens. Additional specimens with RPVs <80 were also examined. The examinations were conducted over 4 sessions with an allotted time of 16 hours. This resulted in a total of 210 specimens being examined.

3.3.6 Note standardisation

Notes and comments were standardised. Each specimen was assigned a visual-difficulty score as follows:

1. Easy to verify with the naked eye.

2. Difficult, but not impossible, to verify.

3. Difficult; required additional contextual information (e.g. sampling location, date, or size relative to the predicted species).

4. Impossible to verify visually; referred for further confirmation.

3.3.7 Genetic verification

Specimens unable to be verified visually (category 4 above) were designated for genetic verification. However, several additional specimens not in this category were selected to allow for validation of the visual based verification conducted by the experts. DNA was extracted in a dedicated historical DNA facility using the protocol outlined by Hall et al, (2023), with NGS library building following the protocol detailed in Marsh et al. (2025), using the "Santa Cruz Reaction" (Kapp et al., 2021) with the modifications of Nguyen et al. (2023). Libraries were shotgun sequenced on an Illumina NovaSeq XPlus 25B lane with a commercial provider, targeting 5-10 million PE reads per specimen. The COX1 barcode gene was recovered using MitoGeneExtractor (Brasseur et al., 2023) which uses exonerate (Slater and Birney, 2005) to map reads to a target reference, in this case the closest reference sequence available on NCBI protein database along with ~40 common contaminant sequences (i.e. bacteria, fungi, human, wolbachia) to help filter out non-target reads.

3.4 Results

3.4.1 Pipeline results

The 100 butterfly model runs achieved a range of F1-scores between 0.9497 and 0.9267 and the 100 moth model runs achieved a range of F1-scores between 0.8486 and 0.8386. The F1-score is the harmonic mean of precision and recall, and provides a balanced measure of classification performance. Out of the original 127,671 butterfly specimens, 17,562 individual specimens were flagged by the model at least once across all 100 runs. The number of specimens that received a RPV of one greatly outnumbers the number of specimens that received a RPV of 100 (Table 1). When the RPV are combined into intervals of 10, over 83% of specimens are categorised with an RPV of 1-10, with the next interval of 11-20, occurring over 6%. Less than 1% of specimens flagged by the pipeline occurred in the RPV interval of 91-100. Out of the original 222,537 moth specimens, 81,788 individual specimens were flagged by the model at least once across all 100

runs. Again, the number of specimens that received a RPV once outnumbers the specimens that were received a RPV of 100 (Table 2). Over 80% of specimens flagged by the pipeline occurred in the RPV interval of 1-10, with the next interval of 11-20, occurring over 9%. Just over 0.1% of specimens occurred in the interval with a RPV of 91-100.

Reoccurring prediction values (RPV) for butterflies and moths **in intervals of 10 for the butterflies and moths.**				
RPV Interval	**Butterfly: Number of Specimens**	**Butterfly: Percentage of Total**	**Moth: Number of Specimens**	**Moth: Percentage of Total**
91-100	171	0.97%	94	0.11%
81-90	157	0.89%	255	0.31%
71-80	129	0.73%	255	0.43%
61-70	121	0.69%	446	0.55%
51-60	179	1.02%	613	0.75%
41-50	242	1.38%	1088	1.33%
31-40	295	1.68%	1834	2.24%
21-30	507	2.29%	3307	4.04%
11-29	1121	6.38%	8150	9.96%
01-10	14639	83.36%	65652	80.27%

Table 2. Reoccurring prediction values (RPV) for butterflies and moths in intervals of 10 for the butterflies and moths.

3.4.2 Visual verification interrogation

3.4.2.1 Error type analysis

In total, 210 specimens were visually inspected: 120 butterflies and 90 moths. 56.67% of the specimens examined had an RPV >80, meaning that they were consistently flagged by the model as being incorrectly labelled (Fig. 13). An additional 493 hybrid butterflies were flagged by the pipeline; however, these technically belong to no official species and were verified to be hybrids by the experts, and these were excluded from the remaining evaluations.

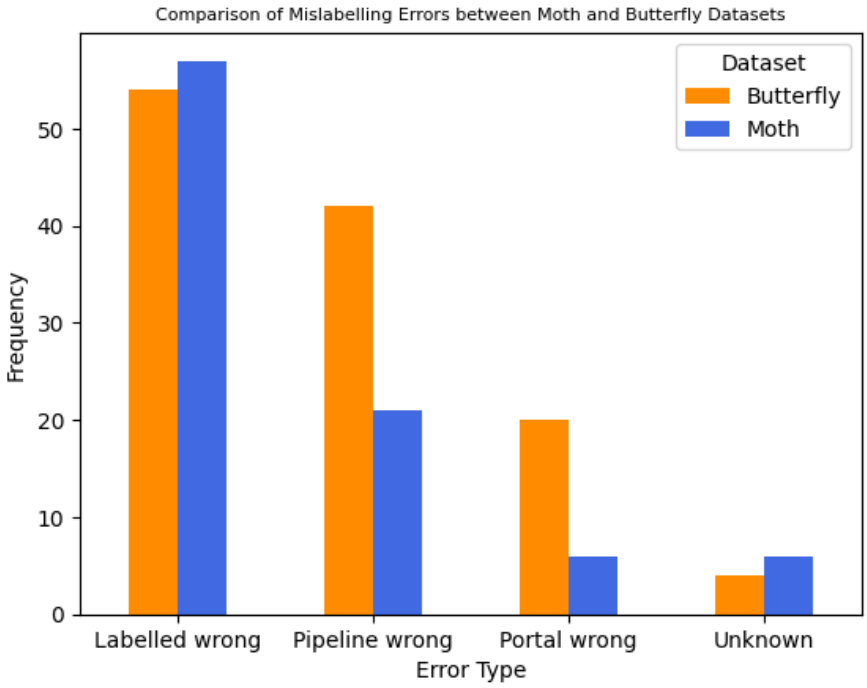


Figure 11. Bar chart showing the combined error results for the butterflies and moths.

The most commonly occurring error among the specimens that were visually inspected by the taxonomists was that the specimens were labelled wrong (54 butterfly, 57 moth) (Fig. 11). This was followed by the pipeline being wrong (42 butterfly, 21 moth), then the portal being wrong (20 butterfly, 6 moth), with the lowest category being unknown (4 butterfly, 6 moth).

3.4.2.2 Difficulty of verification analysis

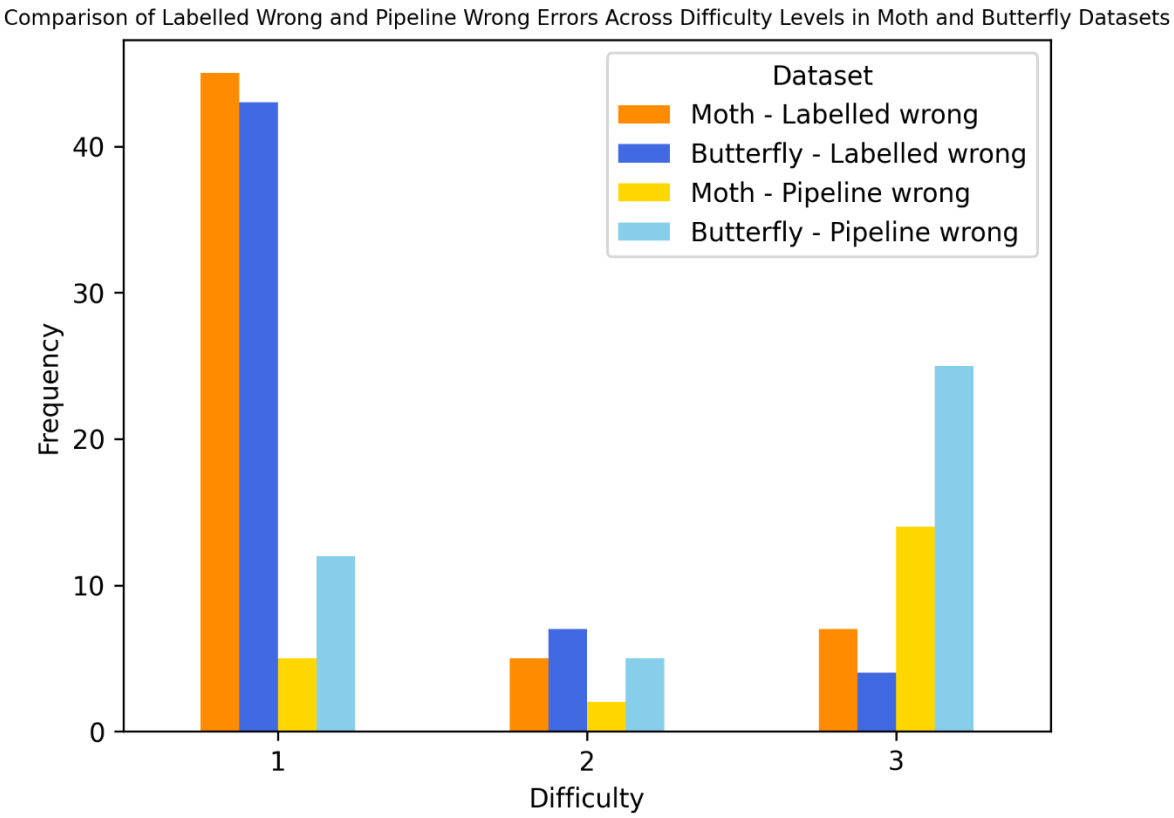


Figure 12. Bar chart showing the difficulty assigned to the visual verifications for moth and butterfly specimens.

In general, specimens that were given a difficulty score of 1 by the taxonomists were more likely to be labelled wrong (Fig. 12). This pattern is seen in reverse when examining verifications with a difficulty score of 3, where the pipeline was more often the reason for the errors. This demonstrates that errors in the labelled wrong category were more likely to be rated as easy to visually verify (score of 1), while errors in the pipeline wrong category were more likely to be rated as difficult to visually verify (score of 3).

3.4.2.3 Relationship between Difficulty and RPV

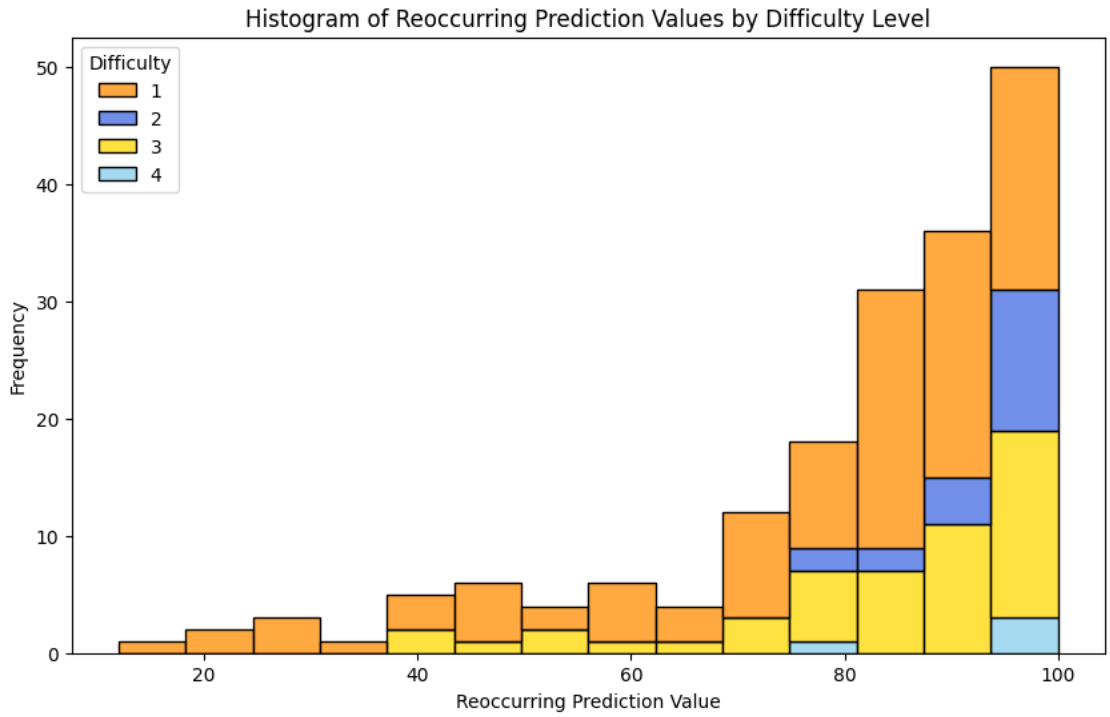


Figure 13. Histogram showing the reoccurring prediction value and difficulty of specimens visually examined.

Most specimens examined had high RPV values, but in general, as RPV decreases, the difficulty level also tends to decrease (Fig. 13). Difficulty Level 1, which contains the most specimens, shows the greatest variability, with prediction values distributed across the entire range. In contrast, Difficulty Levels 3 and 4 are more prevalent among specimens with higher RPVs.

3.4.2.4 Examples of verified labelled wrong specimens

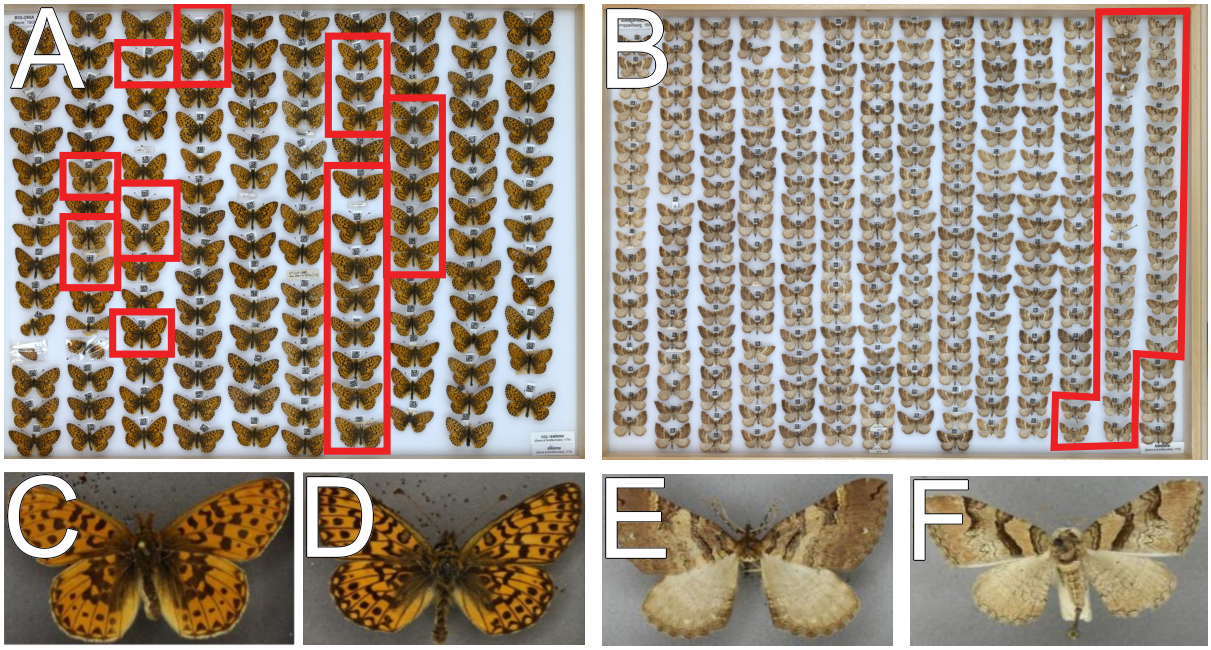


Figure 14. Whole drawer images (A & B) showing labelled wrong specimens (C & E) and their respective species (D & F).

Here we present two examples of when labelling was incorrect. Figure 14A is a whole drawer of *Boloria selene* (Denis & Schiffermüller, 1775) (Fig. 14C) while those highlighted are *Boloria euphrosyne* (Linnaeus, 1758) (Fig. 14D). Guides dedicated to visual morphology separate these two species based on the pattern of the outside edges of the wings with little else considered to separate specimens (European Butterflies, 2024). However, once the difference was noted, experts found it easy to discern between the two and gave these a difficulty of 1. Figure 6B is a whole drawer image of *Earophila badiata* (Denis & Schiffermüller, 1775) (Fig. 14E) while the highlighted specimens are of *Catarhoe rubidata* (Denis & Schiffermüller, 1775) (Fig. 14F). Visual verification of these specimens was, in the opinion of the experts, easy to discern and gave these a difficulty of 1. Moreover, these specimens were all input by a single curator and again, according to the experts, it was a mistake that should have been avoided.

3.4.2.5 Examples of verified pipeline wrong specimens

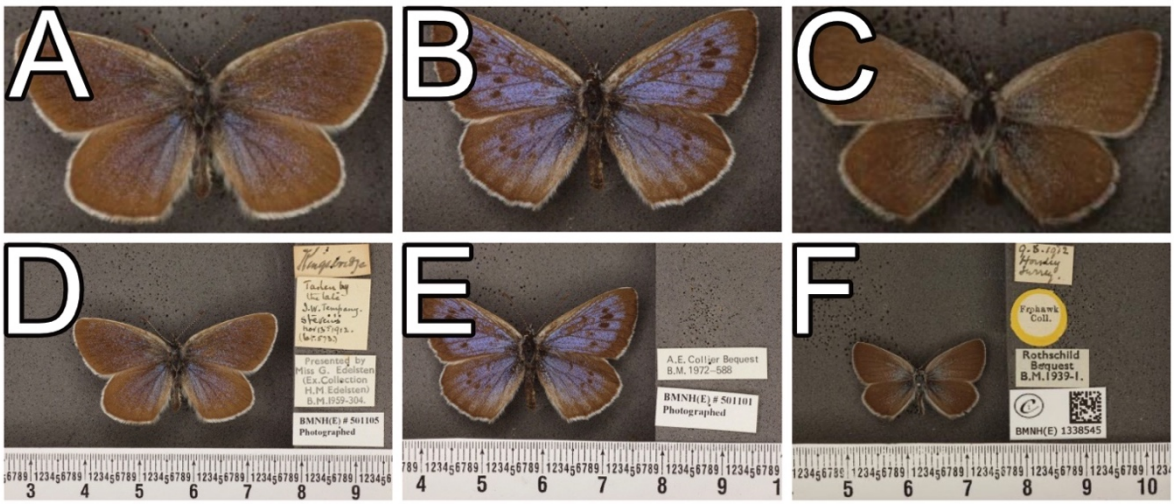


Figure 15. Specimen 'BMNHE_501105' (7A & D) with example of the current species label Maculinea arion (7B & E) and predicted species label Cupido minimus (7C & F).

Specimen 'BMNH(E)501105' (Fig. 15A) belongs to the species *Maculinea arion* (Linnaeus, 1758) (Fig. 7B). The pipeline predicted this specimen as *Cupido minimus* (Fuessly, 1775) (Fig. 15C) with an RPV of 93. Visual verification by the experts confirmed that the pipeline labelled this wrong due to a large size difference between the current species label and predicted species label as can be seen in the images with scalebars and labels (Fig. 15D - F). The experts noted that while the morphology when viewing the cropped images does resemble the predicted species, the specimen in question could easily be verified when viewing it in person or when viewing the image alongside the scalebar.

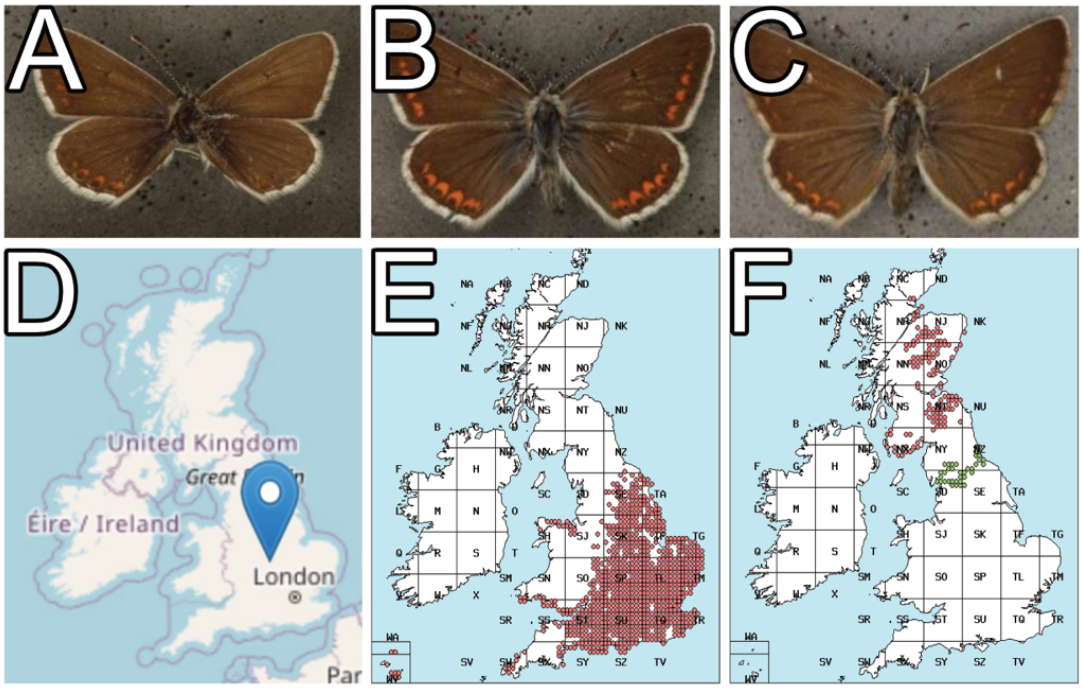


Figure 16. Specimen 'BMNH(E)1390409' (A) and its sampled location (D) with example of the current species label Aricia agestis (B) and collection locations for this species (E). C is an example of its predicted species label Aricia artaxerxes (C) and collection locations of this species (F).

Figure 16A shows specimen 'BMNH(E)1390409' belonging to *Aricia agestis* (Denis & Schiffermüller , 1775) (Fig. 16B). The pipeline predicted this as *Aricia artaxerxes* (Fabricius, 1775) with an RPV of 98 (Fig. 16C). Visual verification confirmed that the pipeline had labelled this wrong because the location that the specimen was sampled from was outside its geographic range. Again, it was noted that while the morphology of the specimen in question resembled the predicted species rather than actual species, the location that the specimen was sampled from would verify that the pipeline predicted it incorrectly. Figure 16D is the location the specimen was sampled from while Figure 16E is the range of the current species label and Figure 16F is the range of the predicted species label.

3.4.2.6 Examples of portal wrong

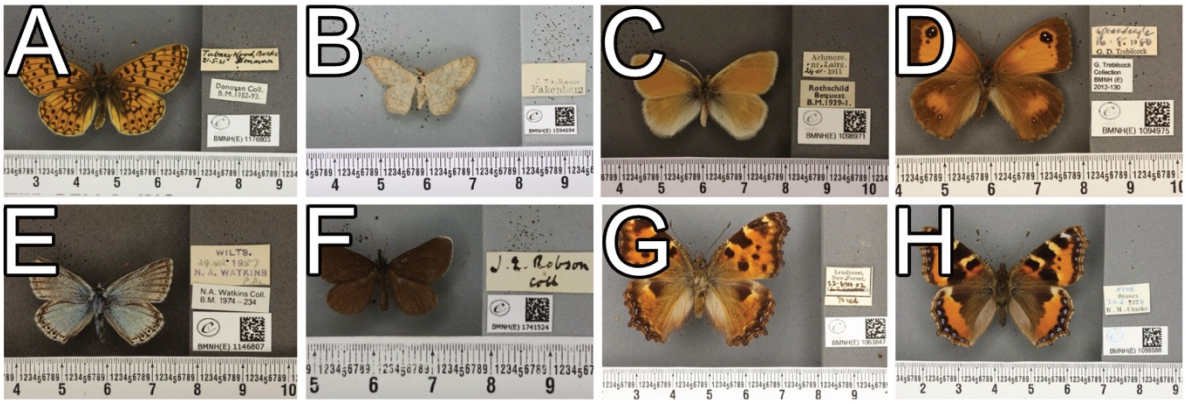


Figure 17. Various examples of issues with specimen storage and retrieval from within the NHM portal.

Figure 17 highlights various errors on the NHM portal in which specimens, their associated information, or their retrieval via the search function from the server storage can be affected. When ID BMNH(E)1176803 (Fig. 17A) is requested on the portal, links for two specimens are retrieved (Fig. 17A and 17B). When ID BMNH(E)1098971 is requested, a single link is retrieved that contains two specimens (Fig. 17C and 17D). Although the ID number matches the ID on specimen 8C, the information on the link belongs to specimen 17D, yet the ID on 17D is different (1094975). Further complicating the mislabelling, specimen 17C, *Coenonympha tullia* (Müller, 1764), is not of the same family as specimen 17D, *Pyronia Tithonus* (Linnaeus, 1758). When searching for ID BMNH(E)1146807, (Fig. 17E), the portal retrieves a completely different ID, and the associated information belongs to specimen 17F. When ID BMNH(E)1063847 is requested, a link for specimen 17G is retrieved. Upon reviewing the information on this link, although the ID number matches the specimen, the attached information (i.e. its taxonomic name, its sampling coordinates, and its drawer number within the collections) belongs to a different species (Fig. 17H).

3.4.3 Genetic verification results

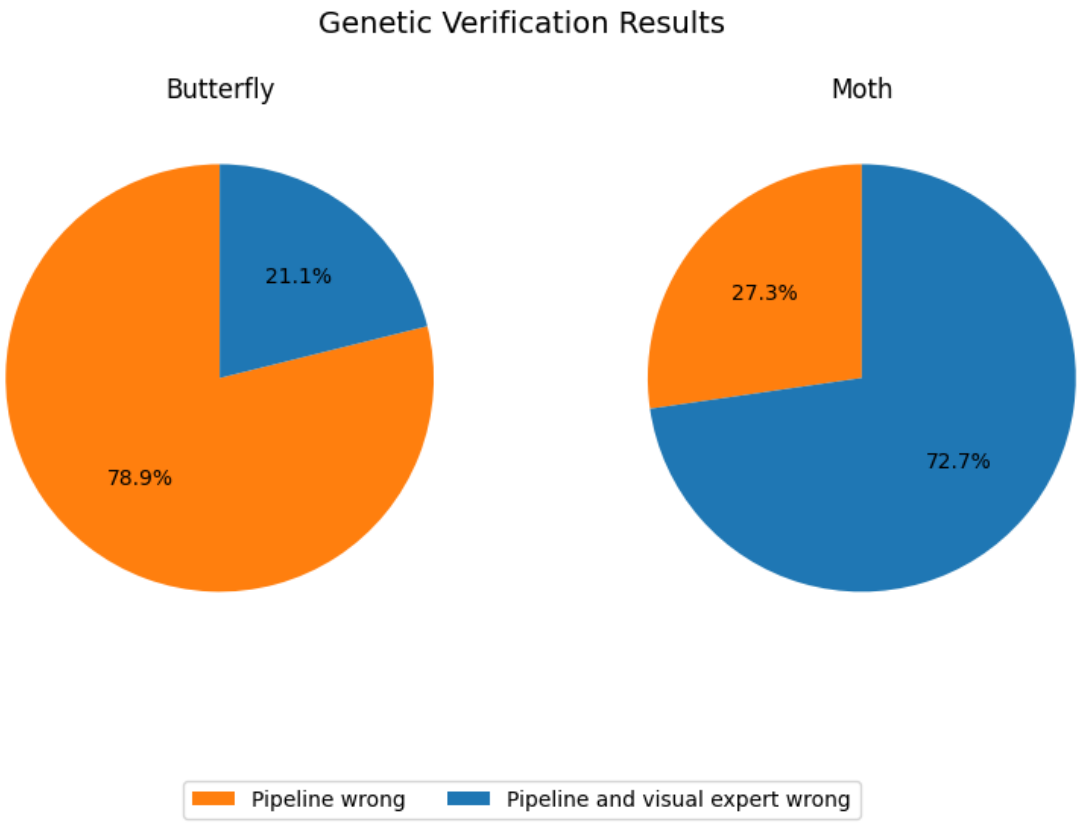


Figure 18. Pie charts showing the results from the genetic verification.

30 specimens were selected for genetic analysis, made up of 19 butterfly specimens and 11 moth specimens (Fig. 18). Among the butterflies, 15 specimens that the visual taxonomists had flagged as incorrectly identified by the pipeline were confirmed as incorrect through genetic analysis. However, 4 specimens contradicted both the pipeline's prediction and the visual taxonomists' assessment, which had supported the pipeline's prediction. For the moths, genetic analysis confirmed that 3 specimens were incorrectly identified by the pipeline, in line with the visual taxonomists' assessment. In contrast, the genetic analysis showed that 8 specimens contradicted both the pipeline and the visual taxonomists. Out of the 4 specimens that were given a difficulty score of 4 (3 butterflies and 1 moth), only two came back from the genetics examination. Both contradicted the pipelines predictions.

3.5 Discussion

A primary challenge of NHMs is the taxonomic identification, curation, and management of vast and continuously growing numbers of specimens (Miller et al., 2004, Mujtaba et al., 2018). Our

study describes a CV pipeline applied to the digitised British and Irish Lepidoptera collection at the NHM, London. This pipeline was developed to automatically identify mislabelled specimens, thereby enhancing the accuracy and efficiency of managing these collections.

Out of the original 127,671 individual butterfly specimens, 17,562 were incorrectly predicted at least once (out of 100 runs, 28.37%). However, analysis of RPV demonstrated that >83% of specimens received an RPV of 1-10, while less than 1% reached an RPV of 91-100. Similarly, for the moth dataset comprising 222,537 specimens, 81,788 were flagged by the pipeline at least once, with over 80% falling within the RPV range of 1-10, while less than 0.12% received an RPV of 91-100. This suggests that while many specimens were flagged by the pipeline, only a small fraction were consistently flagged as being potentially mislabelled. Such specimens (e.g., RPV >80) should be visually inspected by taxonomists and re-labelled if required. This is supported by expert visual examination where 56.67% of the specimens they examine had an RPV >80 (Fig. 12). We confirmed that 147 of the inspected specimens (out of 210; 70%) were indeed incorrectly labelled, either within the collection or during the digitisation process.

In contrast, specimens within the lower RPV ranges (i.e. 1-10) will most likely show pipeline-based errors due to the dataset containing mislabelled specimen images. Although only a small portion of the total number of flagged specimens were examined, 70% of those examined were incorrectly labelled in as either labelled wrong or portal wrong. Even in a scenario where these specimens were the only mislabelled specimens within the dataset, they would ultimately destabilise a CV model's true potential. Research has shown that incorrectly labelled specimens that have been used in the training dataset erode the accuracy of the resulting model (Northcutt et al., 2021). Therefore, it could be assumed that a model known to have incorrectly labelled specimens will undoubtably produce false positive predictions. Future work should focus on this and investigate whether there is a relationship between RPV values and pipeline accuracy. Alternative validation strategies, such as k-fold cross-validation, could also be applied; however, a fixed train–test split was adopted here to reflect a realistic deployment scenario in which models are used to flag individual specimens for expert review rather than to maximise aggregate accuracy metrics.

The mislabelled specimens identified in this study underscores the complexity of managing and curating large NHCs. Our findings align with previous research suggesting that manual labelling errors are not uncommon in such extensive collections with errors reported to be as high as 50% within certain collections (Goodwin et al., 2015). This substantial error rate highlights the critical need for technological solutions (such as that described here) to be used in combination with expert knowledge for the curation and maintenance of large NHCs. Here, we show that automated methods can be used to flag specimens that are potentially labelled differently to their

current status. However, in order to verify and rectify such curation issues, the expert opinion and extensive knowledge of museum curators and taxonomists are needed. We see the collaboration between automated methods and traditional taxonomists as being key for the future curation and maintenance of very large and growing NHCs.

The genetic analysis confirmed that the pipeline made several incorrect predictions, highlighting areas where it aligned with human expertise and also contradicted their predictions. For the butterfly specimens, 15 instances were identified where the pipeline predictions were incorrect, and these errors were accurately caught by the visual taxonomists, showcasing their taxonomic expertise. However, in four cases, the genetic analysis contradicted both the pipeline and the visual taxonomists, indicating that both methods occasionally fail to capture the true identity of certain specimens. Similarly, for the moth specimens, three cases were confirmed where the pipeline predictions were incorrect, and these errors were also identified by the visual taxonomists. In contrast, eight instances showed that the genetic analysis went against the predictions of both the pipeline and the visual taxonomists. These findings suggest that while the pipeline can be effective in identifying potential mislabelling, it is not infallible, reinforcing the importance of a multi-faceted approach to specimen verification.

The synergy of CV, visual, and genetic methods offer robust approaches for managing and curating large NHCs. The combination of these methods is particularly important given the challenges associated with each. Visual verification can be subjective and dependent on the availability and expertise of taxonomic specialists (Austen et al., 2016), while genetic analysis, though precise, can be resource-intensive and sometimes impractical for older or degraded specimens (Karbstein et al., 2024).

Despite the promising results, our study has several limitations. One notable limitation is the current inability of the pipeline to integrate information on specimen size differences or geographical range. For instance, some species may be morphologically similar but vary significantly in size or are endemic to different regions, leading to potential misidentifications by the pipeline. Initial testing showed that the original images which included scalebars and labels interfered with the training of the CV models and resulted in the models occasionally utilising these parts of the images rather than the desired specimen. This was circumvented by cropping the images so that the specimens took up as much of the image a possible, resulting in reduced noise for model training. However, this ultimately resulted in the pipeline unable to differentiate between size as all images are processed as the same size. This limitation suggests that further refinement of the pipeline is necessary to incorporate additional contextual data, such as specimen size and collection location, to improve accuracy. Additionally, experimenting with

systems where CV models focus on specific areas while ignoring excessive noise could be explored.

Our results have revealed a wide range of reasons why specimens within NHCs can become mislabelled, with the biggest being human error. Some specimens showed clear and obvious morphological defining features that should have been, at least in the opinion of the visual-based experts, easy to have been correctly labelled. Due to the age of some of these collections (Paterson et al 2016), the true reasons as to how these errors occurred will never be known. However, current issues where limitations in resources mean that curation staff are unable to dedicate sufficient time to manually verify specimens and manage collections mean that these errors could persist. Specimens that are mislabelled on the portal can also be attributed to human error. The journey of a specimen from its initial input into the collections to its eventual digital representation on the portal would have gone through many different individuals including several generations of curators, photographing teams, or server-level teams, all with varying levels of expertise. The NHM is currently several years into an ambitious project to digitise and upload their NHCs. This highlights that communications from different departments should be a priority when creating such projects and implementing verification steps to avoid errors.

Our study demonstrates that automated methods can be used as important tools for taxonomists and curators to manage very large NHCs. Future work should focus on developing user-friendly interfaces and tools for museum staff and taxonomists to easily interact with and validate the results from the CV pipeline, which could streamline the verification process and free up staff time for other collection management tasks and research.

3.6 Conclusion

In conclusion, our study demonstrates the potential of a combined approach using CV, visual verification, and genetic analysis to significantly improve the accuracy and efficiency of managing NHCs. By automating the initial identification of potentially mislabelled specimens, our CV pipeline offers a scalable solution to the pervasive issue of taxonomic misidentification in large collections. This automation not only enhances the speed and accuracy of specimen verification but also alleviates the burden on human experts, allowing them to focus on more complex tasks that require specialised knowledge.

The integration of AI-driven technologies into museum curation practices represents a significant step forward in preserving the integrity and utility of these invaluable scientific

resources for future research and conservation efforts. Furthermore, our approach underscores the importance of a multi-faceted verification process, combining the strengths of various methodologies to achieve a more reliable and comprehensive system. By continuing to innovate and improve these methods, we can ensure that NHCs remain accurate, accessible, and valuable resources for scientists and researchers worldwide, thereby supporting ongoing biodiversity research and conservation initiatives.

Chapter 4 Genes, shells, and AI: Using computer vision to detect cryptic morphological divergence between genetically distinct populations of limpets

Jack D. Hollister[1,2], David A. Paz-García[3], Rodrigo Beas-Luna[4], Tammy Horton[5], Xiaohao Cai[6], Phillip B. Fenberg[1,2]

1. School of Ocean and Earth Science, National Oceanography Centre, University of Southampton, Waterfront Campus, European Way, Southampton, SO14 3ZH.
2. Natural History Museum, London, Cromwell Road, South Kensington, London, SW7 5BD.
3. Centro de Investigaciones Biológicas del Noroeste (CIBNOR), Conservation Genetics Laboratory, IPN Street 195, Playa Palo de Santa Rita Sur, La Paz, Baja California Sur, 23096, Mexico. Faculty of Marine Sciences, Autonomous University of Baja California, Ensenada, Mexico.
4. Faculty of Marine Sciences, Autonomous University of Baja California, Ensenada, Baja California, Mexico.
5. National Laboratory of Climate Change Biology, SECIHTI, Mexico City, Mexico.
6. National Oceanography Centre, European Way, Southampton, SO14 3ZH.
7. School of Electronics and Computer Science, University of Southampton, University Road, Southampton, SO17 1BJ

4.1 Abstract

Many species are composed of two or more genetically distinct clades, indicating ongoing or past evolutionary divergence. Often however, there are no obvious morphological differences between clades, making it difficult to accurately assess specific aspects of biodiversity or to enact targeted conservation efforts. New advancements in artificial intelligence tools can be used to categorise individuals into their respective genetic clades and to highlight their distinguishing morphological characters that would otherwise be hidden from human observers.

Here, we applied computer vision and explainable artificial intelligence techniques to four limpet species that display well-defined phylogeographic breaks along the Baja California and California coasts. A fine-tuned convolutional network, trained and evaluated over 100 resampling iterations, classified individuals into their genetic clades with median F1-scores of up to 0.96. F1-score performance was markedly higher for true clade groups than the controlled mixed groups, confirming the presence of features specific to the clades. Saliency maps consistently emphasised structures such as the keyhole in *Fissurella volcano* and the ridge tips in *Lottia conus* as distinguishing features, and subsequent shape analyses confirmed significant divergence between clades. These results demonstrate the power of computer vision and explainable artificial intelligence to expose otherwise cryptic morphological diversity and provide a scalable, reproducible workflow that can broaden the biodiversity toolkit and refine eco-evolutionary research across taxa.

4.2 Introduction

Understanding the processes that generate and maintain biodiversity between and within species is a central aim of ecology and evolutionary biology (Govaert et al., 2021). Although genetics, morphology, or a combination of the two are routinely used to delimit taxa and populations, a substantial fraction of diversity remains hidden because genetically distinct lineages may be morphologically indistinguishable, known as cryptic divergence. In such cases, the human eye cannot easily discriminate external traits, and molecular markers are often relied on to provide a diagnostic tool for classifying groups (Lu et al., 2024).

The prevalence of cryptic divergence across the animal kingdom has become increasingly apparent with advances in molecular techniques. A comprehensive meta-analysis of 2,207 cryptic species across major metazoan taxa and biogeographical regions (Pfenninger and Schwenk, 2007), revealed that cryptic species are homogeneously distributed among taxonomic groups rather than concentrated in particular lineages or environments. This finding suggests that morphological stasis upon speciation is seemingly common, independent of phylogenetic relationships or ecological circumstances, and indicates that cryptic diversity predictably affects biodiversity estimates across all animal groups.

There are also high levels of cryptic divergence within species, where there are no obvious morphological differences between genetically distinct populations (Riddle et al., 2000). If there are any morphological differences between cryptically divergent populations within species, it is likely harder to notice them compared to differences between completely separate species,

because the genetic differences are not as significant (Hebert et al., 2004). Given the accelerating loss of species and populations under human impacts, accurate and reproducible methods are urgently needed for detecting morphological differences between cryptically divergent populations or species to inform conservation efforts and provide eco-evolutionary biologists with better tools for understanding genotype-phenotype interactions.

Explainable artificial intelligence (XAI) are techniques that allow for AI model outputs to be reviewed and understood by humans (Aysel et al., 2023). XAI have been particularly advanced within the subfield of computer vision (CV), which now offer a tractable solution to the challenges faced in identifying differences in morphology between cryptic groups (Pinho et al., 2023). *Heatmaps*, a form of CV XAI, provide a visual representation of the regions within an image that most strongly influence model prediction (Aysel et al., 2025). They generate an overlay in which the individual pixels, or spatial areas of pixels, that contribute most to classification decisions are highlighted. One such technique, known as *saliency mapping*, assigns a score to each pixel according to its importance in the final output decision (Alqaraawi et al., 2020, Selvaraju et al., 2020, Jiang et al., 2021). The scored pixels are then rendered in a colour scale, where 'hotter' colours indicate higher importance. These maps therefore serve as a crucial link between model prediction and human interpretation, enhancing the transparency and reliability of automated morphological assessments.

A trained convolutional neural network (CNN) can detect minute shape or colour variations between closely related species, differences which can often elude human observers (Hollister et al., 2025). Furthermore, heatmap systems reveal the pixels most responsible for each classification (Hollister et al., 2023). By coupling automated classification with feature visualization tools, researchers can examine and objectively measure morphological differences, enhancing our understanding of cryptic patterns of biodiversity.

In this study, we developed a CV pipeline to examine four species of limpets with significant population genetic differences along the coasts of Baja California and California (Zarzyczny et al. 2024; Nielsen et al. 2024). Despite clear genetic distinctions between clades, there are no outward morphological differences that have been noted, either in the literature or from field observations of the co-authors, suggesting cryptic divergence. Our primary goal was to train a CNN to accurately classify shells to their respective genetic clades. Next, we aimed to uncover specific features that contribute to this accuracy using saliency maps, which helped guide our analyses and interpretation of shell shape differences between clades. These results can then be used to hypothesise on the potential eco-evolutionary reasons for the morphological differences between clades. Using our explainable AI pipeline, researchers will be in a better

position to explore phenotypic differences between cryptic groups and their ecological and evolutionary significance.

4.3 Methods

4.3.1 Species Selection and Genetic Clade Classification

The following four limpet species were analysed: *Fissurella volcano* (Reeve, 1849), *Lottia conus* (Test, 1945), *Lottia gigantea* (G.B. Sowerby I, 1834) and *Lottia strigatella* (P.P. Carpenter, 1864). The *Lottia* species are members of the Patellogastropoda (true limpets) whereas *F. volcano* belongs to the Fissurellidae (keyhole limpets). Despite their superficial morphological similarities, keyhole limpets and true limpets are not closely related phylogenetically. Fissurellidae can be easily distinguished from the true limpets by the presence of their keyhole. Between *Lottia* species, *L. gigantea* can be easily distinguished from the others due to its large size difference. While there can be some confusion between *L. conus* and *L. strigatella* (Hollister et al. 2023) author PBF is an expert in *Lottia* species from the Baja California Peninsula and is able to visually differentiate all species in this study. Within species however, there is high morphological variability (Hollister et al., 2023).

Specimens were classified into their predefined genetic groups based on previous molecular studies which defined distinct genetic clades separated by clear phylogeographic breaks. In *F. volcano*, *L. conus*, and *L. strigatella*, major phylogeographic breaks along the western Baja California Peninsula was identified using the mitochondrial marker *CO1* (Zarzyczny et al., 2024). By contrast, *L. gigantea* shows no breaks with *CO1* or microsatellites (Fenberg et al., 2010) but does exhibit two breaks identified by genome-wide SNP analysis (Nielsen et al., 2024).

Throughout this study, populations north of their respective breaks are termed the Northern clade and populations south of the breaks, the Southern clade. For *L. gigantea* we restricted sampling to specimens spanning the Californian break because material from the more southerly clade is scarce. These clade designations provided the framework for all subsequent morphological and CV analyses.

4.3.2 Specimen Collection

Specimens were obtained from two primary sources: field sampling and natural history collections. This dual approach provided a comprehensive sample across phylogeographic

breaks while accommodating logistical constraints that limited field sampling at every location. Field collections yielded representatives of the Northern and Southern clades of *F. volcano, L. conus* and *L. strigatella*. Additional specimens were sourced from the Natural History Museum of Los Angeles County (LACM), comprising both clades of *F. volcano, L. conus* and *L. gigantea*. All museum material had been morphologically identified by museum taxonomic experts. Specimen counts were: *F. volcano* = 552 (181 Northern and 371 Southern); *L. conus* = 974 (345 Northern and 629 Southern); *L. gigantea* = 352 (162 Northern and 190 Southern); *L. strigatella* = 789 (215 Northern and 574 Southern). Location counts are provided in the supplementary Data (Sup. Table 1).

Each specimen was photographed in dorsal and ventral orientations using a Panasonic Lumix DC-G9 with an OM SYSTEM 90 mm macro lens on a black background under standardised lighting and magnification. Between three and twenty-one photographs (depending on specimen size) were captured per specimen and focus-stacked using Helicon Focus software to produce a single high-resolution image with consistent depth of field for subsequent CV analyses. All shells, whether obtained from field collections or museum holdings, were dry and photographed under identical imaging conditions (camera, distance, angle, background, and lighting) to ensure consistency across sources. Dry shells are generally stable under standard museum storage conditions, and no published reports describe substantive morphological alteration of dry-stored limpet shells over time, although minor surface variation between sources cannot be excluded.

4.3.3 Model Selection and Configuration

For image classification, we employed the VGG16 neural network architecture (Simonyan and Zisserman, 2014), initially trained on the ImageNet dataset (Deng et al., 2009). Custom top layers were added to adapt the model to the specific requirements of this study. Following preliminary testing, hyperparameters were tuned to optimise performance across all species and orientations. The optimised parameters were consistently applied throughout the pipeline. To ensure reproducibility, the pipeline was seeded with a fixed random state. To ensure the robustness of our findings and to verify that the results were not influenced by random chance, each classification configuration was repeated 100 times. In each iteration, the pipeline randomly sampled training (set to 120 images per class) and validation (set to 30 images per class) images from the total pool of available images required for classifier construction from each respective class (Northern and Southern). The remaining images were reserved as test data for evaluation purposes, referred to as ‘Test-full’. To address potential class imbalance within the test dataset, a further subset was evenly sampled across all classes, referred to as ‘Even-test’

(set to 20 images per class). To address potential sub-class imbalance within the clades, where locations that had large numbers have the potential to create location-specific morphological features rather than the desired clade-specific morphological features, we set a limit of 100 maximum specimens per location. We randomly sampled 100 specimens from these locations and put the remainder into the test sets. Images assigned to training and validation batches were augmented to generate synthetic images. This augmentation process, which included operations such as rotation, flipping, and scaling, is well-documented to enhance the performance potential of image classifiers by increasing the diversity of the training dataset (Shorten and Khoshgoftaar, 2019, Xu et al., 2023).

4.3.4 Mixed-Group Validation

To provide a control group, we trained a parallel set of mixed-group models. All network settings were identical to the original clade-based models. For each species, its two respective clade labels were replaced by two synthetic classes created through random mixing: every mixed group contained equal numbers of Northern and Southern images. Each mixed class therefore represented a uniform distribution of the original categories, so any clade-specific signal should be removed.

Performance on these control models serves as a benchmark for model behaviour for several critical reasons established in recent deep learning research. Systematic experiments demonstrated that convolutional networks can easily fit random labelling of training data, achieving near-perfect training accuracy even when no meaningful relationship exists between images and labels (Zhang et al., 2021). However, while these networks could memorize the random associations during training, they achieved test performance no better than random chance, producing an accuracy of ~10% on the 10-class CIFAR-10 dataset (Zhang et al., 2021). This demonstrates that networks learning from natural data with genuine structure behave qualitatively differently from those fitting arbitrary random associations. While deep networks are capable of memorising noisy data, they tend to prioritize learning simple patterns first, and that networks behave differently when learning from structured versus random data (Arpit et al., 2017). This preferential learning of meaningful patterns when genuine structure exists provides the theoretical foundation for using mixed-group controls to distinguish real clade specific signals from spurious correlations.

If our CNN truly exploits clade-specific morphological features, its accuracy could reach high levels on the original task but fall to chance levels 0% for two-class problems) on the mixed-group task. If there are no real morphological differences between groups, both the original and

mixed-group models will perform similarly, with moderate accuracy. This is because the classifier will pick up on false signals caused by factors like imaging differences, batch effects, or technical issues, rather than true clade differences. The mixed-group control effectively tests whether the model is identifying genuine clade-specific features rather than simply memorising arbitrary training examples (Cawley and Talbot, 2010, Ying, 2019, Oyedotun et al., 2017, Advani et al., 2020).

4.3.5 Size Differentiation

Because shell outline and erosion can vary with specimen size (Oróstica *et al.*, 2021; Vasconcelos *et al.*, 2021), any systematic size difference between clades could bias the CV models. To test for such bias, we measured the major-axis length of every shell with digital callipers (mm) and compared size distributions between northern and southern clades within each species using the Mann–Whitney U test. Size ranges overlapped broadly in all cases and none of the pairwise comparisons were significant ($P > 0.05$). Limpets are generally not known to exhibit external sexual dimorphism (Henriques *et al.*, 2017) unless they have protandric hermaphrotidism (Wright 1988). *Lottia gigantea* shows size-related sexual dimorphism linked to protandrous sex change but there are no shell characteristics that distinguish the sexes (Wright and Lindberg, 1982). However, we included a broad range of shell sizes within each clade but excluded the largest specimens and those with heavy erosion, aiming to minimise any potential influence of sex-related and erosion variation in this species. We therefore assume that size-related cues, including those arising from sexual dimorphism, are unlikely to confound the classifier.

4.3.6 Model Attention Interrogation & Shape Analysis

Previous work shows that CNNs trained on morphological datasets can reveal the image regions most diagnostic for classification via XAI heatmaps (Hollister et al., 2023). The SmoothGrad saliency algorithm was used for this study (Smilkov et al., 2017). Preliminary runs confirmed that the saliency maps centred on the specimens rather than the background, validating their use for downstream analysis.

Guided by these maps, we carried out a mask-based shape analysis. Object regions were first detected with YOLOv8 (Sohan et al., 2024) and then precisely segmented with the Segment Anything model (Kirillov et al., 2023). All masks were rescaled to the same major-axis length (1.0)

while preserving aspect ratio, ensuring that subsequent metrics captured shape rather than size. The metrics extracted were:

1. **Circularity -** equals 1 for a perfect circle and declines as outlines become more irregular, defined as: $Circularity = \frac{4\pi\ x\ Area}{(Perimeter^2)}$
2. **Eccentricity** - distance between ellipse foci ÷ major-axis length; 0 for a circle, increasing with elongation.
3. **Solidity** - Area ÷ Convex-hull area; values near 1 indicate a nearly convex outline, lower values indicate pronounced indentations.
4. **Extent** - Area ÷ Bounding-box area; measures how fully the shape fills its minimal enclosing rectangle.
5. **Minor-axis length** - width of the best-fitting ellipse, normalised to the same scale as the major axis (range 0–1).

Visual representatives can be observed in the supplementary data (Sup. Fig. 1).

4.3.7 Data Analysis

All code and models were run in Python, graphs were generated with the Matplotlib package and statistics were generated using the SciPy package. Model performance was evaluated using the F1-score, which is the harmonic mean of precision and recall and provides a balanced measure of classification performance, particularly useful for imbalanced datasets and is defined as:

$$F1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision\ +\ Recall}$$

4.4 Methods

4.4.1 Model F1-Score Analysis

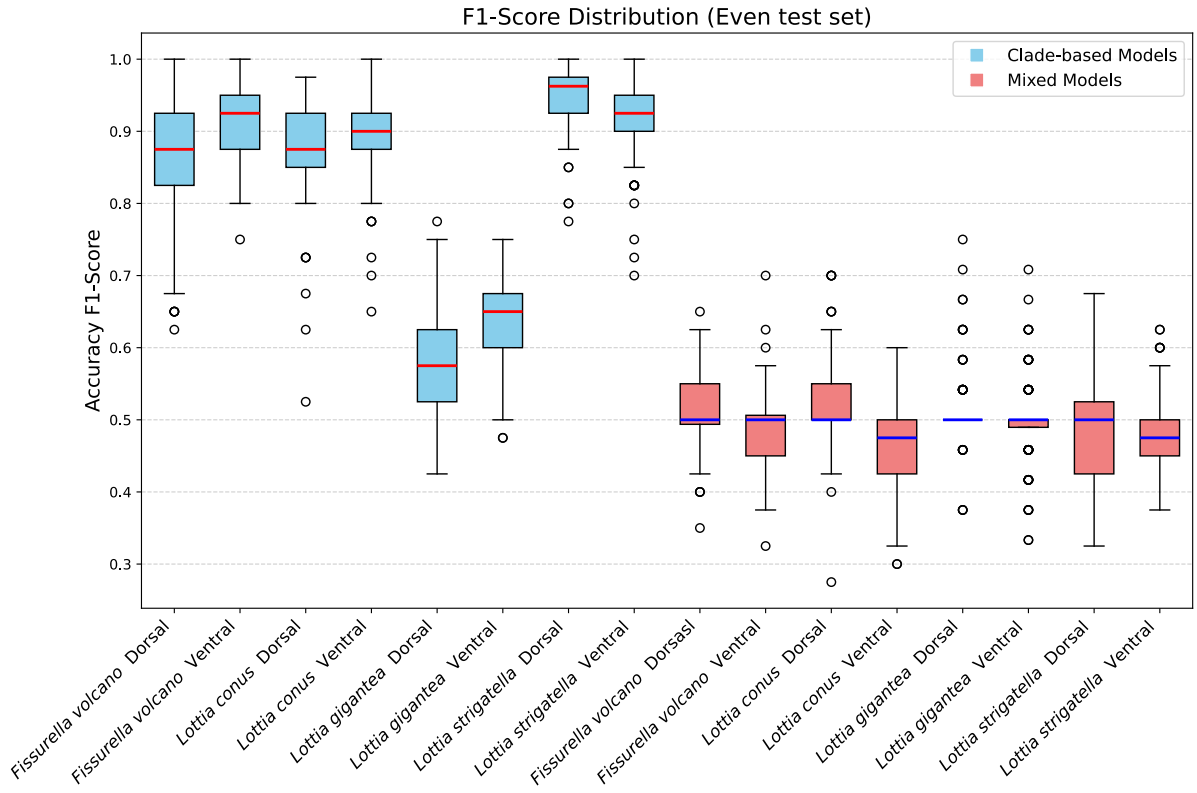


Figure 19. box plots for model combination F1-scores across 100 runs for the even test datasets. Blue boxes are the clade-based models per species and orientation, and the red boxes are the mixed-group controls. For each species and orientation, the F1-scores are significantly greater for the clade-based models compared to the mixed-group controls.

The Even-test results for clade-based models are significantly different from the mixed-group controls (Mann–Whitney U, $P < 0.001$; Fig. 19). Among the clade-based models, the highest median F1-scores were obtained for *L. strigatella* (dorsal: 0.963; maximum 1.000) and *F. volcano* (ventral: 0.925; maximum 1.000). The lowest performance was recorded for *L. gigantea* (dorsal: median 0.575; maximum 0.775). Mixed-group controls consistently underperformed, all of which have median values of ~ 0.500; the largest drop occurred for *F. volcano* (ventral: median 0.500; maximum 0.700).

The Full-test evaluation showed the same pattern (Sup. Fig. 2). *L. strigatella* remained the top performer (dorsal median 0.964; ventral median 0.945), whereas *L. gigantea* (dorsal median 0.575; maximum 0.688) again ranked lowest. For example, the mixed-group control for *L. conus* (ventral) achieved only 0.500 (median) and 0.576 (maximum), compared with 0.904 and 0.938 for its clade-based counterpart, confirming reliance on clade-specific features. Comparing the

Even-test and Full-test scores yielded no significant difference (Kruskal–Wallis test $P > 0.05$), indicating that overall model performance is insensitive to test-set size or class balance. All performance-based subsequent analyses therefore report the Even-test metrics and a boot-strap analysis using these 100 iterations can be seen in the supplementary data (Sup. Table 2)

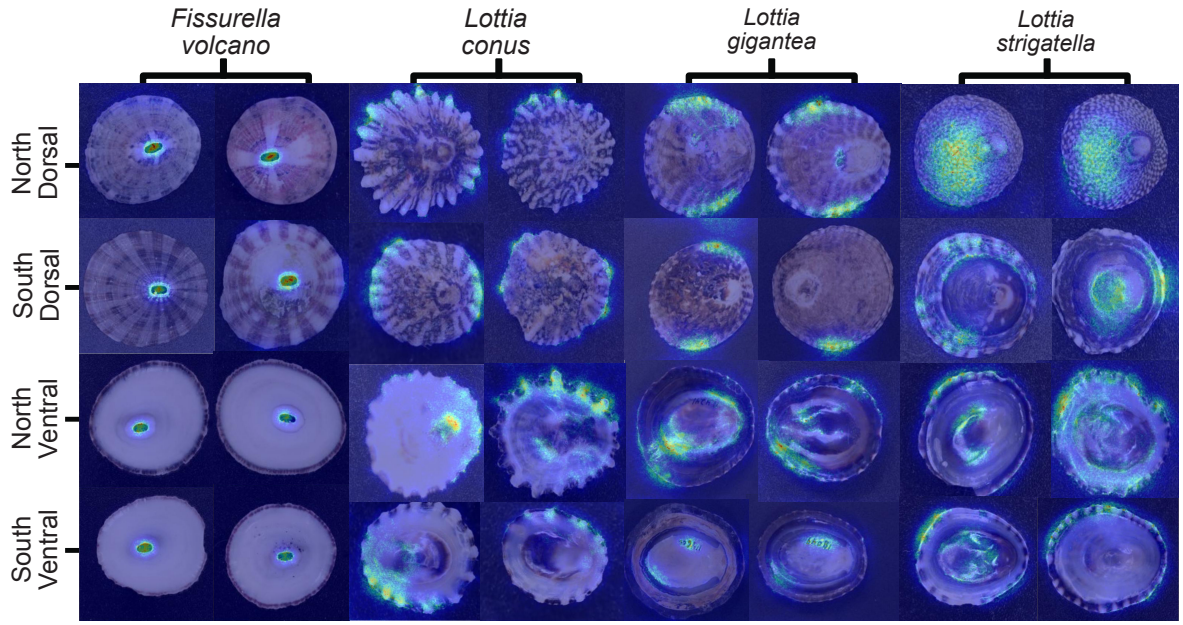


Figure 20. Examples of saliency maps showing all species, groups and orientations. The highlighted parts of the shells are where the model focusses attention for distinguishing between clades. *Lottia gigantea* is larger than all other species, with an average size of 45mm in length for sampled individuals. The average sizes of the sampled individuals for the other species are *L. conus* (9.7mm), *L. strigatella* (9.9 mm) and *F. volcano* (18.3 mm)

4.4.2 *F. volcano* morphological variation

Saliency mapping for both dorsal and ventral orientations consistently highlighted the keyhole (Fig. 20). We therefore selected this structure for detailed, mask-based shape analysis. All five metrics—circularity, eccentricity, solidity, extent and minor-axis length—differed significantly between the Northern and Southern clades (Mann–Whitney U, $P < 0.001$; Fig. 21b). Northern keyholes were less circular, more elongated, more indented, filled a smaller proportion of their bounding box and had a shorter minor axis, whereas Southern keyholes showed the opposite pattern.

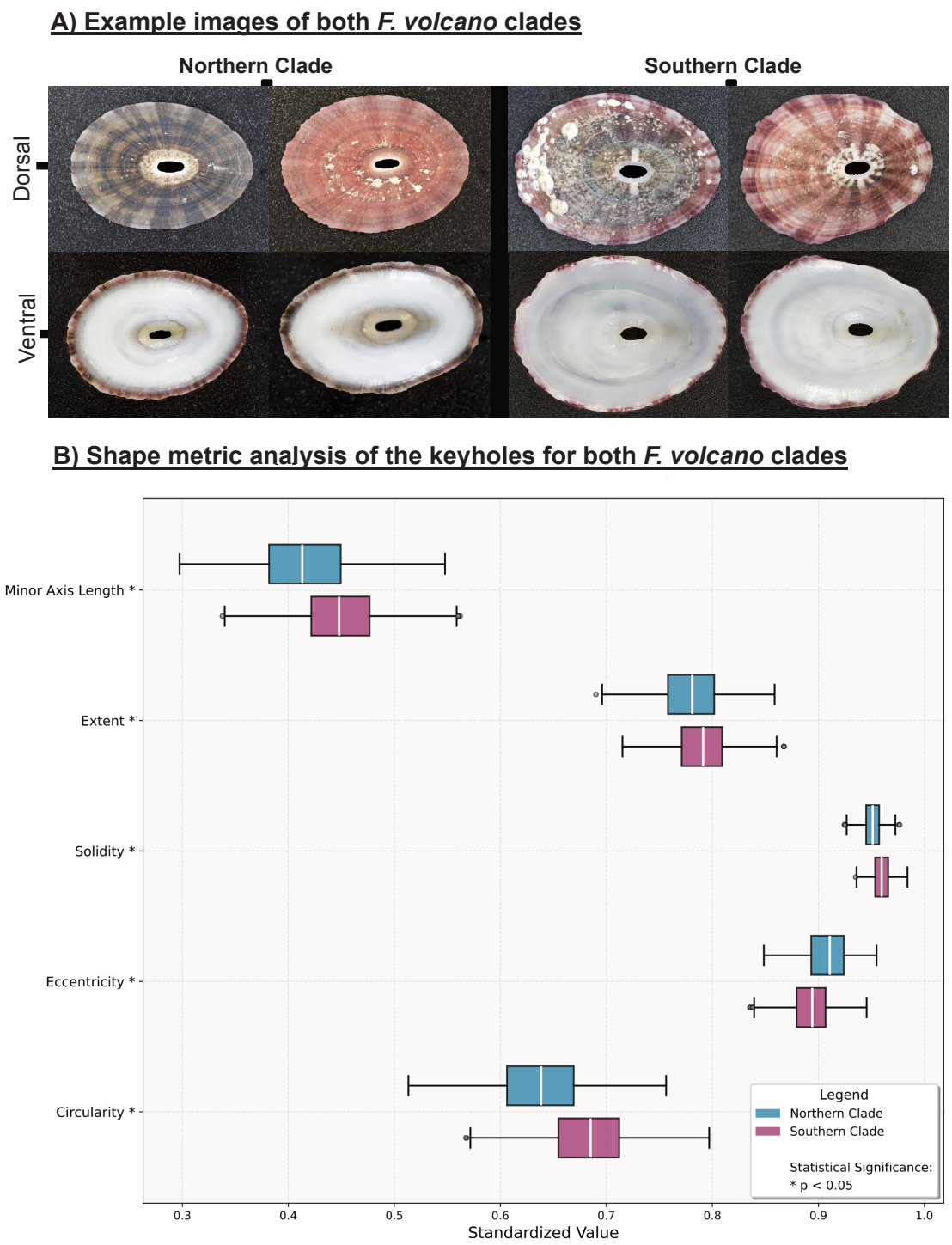


Figure 21. Example images of whole shell specimens from each clade and both perspectives, with the keyholes clearly visible on the apex of each shell (A). Shape metric analysis of *F. volcano* keyholes (B). The saliency maps consistently highlighted the keyholes when distinguishing between clades (see Fig 20).

To further quantify these shape differences, we compared the average outlines of each clade using the Karcher-mean. A Karcher-mean represents the average shape derived from all specimen examinations, providing a single outline that best captures the overall form while accounting for variation among individual shapes. A Karcher-mean shape comparison reinforced

these differences (Fig. 22a). Six corresponding outline points were defined along the shell margin at the 25th, 50th, and 75th percentiles of maximum height on both left and right sides of the shell, ensuring consistent spatial correspondence across specimens for subsequent alignment and shape comparison. These showed that the Northern keyhole is 12.0% narrower along the minor axis, and the mean upper- and lower-quartile distances are 42.5% greater than the central distance, emphasising its more irregular outline. Location-specific Karcher-means display the same clade-level contrast (Sup. Fig. 3).

When the keyhole region (plus a small bounding margin) was cropped from every image and re-submitted to the classifier, the median F1-score fell to ~0.70. When looking at just the keyholes, salience maps were centred on the keyhole perimeter, particularly where inter-clade shape differences occur (Fig. 4b), indicating that the keyhole is the principal, but not sole, feature underpinning discrimination.

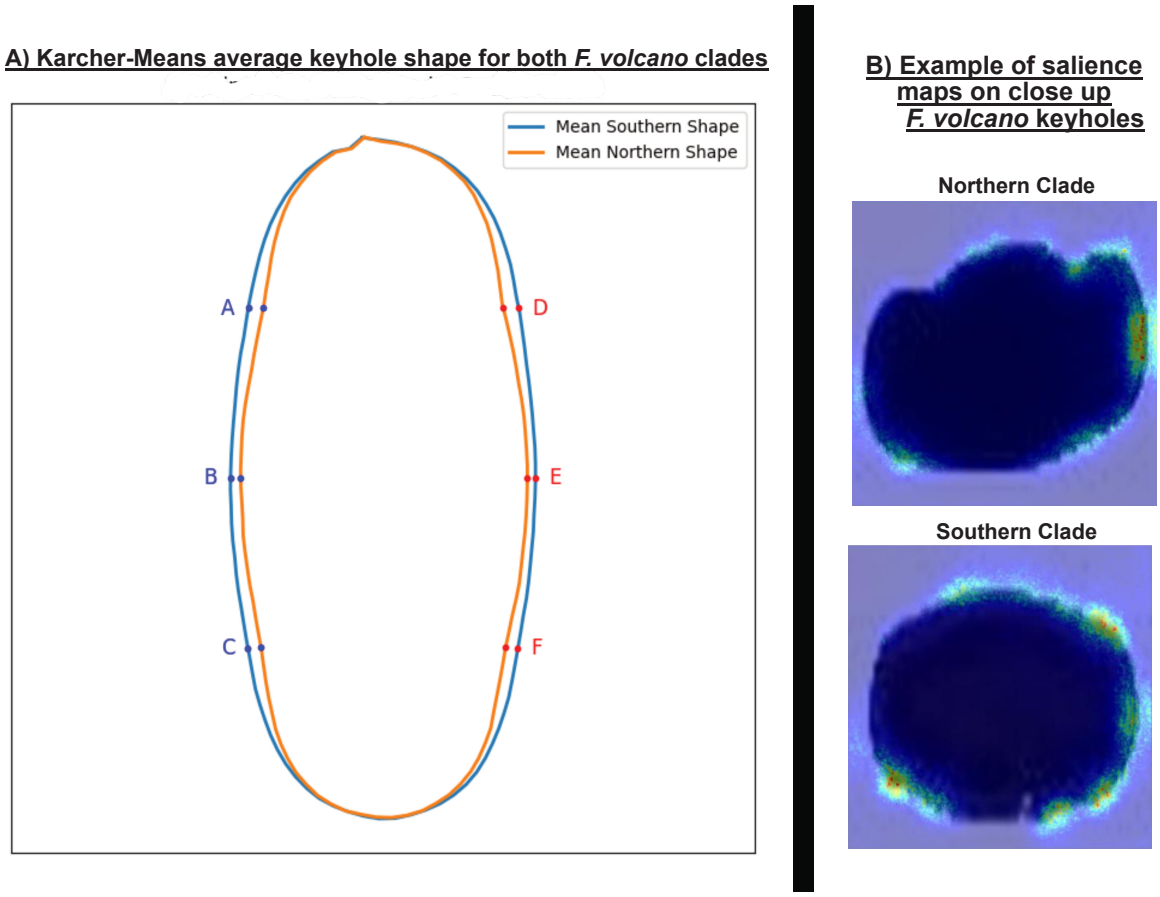


Figure 22. Karcher-mean depiction of the mean keyhole shape of the two *F. volcano* clades (A) and saliency map examples conducted on just the keyholes (B). The keyholes of specimens from the northern clade are more indented and less circular compared to the keyholes of specimens from southern clades.

4.4.3 *L. conus* morphological variation

Saliency maps for *L. conus* converged on the distal ridge tips in both dorsal and ventral orientations (Fig. 20). Guided by this pattern we compared whole-shell outlines between clades. Masks were generated for every shell, rescaled to a common major-axis length, and the standard shape metrics extracted. Circularity, solidity and extent differed significantly between the Northern and Southern clades (Mann–Whitney U, $P < 0.001$), whereas eccentricity and minor-axis length showed no significant variation (Fig. 23b). Thus, Northern shells are on average more irregular, have deeper indentations and occupy a smaller proportion of their bounding box than Southern shells, while overall elongation remains comparable.

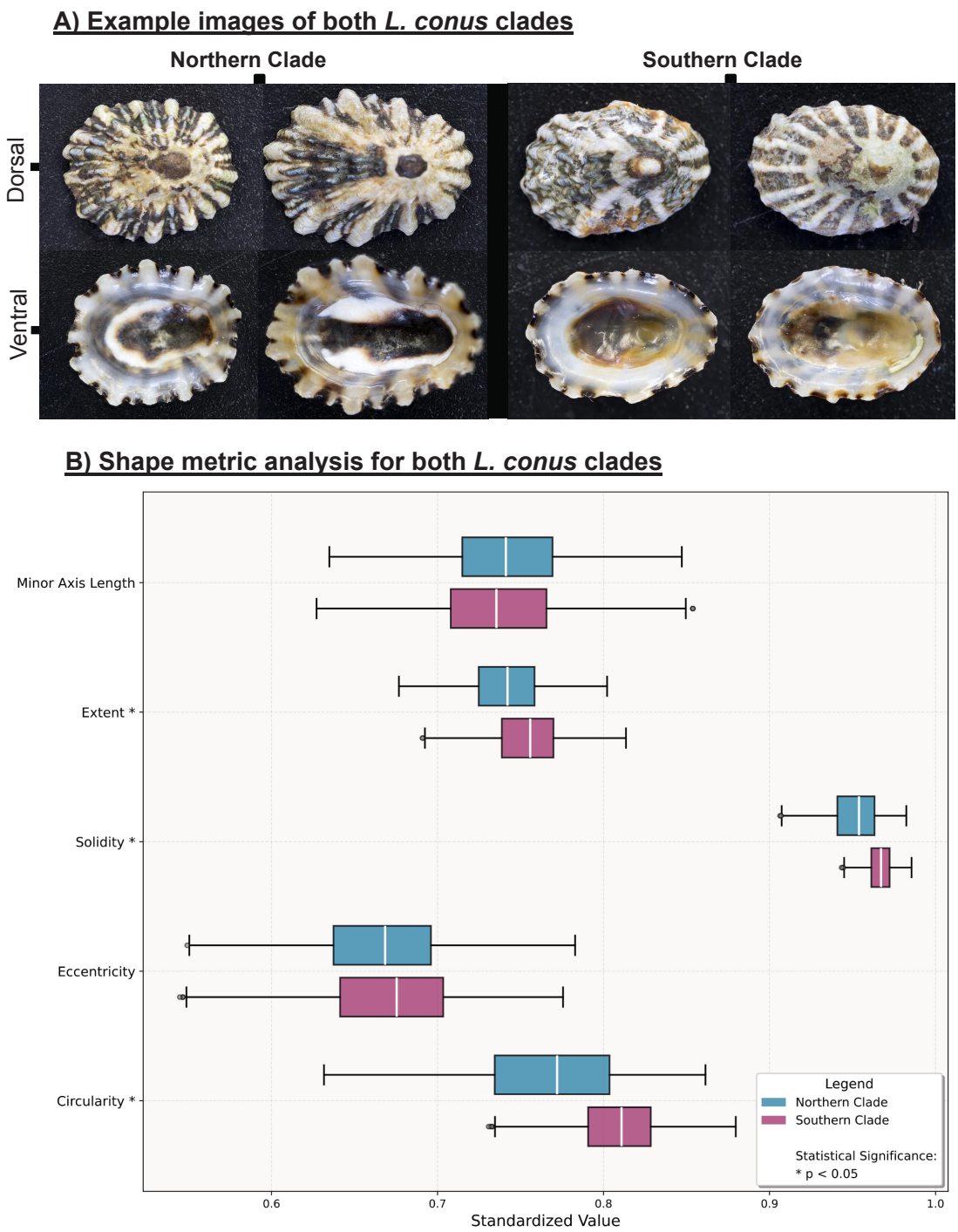


Figure 23. Example images of shells from both L. conus clades and perspectives (A) and shape metric analysis of L. conus shells (B).

4.4.4 *L. gigantea* and *L. strigatella* morphological variation

For *L. gigantea* and *L. strigatella* the saliency maps were diffuse, with attention often distributed along the shell perimeter rather than on a single, discrete structure (Fig. 20). Consequently, we applied the same whole-shell, mask-based shape analysis to these species to quantify any outline differences between clades.

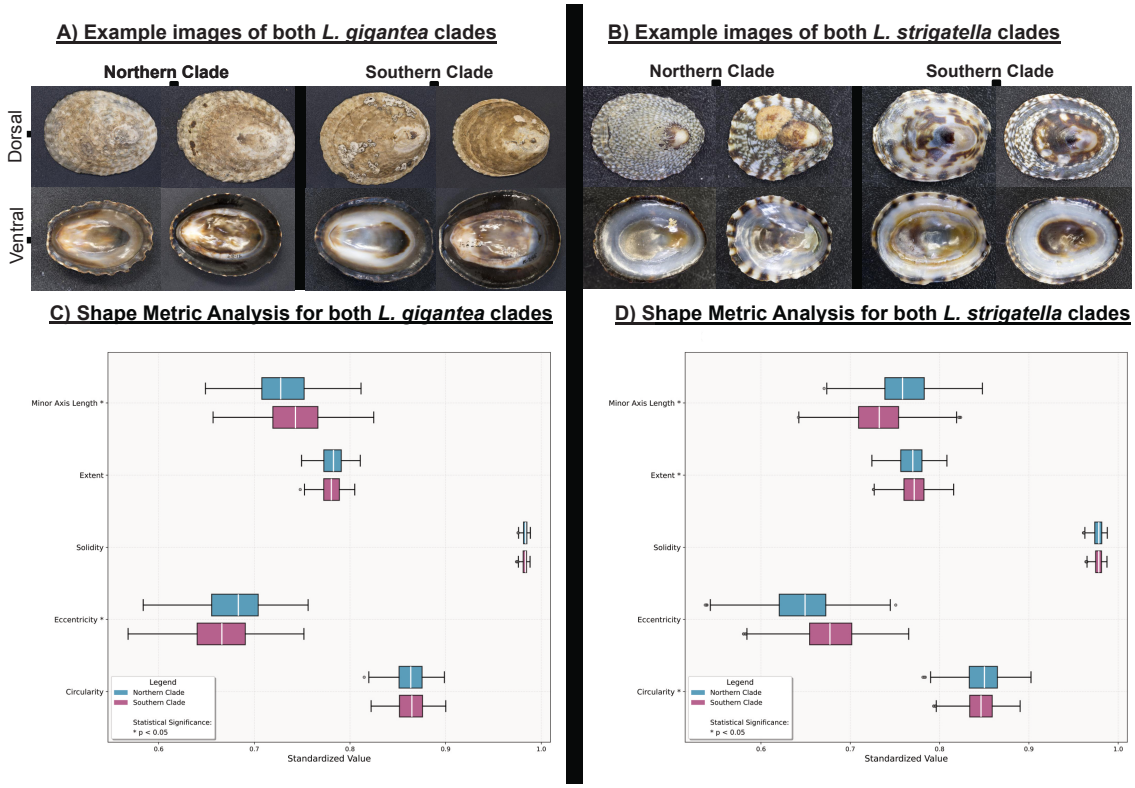


Figure 24. Example images of shells from both *L. gigantea* clades and perspectives (A) and *L. strigatella* (B). Shape metric analysis of *L. gigantea* (C) and *L. strigatella* (D) shells.

For *L. gigantea,* the Northern clade shells were less elongated (lower eccentricity) and displayed a larger normalised minor-axis length than those of the Southern clade, whereas circularity, solidity and extent did not differ significantly. For *L. strigatella*, the Northern clade shells were more regular in outline (higher circularity), had a larger minor-axis length and a lower extent (less compact) than Southern shells, while solidity was nonsignificant between clades.

4.5 Discussion

4.5.1 *CV As A Powerful Tool For Biodiversity research*

Understanding and preserving biodiversity requires advanced methods to detect and analyse cryptic morphological diversity among genetically distinct clades (Grupstra et al., 2024). In this study, we explored the potential of explainable artificial intelligence, specifically CV, to reveal cryptic divergence among genetically divergent clades of limpets. Our findings demonstrate that CV methods can not only accurately classify individuals from four limpet species into their

respective genetic clades (Fig. 19) but also identify and quantify morphological features of clades that were yet to be recognised or reported by human observers. The successful application of CV in this context underscores its value as a powerful tool for biodiversity research, providing new insights into the eco-evolutionary processes shaping morphological traits within cryptically diverse species.

Although demonstrated here with limpet species, the workflow is transferable because it operates on two-dimensional images in which the relevant diagnostic information lies within a single plane. Numerous taxonomic groups satisfy this criterion, including, but not limited to, insect wings (Sauer et al., 2024) and other pinned or slide-mounted materials (Hollister et al., 2022), herbarium sheets (Younis et al., 2020), and thin-section microfossils (Piva et al., 2024). Major natural history repositories, including the Natural History Museum, London, are now digitising such objects at scale, thereby supplying abundant datasets for further investigation (Blagoderov et al., 2012, Allan et al., 2019, Hollister et al., 2025). Researchers wishing to adopt the pipeline therefore need only modest adjustments. For example, substitute an appropriate training image set, fine-tune the hyper-parameters, and rerun the training–validation cycle on a pre-weighted convolutional network. When coupled with heatmap-guided feature visualisation and mask-based shape analysis, which are both easily automated with contemporary detection and segmentation models (Kirillov et al., 2023, Sohan et al., 2024), this approach enables the objective localisation and quantification of defining characters in any planar specimen. Although image augmentation introduced some positional variation, all images were captured under controlled and standardised conditions, with the camera positioned directly above each specimen at a fixed distance and lighting setup. Such consistency is critical for reliable feature detection, as differences in angle, illumination, or scale could alter the appearance of diagnostic features and reduce model reproducibility. Assessing how well the method performs under less standardised imaging conditions would therefore be a valuable future test.

4.5.2 *Detecting Clade-Specific Signals*

Comparisons between the clade-based models and the mixed-group controls show that the pipeline is responding to genuine, clade-specific signals. Mixed-group controls have significantly lower F1-scores (F1 = ~0.5). This means their performance is similar to random guessing for problems with two classes (Zhang et al. 2021). This confirms that F1-score performance in the original models is driven by clade-specific characters rather than generic image features (Khalid et al., 2014, Dhal and Azad, 2022). These characters, though subtle, are evidently consistent enough to support reliable automated discrimination. Most original configurations achieved very

high F1-scores, and several runs reached perfection (F1-score = 1.0) for both dorsal and ventral views of *L. strigatella* and *F. volcano,* while the *L. strigatella* dorsal perspective reached an average F1-score of 0.96 across its 100 iterations. The F1-score performance for *L. gigantea* was appreciably lower (Fig. 19). The smaller difference in F1-scores between clade-based and mixed-group models potentially reflects a relatively recent population divergence in this species. Genome-wide SNP data indicate a population split, but *COI* data show no corresponding division, suggesting that perhaps this separation is evolutionarily young. Such discordance between nuclear and mitochondrial markers often occurs during early stages of population isolation, when genomic differentiation emerges before fixed morphological or mitochondrial differences accumulate (Avise, 2000, Toews & Brelsford, 2012). The absence of pronounced morphological divergence in *L. gigantea* therefore implies that while this species may have some geographically structured genomic divergence, where selection or restricted gene flow has begun to structure genomic variation, phenotypic differentiation has yet to accumulate. This contrasts with the other species, where older, mitochondrial-level separations coincide with measurable morphological divergence, producing higher F1-scores for clade-based models. Note however, our clades for *L. gigantea* were only based on the Californian phylogeographic break which is less clear geographically (in the Los Angeles region) and with lower support than the more distinct break in the central Baja California Peninsula (Nielsen et al. 2024). If more specimens were available for this southernmost clade, the models might have performed better. In addition, the average length of the *L. gigantea* shells analysed here was 44.6 mm, significantly larger than those of the other species: *L. conus* (9.7mm), *L. strigatella* (9.9 mm) and *F. volcano* (18.3 mm). *Lottia gigantea* individuals frequently exhibit pronounced shell erosion (Kido and Murray, 2003, Mann et al., 2012), which may remove some details.

The absence of any significant difference between the Even-test and Full-test evaluations confirms that the morphological signal is independent of sample size or class imbalance. F1-scores nevertheless showed variable dispersion across their respective iterations: the broadest range (0.450) arose in the *L. conus* dorsal models, whereas the narrowest (0.225) was recorded for *L. strigatella* dorsal models, indicating a more stable response in the latter. Variation of this kind is often linked to dataset quality (Gong et al., 2023, Picard et al., 2020). Consistent imaging protocols and balanced classes are critical to reliable performance (Folmsbee et al., 2019, Glučina et al., 2023, Zhang et al., 2022), yet limpet shells are inherently variable and may be eroded, damaged or obscured by surface deposits (Hollister et al., 2023). Although severely damaged specimens were removed, residual heterogeneity remained, so some random training–validation splits inevitably contained fewer informative features. Implementing 100 independent resampling iterations was therefore essential for exposing and averaging over this variance, and similar protocols are recommended when benchmarking image classification pipelines.

However, alternative methods for examining the data are available such as k-fold cross-validation that could have been used to maximise training data usage and reduce dependence on any single partition.

4.5.3 *Relevant Characters For Cryptic Morphological Divergence*

Saliency mapping highlighted the image regions that most influenced the classifier and thus pointed to characters of possible ecological or evolutionary relevance. In *F. volcano*, the maps converged on the keyhole; shape metrics confirmed significant clade differences in circularity, eccentricity, solidity, extent and minor-axis length, with Northern keyholes narrower and more indented than Southern ones. When images were cropped to the keyhole alone, saliency shifted to the aperture perimeter and the F1-score fell to ~0.70, indicating that the keyhole is the principal, but not exclusive, discriminant. Importantly, these clade-specific shapes can be observed at each location (i.e., specimens from each Northern location are narrower and indented and specimens from each Southern location are more oval shaped). This suggests that these morphological differences are clade-specific and therefore likely related to the genetic differences between clades (Zarzyczny et al. 2024), but further research is required.

For *L. conus*, the maps consistently highlighted the ridge tips. Corresponding shape analyses showed that Northern shells were more irregular, had lower solidity and occupied a smaller proportion of their bounding box, signifying greater concavity relative to Southern shells. In *L. gigantea* and *L. strigatella* the saliency maps were less convergent, with attention dispersed across the shell surface and occasionally concentrated along the perimeter. Even so, quantitative metrics detected clear clade-level shell-based divergence. *Lottia gigantea* clades differed significantly in eccentricity and minor-axis length, whereas *L. strigatella* clades diverged in circularity, extent and minor-axis length. The diffuse saliency pattern implies that the discriminating information is spread across the shell or resides in attributes not captured by outline geometry, such as colour bands or surface patterning. This interpretation is supported by the exceptionally strong and stable performance of the *L. strigatella* models, whose high mean F1-score and narrow dispersion suggest additional, non-geometric cues underpin effective classification in that species. While many of the highlighted regions correspond to biologically interpretable shell features, not all saliency responses necessarily reflect genuine morphological signal. Saliency maps identify regions that most strongly influence model decisions rather than features of confirmed biological relevance, and activation can occasionally arise from background texture, residual reflections, or minor lighting differences. This limitation is inherent to most image-based explainability methods and should be considered when interpreting fine-

scale patterns of activation. Furthermore, AI and CV models operate purely as mathematical systems that detect and process numerical patterns; they do not possess an intrinsic understanding of the biological meaning of these patterns. Nevertheless, the overall consistency of highlighted regions across models and orientations suggests that the major areas of importance are clade specific rather than stochastic or artefactual.

The adaptive drivers of the clade-level differences documented here remain unresolved. Furthermore, we do not currently know whether morphological differences between clades are a result of genetic differences or due to phenotypic plasticity (i.e., caused by differences in environmental or ecological conditions between clades). Morphological traits generally evolve under selective pressures arising from environmental conditions, resource acquisition or predation (Boaventura et al., 2002, Vermeij, 1973, Trussell, 1996). In Fissurellidae, the keyhole serves as an exhalant opening for waste or respiration (McLean, 1984). A latitudinal survey across its congener (*F. radiosa*) shows that the keyhole narrows towards the cooler portion of its geographic range (albeit with limited spatial sampling). While clade differences in keyhole shape detected in *F. volcano* generally matches this trend, the precise advantage of keyhole shape remains to be studied empirically. Nor is it known whether keyhole shape is a result of the genetic differences between clades, phenotypic plasticity (i.e., caused by water temperature or other environmental differences between locations/regions) or a combination of both. For *L. conus*, comparable data are limited. Some studies suggest that limpet shell morphology is shaped by the need to maintain attachment to the substrate, influenced by factors such as wave exposure and the physical characteristics of the surface (Sempere-Valverde et al., 2024, Paulo Cabral, 2007). However, these studies note that limpet shells can become more or less conical under different environmental regimes but do not explain why the ridge tips themselves become more elongated and projecting, rather than remaining broadly rounded.

Future research could explore many extensions to the current framework. One practical direction would be to test the classifier on images captured under varying camera angles and lighting conditions to assess how robust the workflow remains under less standardised imaging. Beyond this, future developments in three-dimensional imaging could overcome such limitations by recording complete shell geometry, allowing morphological differences to be examined independently of viewing angle. For instance, three-dimensional imaging that records shell height and curvature would permit finer quantification of morphological divergence and would allow research into subjects that do not sit on a single plane. Morphology-based assessments that incorporate colour or surface-pattern information (Williams, 2017), rather than geometry alone, could expose additional clade-specific characters. More precisely linking morphological differences identified by AI methods to genetic differentiation through targeted genomic sampling could illuminate how genotype-phenotype interactions contribute to observed morphological

variation (He et al., 2024), potentially identifying specific genomic regions that underlie morphological divergence between clades. Together, these advances should deepen our understanding of the trait variation driving evolutionary differentiation and enhance conservation assessments of cryptic biodiversity.

4.6 Conclusion

This study highlights how computer vision, combined with user-friendly AI tools like saliency mapping, can uncover hidden patterns in shapes and forms between cryptic groups. This approach not only enhances our understanding of diversity in nature but also makes complex analysis easier and more precise. The pipeline reliably classified individuals into their genetically defined clades and pinpointed clade-specific shell characters, such as the keyhole in *F. volcano* and the ridge-tip geometry in *L. conus*, demonstrating that it is driven by clade-specific features rather than irrelevant image cues. Although this analysis does not pinpoint the exact reasons behind these differences, it provides a useful framework for exploring subtle variations in any group of organisms, especially when their key characteristics are primarily presented in two dimensions. As natural history collections continue to release large image datasets, the scope for applying explainable AI workflows across diverse organismal groups will grow correspondingly.

Chapter 5 Do You See What I See? Comparing Human and Convolution Neural Network Attention to Butterfly Morphological Features.

Jack D. Hollister [1,2], Geoff Martin [2], Tammy Horton [3], Xiahao Cai [4], Ben W. Price [2], Phillip B. Fenberg [1,2].

1. School of Ocean and Earth Science, National Oceanography Centre, University of Southampton, Waterfront Campus, European Way, Southampton, SO14 3ZH.
2. Natural History Museum, London, Cromwell Road, South Kensington, London, SW7 5BD.
3. National Oceanography Centre, European Way, Southampton, SO14 3ZH.
4. School of Electronics and Computer Science, University of Southampton, University Road, Southampton, SO17 1BJ

5.1 Abstract

CNNs now offer classification accuracy comparable to humans, but the extent to which they attend to the same morphological features that a human would utilise remains unclear. Here, we examined four pairs of closely related British butterfly species using a modified VGG16 classifier trained on a large image dataset. We compared model attention visualised through various heatmap overlays with diagnostic traits reported in three commonly used identification guides and descriptions generated by the large language model, Claude.

Heatmaps consistently emphasised features such as lunules, discoidal spots, and marginal markings, closely aligning with characters used in the literature. Saliency maps focused more on specific features while Grad-CAM maps highlighted broader regions with less focus on specific traits. Paired classifiers, restricted to two species, produced more abundant attention on diagnostic features than full classifiers trained on 59 species, reflecting a trade-off between fine and broad discrimination. Comparisons with Claude outputs aligned with the literature and

the heatmaps, however, additional emphasis on traits not mentioned in the literature such as wing shape were not consistently reflected in the heatmaps.

These findings demonstrate that CNN explainability methods can recover biologically meaningful features and provide a framework for evaluating whether machine attention aligns with morphological features used by humans. This approach has potential applications in validating classifiers, detecting overlooked features, and supporting large-scale biodiversity research.

5.2 Introduction

Butterflies (*Lepidoptera: Papilionoidea*) are among the most visually recognisable insect groups, with depictions dating back to c. 3000 BC in the tomb of Nefer/Kahay (Nazari and Evans, 2015) and early descriptions recorded by Aristotle in 350 BC (Balme and Gotthelf, 2002). Since then, their morphology has been extensively documented, particularly for the purpose of species and group delimitation and ecological studies (Thomas, 2020). A substantial body of literature, including books, manuscripts, online resources, and specialist groups, details morphological differences to aid identification. Morphology, however, is not only used for taxonomic purposes, but also in understanding the biology and ecology of individuals and populations, both in contemporary and historic specimens. Traits such as colouration, markings, and structural features provide insights into feeding, predator avoidance, reproduction, sex, age, condition, size, and health (Dharmaraaj and Kunte, 2025, Oliver et al., 2009).

Traditionally, morphological assessment has relied on expert taxonomists, often specialising in specific groups. More recently, molecular approaches have been used to classify specimens into taxonomic or population units. While molecular methods can achieve high accuracy in delineation, and progress has been made in linking genes to morphology, they typically provide less information on broader ecological characteristics (Hof et al., 2016, Van Belleghem et al., 2021).

CV, a form of AI, has increasingly been applied to classify organisms from images or videos, with performance comparable to human experts (Hollister et al., 2023). These methods can resolve taxa at multiple levels, including cryptic species and within-species groups (Blair et al., 2024). A key limitation, however, is the “black box” nature of AI, where the features used to reach a decision are often opaque (Von Eschenbach, 2021). To address this, CV explainability tools in the form of heatmap overlays can be applied (Cheng et al., 2025, Aysel et al., 2023). These highlight image regions contributing most strongly to classification decisions and previous work

has demonstrated that such methods can successfully identify morphologically defining traits from 2D images of specimens from closely related species (Hollister et al., 2023). However, there is currently no work investigating whether these features highlighted by CV explainability align with the perspective and interpretation of a human classifying the same specimens visually using morphology.

Here, we investigate whether the attention of an image classification pipeline aligns with the morphological features used by humans. Specifically, we examine closely related butterfly species using Grad-CAM and saliency maps and compare the highlighted regions to diagnostic features reported in key identification literature and to descriptions generated by a large language model. By combining heatmap-based visualisations with established identification literature and model-generated descriptions, we can directly assess whether the features emphasised by AI classifiers correspond to those traditionally used in morphological assessment. The following results present eight species, across four pairs, to examine where classifier attention aligns with, or diverges from, human-defined diagnostic characters.

5.3 Methodology

5.3.1 Species Selection

Four pairs of closely related congeneric British butterfly species were selected for analysis: *Aricia agestis* (Denis and Schiffermüller, 1775*), Aricia artaxerxes* (Fabricius, 1793*), Boloria Euphrosyne* (Linnaeus, 1758)*, Boloria selene* (Denis and Schiffermüller, 1775)*, Colias croceus* (Geoffroy, 1785)*, Colias hyale* (Linnaeus, 1758)*, Pieris brassicae* (Linnaeus, 1758) and *Pieris rapae* (Linnaeus, 1758).

5.3.2 Dataset and Model Construction

The dataset, referred to as the “iCollection”, originally comprised ~120,000 images across 59 species, including the eight focal species (Paterson et al., 2016). Previous work identified potentially mislabelled specimens within this dataset (Hollister et al., 2025). After removing these, the dataset was reduced to ~105,000 images across the same 59 species. For model development, 300 images per species were randomly sampled, of which 250 were used for training and 50 for validation, resulting in 17,700 images in total. The remaining images (~87,300) were used for testing and model evaluation. This random sampling and training process was

repeated 10 times with different images to ensure that the results were not a one off and to show the integrity of the workflow.

Image classification models were implemented in Python using the TensorFlow DL framework. A modified VGG16 architecture (Simonyan and Zisserman, 2014) pre-trained on ImageNet weights (Deng et al., 2009) was employed, with custom top layers added. Hyperparameters were tuned during preliminary trials and subsequently fixed throughout. Training followed a two-stage procedure: initially, all VGG16 layers were frozen and only the top layers were trained for five epochs; Then, all but the lowest eight VGG16 layers were unfrozen and training continued until early stopping criteria were met, ensuring that only the best weights were retained. Additionally, pairwise classifiers were then trained using the pruned dataset for each of the four species pairs, with 10 iterations each.

5.3.3 Feature Analysis

Morphological descriptions were collated from three commonly used identification guides: The Butterflies of Britain and Ireland (Thomas, 2020), Butterflies of Britain and Europe: A Photographic Guide (Haahtela, 2019), and the Collins Butterfly Guide (Tolman, 2008). These are hereafter referred to as BBI, BBE, and CBG respectively. Each contains transcriptions of morphological based-identification methods for all 59 species using standard naming practises of features (Fig. 25).

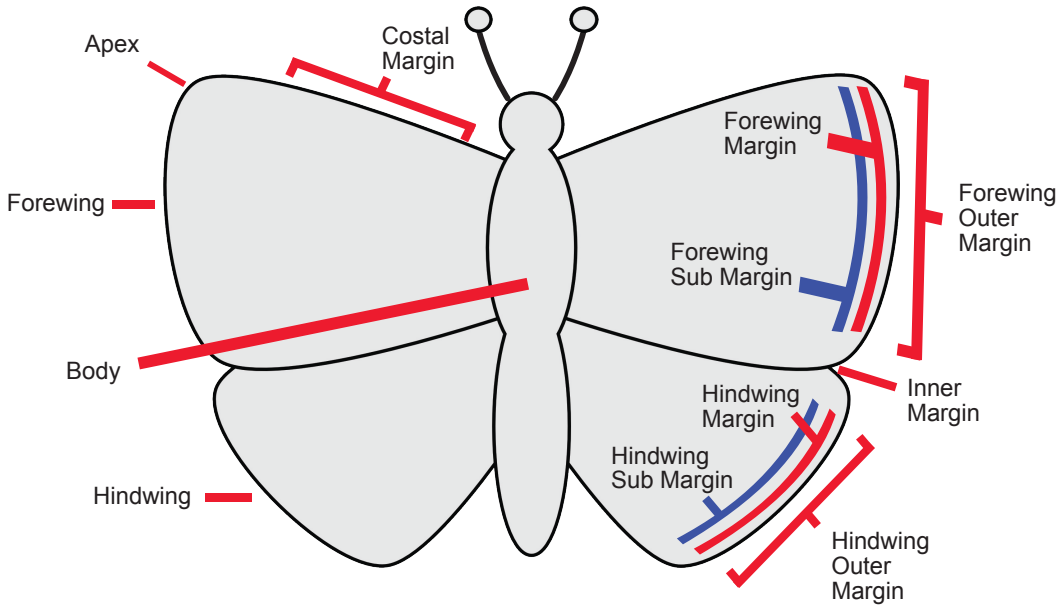


Figure 25. Diagram showing names of specific features and regions used within for transcriptions.

In addition, a large language model (LLM) was incorporated into this study as an independent source of morphological information for comparison with the heatmaps. LLMs are trained on vast datasets that typically include websites, books, and scientific papers (Liu et al., 2024b). Although the exact sources used to train most models are not disclosed, it is likely that they embed knowledge from the same body of literature consulted in this research. To date, however, there have been no published studies that explicitly use LLMs in this way, and their inclusion here therefore represents a novel approach. Many of the core identification resources for British Lepidoptera consist of historic monographs, field guides, and long-established natural history texts that are not subject to modern paywalls, increasing the likelihood that their content is represented within large language model training corpora.

For this study the LLM Claude (Sonnet v4) (Anthropic, 2025) was selected to provide an additional source of information for comparison with the heatmaps. Claude was chosen due to its status as a top-tier LLM with demonstrated high performance on complex reasoning tasks (Jiang et al., 2025). Claude was prompted with the following query for each pair of species:

"Can you describe the morphological differences between Species A and Species B with a focus on the upper side of their wings, please?"

To assess the features utilised by the classifiers, two explainability techniques were employed: Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2020) and

saliency maps (Simonyan et al., 2013). Grad-Cam uses the gradients flowing into the final convolutional layers of the network to weight the importance of feature maps. These weighted maps are then combined to produce a coarse localisation heatmap over the input image. Grad-CAM therefore highlights broader regions that contribute most strongly to the classification decision, rather than individual pixels. Saliency maps provide a pixel-level measure of importance by calculating the gradient of the classification score with respect to each input pixel. The resulting heatmap highlights the specific pixels where small changes would most strongly influence the classifier's output. Saliency maps therefore emphasise fine-scale image details and are well suited to identifying localised features.

The resulting heatmaps were visually interpreted to determine which image regions were highlighted, whether these highlighted areas corresponded to morphologically defining features upon the specimens within the images, and the extent to which they aligned with descriptions from the literature and Claude outputs. Example images and heatmaps for each species can be seen in figures 26 – 33 for the full classifier model (trained on 59 species) and the paired classifier models (trained on each species pair).

5.4 Results

5.4.1 Model performances

Models trained on the unpruned dataset achieved a mean F1-score of 0.926 after 10 iterations. After pruning potentially mislabelled specimens, mean performance increased to 0.996 across 10 iterations. Pruned datasets were therefore used in all subsequent analyses to preserve the integrity of species-specific features. Using the pruned dataset, the following mean F1-scores were obtained for the pairwise groups (Table 3).

Pairwise groups and their respective mean model performance after 10 iterations	
Pairwise Group	**Mean F1-score**
A. agestis vs. *A. artaxerxes*	0.954
B. euphrosyne vs. *B. selene*	0.994
C. croceus vs. *C. hyale*	0.993
P. brassicae vs. *P. rapae*	0.999

Table 3. Pairwise groups and their respective mean model performance after 10 iterations.

5.4.2 *Aricia agestis* versus *Aricia Artaxerxes*

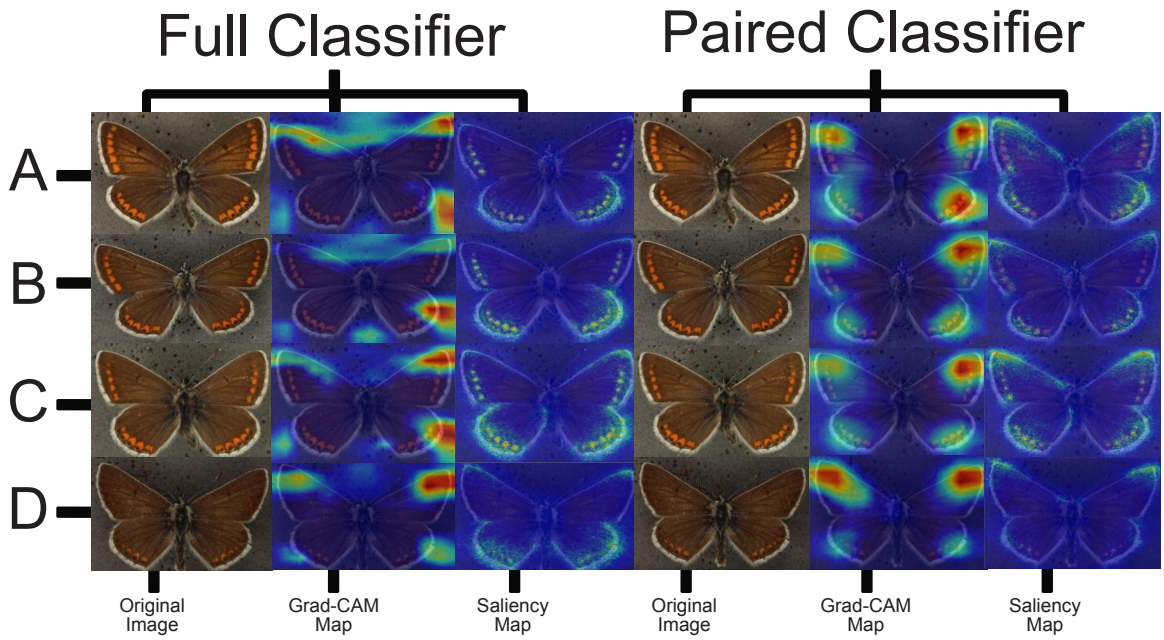


Figure 26. Four example specimens (A-D) of Aricia agestis with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

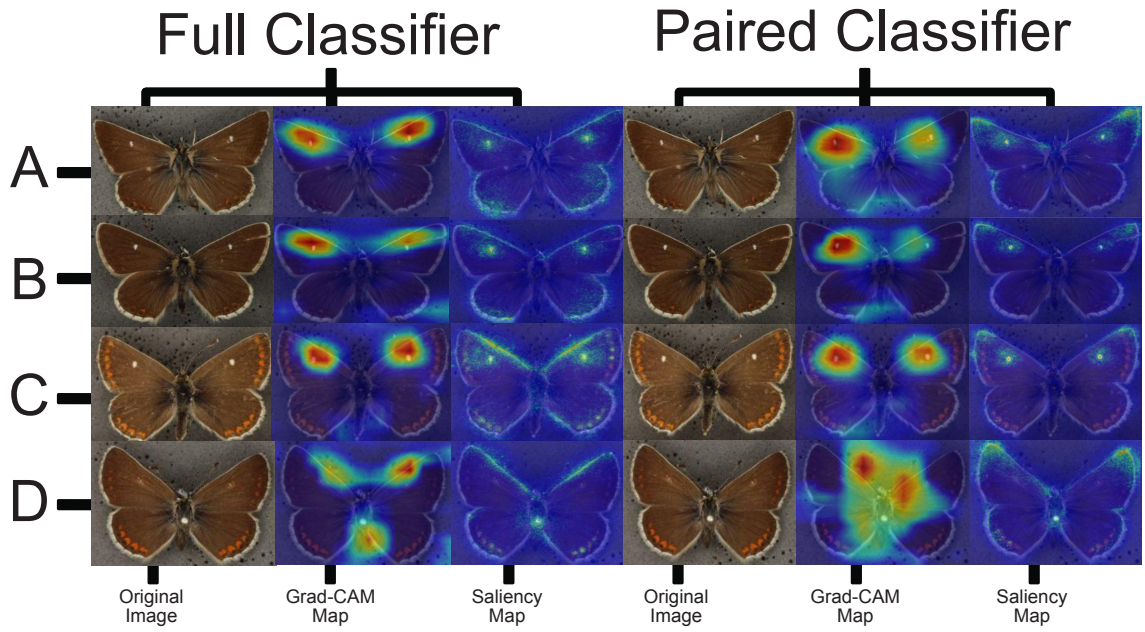


Figure 27. Four example specimens (A-D) of species Aricia Artaxerxes with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

For *A. agestis* and *A. artaxerxes*, both the literature and Claude emphasise orange lunules (CBG refers to them as "submarginal spots"), but the detail differs. In *A. agestis*, BBI notes that males may show less defined lunules than females. Saliency maps from both full and paired classifiers highlight the lunules, although this is less clear in specimen 26D (Fig. 26), where the markings are faint. Paired classifier Grad-CAM maps also attend to the lunules, whereas full classifier Grad-CAM maps do not. In *A. artaxerxes*, by contrast, the literature describes lunules as reduced or absent on the hindwing, and Claude notes that they are less defined overall than in *A. agestis*. Lunules are visible in specimens 27C and 27D (Fig. 27), but they are not highlighted in any Grad-CAM maps. Saliency maps from the full classifier show some attention, while paired classifier saliency maps avoid them entirely. The forewing discoidal spot further separates the species. In *A. agestis*, neither the literature nor the heatmaps emphasise this feature, although faint dark spots are visible on the forewings (Fig. 26). In *A. artaxerxes*, BBI and CBG describe the discoidal spot as white, and all heatmaps highlight it in specimens 27A–C (Fig. 27). In specimen 27D (Fig. 27), where the marking is faint, the models instead attend to a nearby pinhead artefact of similar size and colour. Wing shape is also described differently. Claude notes *A. agestis* as having "more rounded" wings and *A. artaxerxes* as "more angular". In *A. agestis*, both classifiers show attention on the forewing apex and the outer hindwing margins. In *A. artaxerxes*, the two classifiers differ: the full classifier shows attention on the costal margin, while the paired classifier highlights the perimeter of the forewing apex.

5.4.3 *Boloria euphrosyne* versus *Boloria selene*

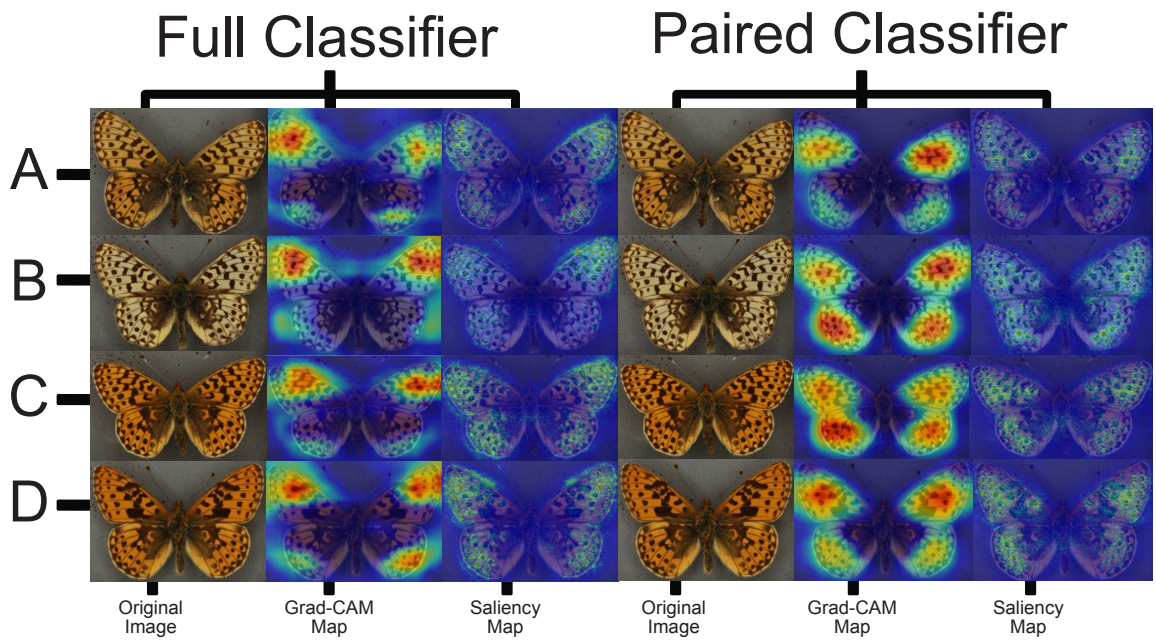


Figure 28. Four example specimens (A-D) of species Boloria euphrosyne with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

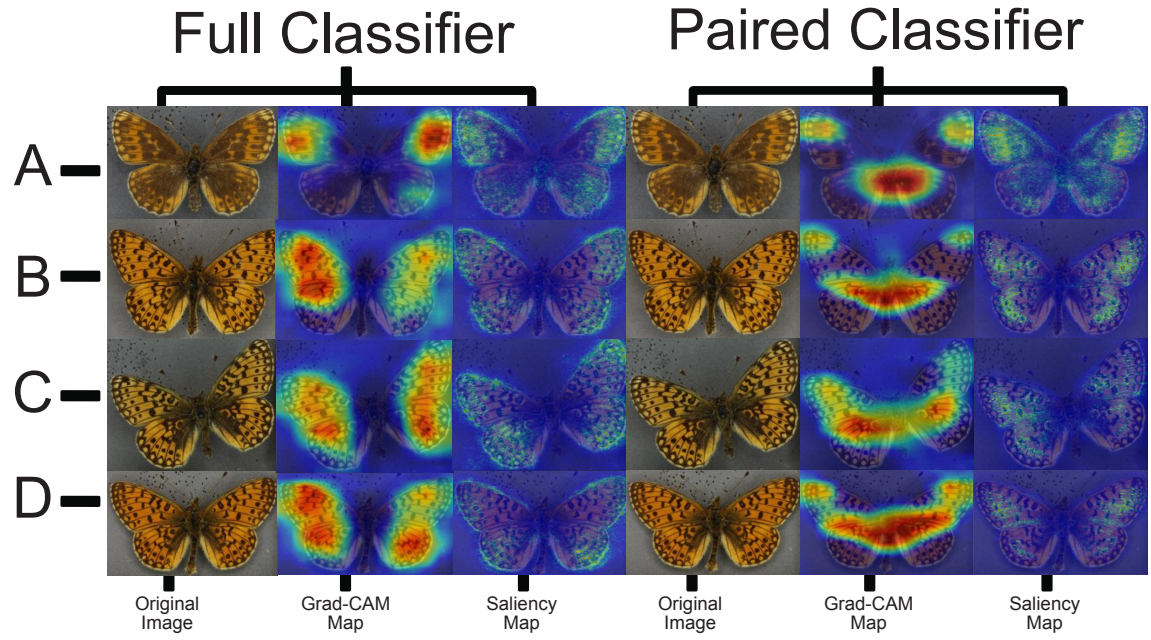


Figure 29. Four example specimens (A-D) of species Boloria selene with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

For *B. euphrosyne* and *B. selene*, the literature provides limited information on the upperside. In *B. euphrosyne*, CBG states that diagnostic characters are on the underside. In *B. selene*, all three

sources (BBE, BBI, CBG) emphasise the underside, with no mention of upperside characters. Claude gives descritipons for both species. In *B. euphrosyne*, BBI describes uniform upperside markings with yellowish marginal spots in females and darker wing bases. BBE mentions the presence of black triangular spots in the margins. BBE and Claude add that both sexes are similar. Claude describes *B. euphrosyne* as brighter, with a tawny orange ground colour and sharper black markings, and *B. selene* as duller, with blurred and less contrasted markings. However, the heatmaps distinguish the two species. In *B. euphrosyne*, Grad-CAM maps from both classifiers show attention on the submarginal areas of the hindwings and forewings. This is reflected in both classifiers; however, the attention is more spread out in the paired classifier heatmaps. For the *B. euphrosyne* specimens, there is also attention on the apices in both classifiers for the Grad-CAM heatmaps. In the full classifier, the attention additionally spreads across the forewings and hindwings, whereas in the paired classifier there is less attention in these areas. There is attention on the hindwing area in the paired classifier for *B. euphrosyne*; however, this area was not focused on in *B. selene*, despite being mentioned in the literature. For both species, the saliency maps show focus on the marginal areas, and while not mentioned in the literature for *B. selene*, this does match BBE's description of triangular outer marginal spots in *B. euphrosyne*. Claude mentions shape differences for both species. Both species present attention on the perimeter of the apex; however, *B. selene* shows further attention along the perimeter of the hindwing outer margin.

5.4.4 *Colias croceus* versus *Colias hyale*

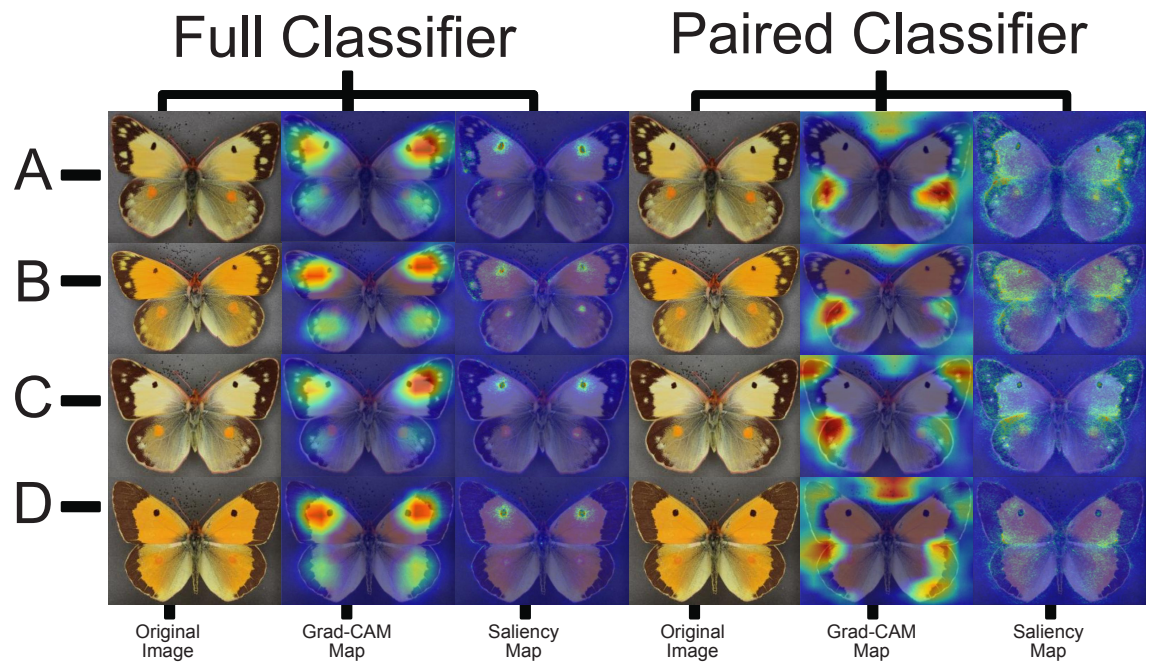


Figure 30. Four example specimens (A-D) of species Colias croceus with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

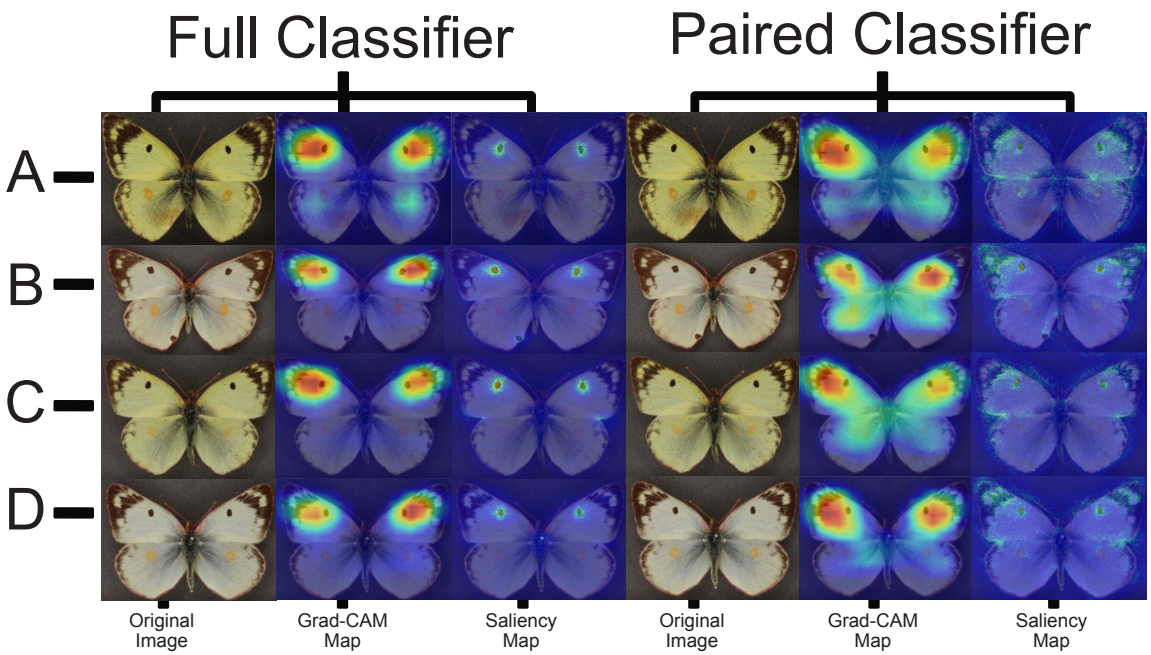


Figure 31. Four example specimens (A-D) of species Colias hyale with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

For *C. croceus* and *C. hyale*, both the literature and Claude emphasise the submarginal regions. In *C. croceus*, BBE, BBI, CBG, and Claude all describe the submarginal areas, noted as entirely black in males and black with irregular yellow spots in females. Almost all Grad-CAM heatmaps attend to these regions, except in specimens 30A and 30B of the paired classifier (Fig. 30). The saliency maps from the paired classifier focus consistently on these areas across all images, whereas the full classifier saliency maps do not. In *C. hyale*, BBE does not describe the upperside, while CBG and Claude both note the submarginal areas as dark. Grad-CAM maps show some attention to the forewing margins, and the paired classifier saliency maps also highlight these regions, whereas the full classifier saliency maps do not. Forewing spots also distinguish the two species. In *C. croceus*, Claude describes a black spot on the forewing, which is reflected in Grad-CAM and saliency maps from the full classifier but is absent in the paired classifier maps. In *C. hyale*, CBG and Claude note the forewing black discal spot, and all heatmaps show attention to this feature. BBI further describes the wing bases as slightly darker. Only the paired classifier Grad-CAM maps show some attention to these areas, while the full classifier and both saliency maps do not. In *C. croceus*, all heatmaps show some focus on the hindwing orange oval spot, with stronger emphasis in the paired classifier maps, although this feature is not mentioned in the literature or by Claude. In *C. hyale*, the hindwing margins are greatly reduced or almost absent, consistent with both the literature descriptions and the heatmaps. Wing shape is mentioned by Claude for both species. In *C. croceus*, this is not reflected in the full classifier maps, although both Grad-CAM and saliency maps from the paired classifier show attention on the inner margins. In *C. hyale*, BBI and Claude also describe wing shape. The paired classifier saliency maps show attention on the inner margin, and the full classifier saliency map shows a small amount of attention in specimen 31C (Fig. 31). Neither Grad-CAM map shows consistent attention to these features.

5.4.5 *Pieris brassicae* versus *Pieris rapae*

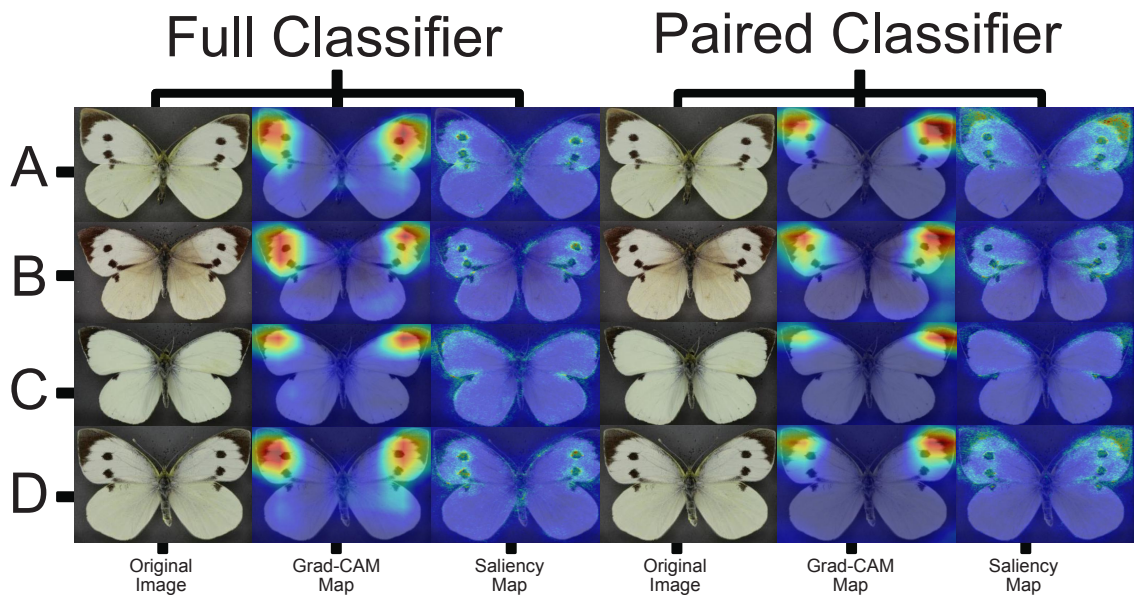


Figure 32. Four example specimens (A-D) of species Pieris brassicae with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

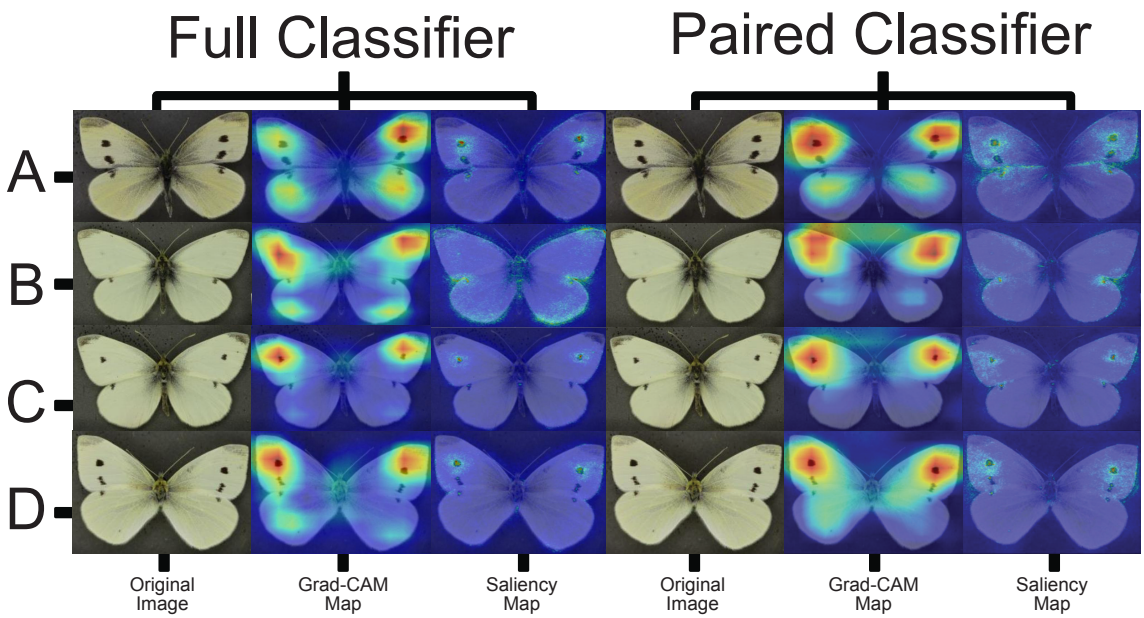


Figure 33. Four example specimens (A-D) of species Pieris rapae with Grad-CAM and Saliency map images from both the full classification model and the paired classification model.

For *P. brassicae* and *P. rapae*, the literature, Claude, and the heatmaps all highlight diagnostic differences on the forewings. In *P. brassicae*, BBE, CBG, and Claude all describe dark forewing tips. Both Grad-CAM maps highlight these regions, although only the paired classifier saliency maps show attention to them. In *P. rapae*, BBE, BBI, and Claude note that the black forewing

margins are reduced and less intense. These margins are visible in the specimens, and while both Grad-CAM maps show some attention to them, neither saliency map does. In *P. brassicae*, BBE and Claude describe two black spots on the female forewing, and both Grad-CAM and saliency maps from the full and paired classifiers highlight these spots when present. In *P. rapae*, BBE, BBI, and Claude state that females have two spots, whereas males have one. All heatmaps consistently focus on the single spot present in both sexes. For the second spot in females, neither saliency map shows attention, while both Grad-CAM maps show only minor attention. Claude notes that *P. brassicae* has "more rounded" wings, and both saliency maps show attention to the costal margin, forewing outer margin, and inner margins. By contrast, Claude describes *P. rapae* as "narrower and more pointed" compared to *P. brassicae*. The heatmaps show no clear attention to wing shape or perimeter however both Grad-CAM maps display broad focus across the fore- and hindwings without emphasising specific patterns, while both saliency maps avoid these regions.

5.5 Discussion

This study examined whether CNN classifiers attend to the same morphological features that humans use when distinguishing closely related butterfly species. By comparing two CV explainability visualisation tools, Grad-CAM and saliency maps, with descriptions from three standard identification guides and a large language model, we evaluated the extent to which machine-derived attention aligns with human-based interpretation. Accordingly, the analyses presented here are intended as an exploratory assessment of the extent to which machine attention aligns with human-defined diagnostic characters and model-generated descriptions, rather than as an exhaustive or definitive account of species-level butterfly morphology.

Across all species examined, most attention highlighted by the heatmaps corresponded with features discussed in the literature. Saliency maps consistently emphasised diagnostic characters such as lunules, discoidal spots, and marginal markings. In contrast, Grad-CAM maps more often displayed broad, region-level attention across the forewings or hindwings, with less emphasis on discrete traits. This distinction reflects the underlying principles of the two methods: saliency maps quantify pixel-level sensitivities, whereas Grad-CAM reflects gradients at the feature-map level, producing coarser representations (Simonyan et al., 2013, Selvaraju et al., 2016, Selvaraju et al., 2020).

The comparison also revealed areas of divergence between model attention and human interpretation. In *C. crocues*, heatmaps consistently focused on hindwing orange oval spots,

despite these not being mentioned in the literature or Claude outputs. Again, in *P. rapae*, the Grad-CAM of both classifiers is showing a broad focus across the wings which is not mentioned in the literature or from Claude. We also see focus on novel artefacts, that resemble features. In *A. artaxerxes*, heatmaps highlighted a specimen pinhead that visually resembled a discoidal spot. Such cases suggest that CNNs may incorporate features overlooked or previously considered taxonomically irrelevant by humans, raising the possibility that explainability methods could uncover underappreciated characters or even new delineating features. At the same time, they highlight the need for caution, as models may also rely on spurious correlations introduced during training or respond to non-biological artefacts (Geirhos et al., 2020, Hollister et al., 2025).

A clear distinction also emerged between paired and full classifiers. Paired classifiers consistently produced more areas and pixels highlighted across both Grad-CAM and saliency maps. In contrast, full classifiers, which were required to discriminate among 59 categories, distributed their capacity across a broader set of features and decision boundaries. This appeared to result in fewer visible regions of attention, although diagnostic features were still highlighted. For example, in *P. brassicae* the full classifier saliency maps showed attention only on the hindwing spots, whereas the paired classifier highlighted both the hindwing spots and the forewing tips, with both traits referenced in the literature. Although the paired classifier heatmaps did not always yield sharper diagnostic focus, the results suggest that restricting the task to two categories enables the network to allocate additional internal resources to detecting multiple defining features (Del Moral et al., 2022, Huh et al., 2016). This implies a trade-off: classifiers trained on many species may be more effective at large-scale identification but reduce the opportunity to emphasise traits specific to closely related taxa. However, it should be noted that the full classifier model and the paired model classifiers resulted in almost equally high F1-scores meaning their ability to correctly predict specimens was on par.

In addition, comparisons with Claude outputs highlighted further nuances. While Claude often reiterated features described in the literature, it also emphasised traits such as wing shape that were not consistently reflected in the heatmaps. Large language models such as Claude are trained on vast quantities of data, including books, scientific journals, and other sources that may contain morphologically defining features relevant to the species examined in this study, but which were not present in the selected identification literature (Liu et al., 2024b). However, LLMs are also known to amplify their outputs, adding additional details or generating incorrect information in a process referred to as “hallucination” (Orgad et al., 2024; Ji et al., 2023). Some LLMs have the capacity to search the web when prompted and provide cited sources, which can lend greater transparency and credibility to their outputs (Shi et al., 2025; Liu et al., 2024a). For the purpose of this study, such functionality was deliberately not used. Instead, the LLM was restricted to a single, fixed prompt per species pair, without iterative refinement or external

retrieval, in order to provide a controlled comparison between the model's internalised knowledge, the identification literature, and the CNN-derived heatmaps. It should be noted that many contemporary "co-scientist" or agentic LLM systems rely on iterative prompting, retrieval-augmented generation, and human-in-the-loop feedback to progressively refine outputs (Boiko et al., 2023). The simplified use adopted here therefore reflects an intentionally conservative and exploratory application of LLMs, aimed at contextual comparison rather than authoritative morphological synthesis.

An important observation is that clear differences emerge in the heatmap attention patterns between paired species, with species-defining features highlighted even when specimens appear highly similar. For example, *A. agestis* and *A. artaxerxes* share many traits, including overall colouration, orange lunules, forewing spots, and similarly coloured margins. Despite this, the heatmaps revealed distinct focal regions: in *A. agestis* attention was concentrated on the hindwing margins, whereas in *A. artaxerxes* it was directed towards the costal margins. The models also differentiated spot patterns, showing little attention to the black spots in *A. agestis* but highlighting the white spots in *A. artaxerxes*. Furthermore, stronger attention was placed on the lunules of *A. agestis* compared with *A. artaxerxes*, even though these markings occur in both species. These distinctions, combined with the high F1-scores achieved by both paired and full classifiers, indicate that the models were able to capture and exploit subtle morphological differences sufficient to accurately delineate between the various pairs. We also see attention patterns of distinguishing features that vary between the species that also show sexual differences. For example, in *P. rapae* and in *P. brassicae* females possess two spots on their forewings. However, the males in *P. brassicae* do not contain this spot while the males in *P. rapae* can contain this spot. In *P. brassicae*, the attention of both heatmaps in both classifiers focuses on both forewing spots in females, however, in *P. rapae* the attention only focuses on one spot in females which is the same one that is present in some of the males. This therefore provides evidence that specific features can be both focused on and ignored when similar features are shared between morphologically similar specimens but also contain small but unique features.

Overall, our findings reinforce previous work showing that CNNs can recover biologically meaningful features from specimen images (Hollister et al., 2023) but extend this by systematically comparing machine attention with human reasoning. The strong correspondence between the heatmaps and diagnostic traits used by humans suggests that explainability methods can be valuable tools for validating image-based classifiers in biodiversity research. However, both Grad-CAM and saliency maps are constrained by their mathematical formulations, and neither fully captures global structural traits such as wing shape or the whole colour of the specimen. In addition, the analysis was restricted to four species pairs, and results

may not generalise across other taxa or imaging conditions. It should also be noted that this analysis only focused on the upper side of the wings whereas a lot of the literature specifically noted that defining features can be easily obtained from the undersides. Furthermore, this type of analysis did not take size into account, as all specimens were standardised to the same image size during pre-processing, and it has been shown that some specimens of butterfly, while visually similar in appearance, are measurably different in size (Hollister et al., 2025). However, it could be equally argued that despite only having access to the upper side, or being unable to differentiate size, the accuracy of the classifiers is extremely high (Table 3).

5.6 Conclusion

This study demonstrates that CNN classifiers attend to many of the same morphological features that humans use to distinguish closely related butterfly species. Saliency maps showed the strongest correspondence with field guide descriptions, while Grad-CAM maps revealed broader regional focus. Paired classifiers highlighted more diagnostic features than the full classifier, reflecting the trade-off between fine-grained discrimination and large-scale classification. While broad classifiers remain valuable for large-scale species identification (Hollister et al., 2025), pairwise or restricted classifiers can provide deeper insight into the traits underlying species boundaries. LLMs such as Claude also show potential as complementary sources of morphological descriptions, although their outputs require careful validation. Features such as web-based reference retrieval may help ensure reliability in future applications. Overall, explainability methods such as Grad-CAM and saliency maps can help bridge machine learning and human interpretation in taxonomy. Their integration into biodiversity workflows offers opportunities not only to validate model outputs, but also to identify cryptic traits (Pinho et al., 2023), refine datasets (Wan et al., 2024), and enhance large-scale specimen curation (Hollister et al., 2025).

Future work should extend this approach by applying it at scale to large museum collections and across a wider range of taxa, while also incorporating alternative explainability techniques and multimodal datasets and techniques (Bayoudh et al., 2022, Bayoudh, 2024). Such developments would not only further validate the methodology but also support the detection of misidentified specimens (Hollister et al., 2025), the recognition of cryptic taxa, the discovery of overlooked morphological traits, and ultimately strengthen its value as a tool for both CV research and biological discovery.

Chapter 6 Conclusion

The overarching aim of this thesis was to investigate how CV and XAI can be applied to morphological interpretation in biodiversity science. Traditional taxonomy, long reliant on expert-led morphological observation, faces persistent difficulties that arise both from biology and from practice. Morphological difficulties include cryptic similarity and phenotypic plasticity, while practical constraints such as mismanagement and overburdened workloads can lead to misidentifications (Bickford et al., 2007; Hebert et al., 2004; Meyer and Paulay, 2005). At the same time, molecular approaches have uncovered hidden diversity but often leave unanswered the question of which visible traits distinguish lineages (Dayrat, 2005; Padial et al., 2010). This creates a critical gap between genotype and phenotype that hampers integrative biodiversity research. The central hypothesis tested here was that CV models, when coupled with XAI methods, could not only achieve high classification accuracy but also visualise morphological signals in ways that are transparent, interpretable, and scalable. By combining DL with heatmap-based interpretability methods (Simonyan et al., 2013; Selvaraju et al., 2020), this work aimed to assess whether AI-based approaches could complement, or in some cases surpass, traditional identification workflows, while still remaining interpretable to human experts.

6.1 Review of Key Findings

In Chapter 2, CNNs were trained on dorsal and ventral images of limpets from the Baja California peninsula, achieving accuracies (97.9%) marginally higher than expert human identifications (97.5%), and doing so at speeds unattainable for a human (Hollister et al., 2023). Heatmap visualisations confirmed that model attention corresponded to diagnostic features recognised by ecologists, providing reassurance that classifiers could focus on biologically meaningful regions.

In Chapter 3, CV was applied to digitised butterfly collections, where a workflow for flagging mislabelled specimens was developed (Hollister et al., 2025). This directly addressed a critical challenge in NHCs, where errors can propagate into downstream biodiversity analyses (Goodwin et al., 2015) and showed that automated screening can provide scalable solutions for dataset integrity while also laying the groundwork for how the different scientific disciplines of visual taxonomy, genetics, and CV can work in harmony.

Chapter 4 examined genetically distinct but morphologically similar limpet clades. Models achieved high accuracy in distinguishing clades, and saliency maps highlighted consistent shell regions potentially linked to evolutionary adaptations (e.g., the keyhole shape in *F. volcano)*; This created an opportunity to extract significant empirical evidence and showed that even when phenotypic differentiation is subtle and invisible to human observers, it can nonetheless be recovered and visualised through AI, offering a route to help bridge the gaps between genotype and phenotype (Pfenninger and Schwenk, 2007; Padial et al., 2010).

Finally, Chapter 5 compared the attention from heatmaps with traits defined by experts and described in the taxonomic literature for closely related British butterfly species, as well as with the internal knowledge of a large language model (LLM). Areas of overlap were evident and abundant, with features highlighted in common across the literature, the LLM, and the classifier. However, differences also emerged: heatmaps occasionally highlighted regions not mentioned in the literature, which could represent overlooked diagnostic traits but could equally reflect attention to novel artefacts in the images. Despite these differences, high F1 scores were achieved, demonstrating that models can accurately distinguish between closely and morphologically similar species while capturing discriminative features consistent with expert-derived traits and highlighting new areas of potential interest (Wäldchen et al., 2022; Behrens et al., 2023). This underscores both the promise and the challenge of XAI where models may detect meaningful signals not emphasised in human taxonomy, raising questions about how such traits might be incorporated into formal diagnoses.

6.2 Synthesis of Chapters

When considered together, the data chapters show how CV and XAI can contribute to taxonomy and biodiversity science by combining accuracy, scalability, and interpretability in ways that traditional methods cannot. Across both limpets and butterflies, models consistently matched or exceeded expert performance (Hollister et al., 2023, 2025), doing so with vastly greater efficiency and at a scale that far outpaces the practical limits of human identification. This highlights a fundamental difference between humans and machines: whereas expert performance is constrained by fatigue, workload, and time pressure, well-designed CV models can operate indefinitely at peak accuracy and without subjective bias. For example, in Chapter 2 the expert involved noted that their misidentifications were most likely linked to fatigue during extended periods of classification (Hollister et al., 2023). Similarly, in Chapter 3, several of the mislabels flagged by the CV pipeline were traced to human error, spanning multiple generations of Lepidoptera curation and extending into modern departments, including digitisation and data-

hosting teams (Hollister et al., 2025). This demonstrates that, even when many experts and departments are involved, human-based errors remain inevitable, and therefore an independent, systematic system for specimen classification and validation offers the potential to become a valuable safeguard if utilised correctly.

When incorporating heatmaps into the analytical pipeline (Simonyan et al., 2013; Selvaraju et al., 2020), model predictions went from being blindly reported as being highly accurate to becoming linkable to visible features from model outputs, thereby acting as a form of model validation. This demonstrates that DL does not have to remain a "black box" (Rudin, 2019; Von Eschenbach, 2021) and that, with the right tools, the visual features underpinning predictions can be made transparent and available for expert scrutiny (Aysel et al., 2025, Doshi-Velez and Kim, 2017). In the case of limpets in Chapters 2 and 4, this provided reassurance that models were attending to traits already considered taxonomically meaningful, such as the keyhole, ridges, and perimeter shape (Simison and Lindberg, 2003; Crummett and Eernisse, 2007; Hollister et al., 2023). Similarly, in chapter 3 and 5, attention maps highlighted traits on butterflies long recognised in taxonomy, such as wing spots and lunules (Tolman, 2008, Haahtela, 2019 Thomas, 2020,). Having a toolset that can explicitly define the features distinguishing specimens has important implications for taxonomy: it can be used to verify identifications alongside human expertise, to highlight diagnostic features in newly described species, or has the potential to support the creation of identification keys for taxa lacking dedicated experts (Padial et al., 2010).

In Chapters 2 and 5, clear differences in the allocation of heatmap attention between closely related and visually similar species and genus pairs were presented and discussed. For instance, figure 5 in chapter 2 shows Grad-CAM images for *F. rubropicta* and *F. volcano*. Even though both are similar in appearance, a clear difference in the allocation of features via the heatmaps is evident with the area around the keyhole being highlight for *F. rubropicta* whereas the keyhole is highlighted for *F. volcano*. This was suggested to be due to having more pronounced ribbing *F. rubropicta,* although this was not further interrogated (Hollister et al., 2023). Again, in chapter 5, in the *P. brassicae and P. rapae* comparison there are differences between the two species heatmap allocations. Both species females can have two spots on their forewing, however, the males in *P. rapae* can contain one spot while the males in *P. brassicae* contain none. This appears to be distinguishable as the heatmaps for *P. rapae* highlight only the spot present on both sexes but highlight both for the females in *P. brassicae*. Furthermore, Chapter 2 was able to assign statistical significance to these differences using heatmap intensity values obtained by aggregating the values of each pixel when scored based on brightness (Fig. 8) (Hollister et al., 2023). By contrast, in Chapter 4, when observing clades within the same species, heatmaps highlighted the same features across clades (for example, the keyhole in *F. volcano*), suggesting no difference. Yet, once further interrogation of these highlighted features was carried out using

shape analysis, clear and statistically significant differences were revealed. This means that XAI can both validate existing taxonomic characters and guide more detailed morphological analyses, helping to uncover subtle forms of divergence that may otherwise be overlooked.

In addition to qualitative inspection of attention maps, Chapter 2 also explored a simple quantitative summary of model attention using heatmap intensity values, obtained by aggregating pixel-level brightness across each heatmap and comparing distributions between classes. This provided an initial, statistically tractable way to test whether model attention differed systematically between taxa, and offered a potential route for using attention summaries as a secondary diagnostic signal (for example, to flag predictions that fall outside typical intensity ranges for a given class). However, this approach was treated as exploratory and was not pursued as a primary analytical thread in subsequent chapters. Attention intensity is sensitive to choices in normalisation, image framing, and background structure, and may conflate changes in localisation with changes in overall map magnitude, limiting comparability across datasets, architectures, and imaging regimes. Later chapters therefore prioritised the spatial localisation of attention and, where necessary, direct extraction of interpretable morphology from highlighted regions, rather than relying on global heatmap-intensity summaries.

At the same time, Chapters 3 and 5 revealed instances where models attended to novel artefacts rather than diagnostic features. In Chapter 3, for example, early testing showed that models sometimes focused on specimen labels instead of the specimens themselves (Hollister et al., 2025). This artefactual attention was subsequently corrected through preprocessing steps where images were cropped to remove these labels, illustrating the value of explainable methods for identifying and mitigating such issues. In chapter 5, attention was shown to focus on a pin head that closely resembled the marking of the species of butterfly it was assessing. These findings demonstrate the utility of XAI not only for validating model predictions but also how it can be used for improving dataset quality by revealing hidden artefacts. Data quality is one of the key determinants of robust AI models (Murdoch et al., 2019) and having a tool that can flag attention issues before full-scale training could not only lead to better model performance but also reduce costs by avoiding wasted computation on flawed data.

6.3 Limitations

Despite these advances, several limitations must be acknowledged. All analyses relied on curated digital images, whether of museum specimens or field-collected material, which introduces biases in terms of image quality, consistency, and the accuracy of existing

identifications (Wan et al., 2024). Historical mislabelling in collections is well documented (Goodwin et al., 2015), and even with careful verification, such errors can propagate into training datasets and influence model performance (Rädsch et al., 2023). In Chapter 3 we also observed that the iCollections dataset was not only constructed by curators, which could have introduced errors, but also by digitisers and staff uploading images, all of whom contributed mistakes that the pipeline was able to highlight. Such issues could potentially influence the results presented here but also affect the work of others who download and reuse these digital datasets. However, in Chapter 5 we also see that by removing possible mislabels within datasets, the potential accuracy of models can be substantially increased, demonstrating how careful filtering and error correction can markedly improve performance.

Interpretability itself also carries significant caveats. Although heatmaps provide visualisations of model attention (Simonyan et al., 2013; Selvaraju et al., 2017), they do not imply that the AI 'understands' what it is observing, input is only numbers and patterns. For example, Adebayo et al. (2018) showed that some saliency methods are insensitive to model parameters or training data, producing identical explanations even when the model was randomised. Similarly, Szczepankiewicz et al. (2023) found that method reliability varies strongly with dataset and architecture. These methods are therefore best treated as heuristic guides that require corroboration from taxonomic expertise (Murdoch et al., 2019; Pichler and Hartig, 2023), rather than as definitive explanations of model reasoning. They provide useful visual cues, but they can also mislead if not properly diagnosed, particularly in cases where external morphology is subtle, varied, or where imaging artefacts dominate.

Equally, while limpets and butterflies provided tractable case studies due to their availability of specimens and existing taxonomic frameworks, it remains uncertain how well these approaches will generalise to taxa with less distinctive morphology or with lower sample sizes (Trail, 2021). There may be fundamental limits to what CV can achieve: some groups may simply exhibit too little external morphological divergence for image-based models to consistently separate them, regardless of dataset size or algorithmic refinement. In such cases, even the best CV methods may plateau in performance. For these taxa, molecular or anatomical data will remain essential for reliable delimitation, with CV serving primarily as a complementary tool for flagging candidate divergences rather than providing definitive identifications.

The CV models and methods used within this thesis do not always transfer easily to new specimens or taxa. The models developed in all chapters were specific to the specimens they were trained on. As soon as a new species or group is introduced, the existing model will misclassify it into one of the classes it already knows. Each time new taxa are added, retraining would be required, rendering earlier models partially redundant. There are also practical limits on

how many classes a single CV model can robustly incorporate. Studies show that as the number of categories increases, classification accuracy declines unless per-class sample sizes are very large (Luo et al., 2019). Likewise, as the number of classes grows, networks exhibit increasing confusion between similar categories, reflecting the difficulty of learning complex class hierarchies (Alsallakh et al., 2017). A model encompassing every species in the world is therefore unachievable; instead, many smaller, taxon- or context-specific models will be required, with knock-on effects for dataset management, storage, computational cost, and long-term maintenance (Barbedo, 2018; Blair et al., 2024).

While the high accuracies obtained in this thesis using 2D images are encouraging, it is important to recognise that 2D views inherently limit the information available. For example, in the butterfly studies only the upper side was visible; features on the underside, wing edges, or lateral aspects were obscured. In Chapter 5, the taxonomic literature highlighted precisely these areas as containing key diagnostic information, with some sources even stating that underside characters alone could be used to distinguish certain species or groups. Moreover, when handling a specimen physically, a taxonomist can examine minute traits from all angles, detect tactile qualities such as texture or rigidity, and manipulate parts to view hidden structures, none of which are accessible via static 2D images, which again emphasises the need for a human in the loop workflow. By contrast, systems designed to generate 3D models (Nguyen et al., 2017) show that capturing specimens from multiple angles reveals additional geometric and structural information that improves resolution of diagnostic traits. Nonetheless, the high accuracies achieved with 2D images in this thesis, particularly when specimens had already been physically curated and prepared, suggest that for many taxa such limited views may still be sufficient for classification tasks, albeit with important caveats.

Finally, practical challenges remain around computational resources and reproducibility. Training DL models is resource-intensive, which may limit uptake in institutions without access to high-performance computing (Ahmed et al., 2024). Biodiversity studies also vary widely in how transparently they report data partitioning, model training, and code availability (Lipton, 2018; Zhou et al., 2021). At present, it remains unresolved whether attention visualisations consistently map to stable morphological traits across taxa, or whether they sometimes highlight artefacts of imaging. Similarly, while classifiers can recover subtle morphological variation that may be linked to genetic divergence, it is unclear whether such differences are universally interpretable within existing taxonomic frameworks. These knowledge gaps highlight the need for systematic cross-taxon evaluations of explainable models, as well as rigorous tests of whether AI-derived traits can be formalised into diagnoses or linked directly to functional or ecological significance.

6.4 Future Directions

First, there is a clear opportunity to scale the explainable CV workflow across other specimen types now being digitised at pace by major repositories; pinned insects, herbarium sheets, slide-mounted preparations, and microfossils, and particularly those where the majority of diagnostic information is captured in two dimensions (Blagoderov et al., 2012; Nelson and Ellis, 2019). Chapters 2, 4 and 5 have demonstrated that interpretable heatmaps can focus on biologically meaningful characters (Hollister et al., 2023), and in analogous 2D contexts the same pipeline should transfer with only minor adjustments. As museums continue to release large, standardised image datasets, the throughput and comparability required for multi-taxon deployment are increasingly available, and this thesis shows that such deployments can return actionable curatorial and taxonomic value at scale (Hollister et al., 2025).

A further avenue lies in extending explainable CV beyond single-plane 2D imaging to encompass whole specimens through multi-angle or 3D reconstructions. Current pipelines rely on dorsal or frontal views, which inevitably obscure key characters located on ventral surfaces, lateral aspects, or internal structures, as highlighted in Chapter 5 where underside wing features were central to taxonomic diagnoses. In contrast, 3D imaging technologies such as photogrammetry, structured light scanning, or micro-CT allow specimens to be rotated, examined from all directions, and even measured volumetrically, recovering traits inaccessible to flat images (Nguyen et al., 2017; Falkingham, 2012). Recent work in natural history digitisation shows that multi-view imaging can improve trait resolution and model robustness, particularly for groups where shape and curvature carry diagnostic value (Younis et al., 2020; Lürig et al., 2021). For XAI, 3D data would make it possible to localise attention across entire surfaces, link highlighted regions directly to geometric descriptors, and explore new dimensions of morphological variation. Integrating these approaches would not only improve classification accuracy but also align computational pipelines more closely with the holistic way taxonomists interact with specimens, inspecting from all angles, manipulating features, and weighing multiple traits simultaneously.

From a technical perspective, the next step is to strengthen the link between attention visualisations and quantitative morphology. Saliency and Grad-CAM provide intuitive ‘where’ signals (Simonyan et al., 2013; Selvaraju et al., 2020) but coupling them with segmentation-based extraction of shapes or subregions, as trialled in Chapter 4 (keyholes in *F. volcano* clades), would convert qualitative heatmaps into measurable traits that can be compared among individuals, populations, and clades. This would enable formal tests of whether highlighted regions align with established diagnostic structures or reveal new, consistent characters, moving explainability

from post-hoc illustration towards a reproducible framework. In parallel, dataset-refinement practices such as active curation loops that can highlight mislabels should be utilised as standard practice within museums and NHCs to ensure that models become both more accurate and more trustworthy, particularly for legacy datasets (Goodwin et al., 2015).

Another priority is integrative biology: using interpretable CV to bridge genotype and phenotype. The clade-based analyses here show that networks can detect subtle, spatially consistent divergence that aligns with independent genetic structure (Pfenninger and Schwenk, 2007; Padial et al., 2010). A natural extension is to pair heatmap-guided traits with targeted genomic sampling to test whether the highlighted regions map to loci involved in development or to environmental gradients implicated in local adaptation (Van Belleghem et al., 2021). Equally, future work must recognise that in some groups external divergence will remain minimal: for these taxa, hybrid pipelines that combine interpretable CV with molecular or anatomical data will remain essential for robust delimitation.

Finally, usability and reproducibility should be primary goals. Purpose-built interfaces that present predictions, uncertainty, and heatmaps alongside voucher metadata will make verification faster and more transparent, helping curators without overwhelming staff time (Hollister et al., 2025; Von Eschenbach, 2021). The field would benefit from lightweight, shareable reporting standards that cover train/validation/test partitions, augmentation regimes, XAI settings, uncertainty calibration, and code availability, making results reproducible across institutions with differing resources (Lipton, 2018; Zhou et al., 2021). Together, these directions aim to deliver interpretable CV that is not only accurate in the lab but a dependable and genuinely useful tool within the day-to-day work of NHCs.

6.5 Broader Implications

The potential applications extend beyond taxonomy and collections. As biodiversity monitoring intensifies under global change, the capacity to extract and validate morphological signals at scale could provide indicators of population level differences in morphology, which would be particularly important for populations that are under ecological stress or are genetically unique. Interpretable CV pipelines therefore hold promise not only for systematics but also for conservation planning, environmental assessment, and the preservation of biodiversity baselines in a rapidly changing world. At the same time, explainability addresses a central socio-technical concern in biodiversity informatics: trust. By making explicit which pixels and areas drive a decision, heatmap-guided workflows offer a common language for collaboration between

modellers and domain experts, turning model outputs into hypotheses about morphology rather than inscrutable labels (Doshi-Velez and Kim, 2017; Rudin, 2019). As these methods propagate through museums and research groups, transparent reporting and dataset-refinement practices will be essential to ensure that speed does not come at the expense of rigour (Murdoch et al., 2019; Pichler and Hartig, 2023). In combination, these developments point towards a future in which expert knowledge and machine inference reinforce one another: curators and taxonomists define the questions and make final decisions on model outputs, while models deliver scale, consistency, and highlight candidate morphological features (Wäldchen et al., 2022; Behrens et al., 2023).

At a time when biodiversity is declining at unprecedented rates and the demand for reliable, large-scale data is greater than ever, the ability to combine CV with XAI offers a timely contribution. NHCs, once considered static archives, are being transformed into dynamic research infrastructures through digitisation (Blagoderov et al., 2012; Nelson and Ellis, 2019). Yet these vast image datasets require methods that can both scale and explain their outputs if they are to inform ecological baselines, conservation priorities, and evolutionary research. The work presented here demonstrates that interpretable CV can help meet this challenge: by accelerating identifications, improving data quality, and revealing previously hidden morphological signals, while contributing to the broader goal of sustaining biodiversity in an era of rapid global change (Bickford et al., 2007).

6.6 Concluding Remarks

This thesis has shown that CV and XAI can move beyond proof-of-concept demonstrations and become practical tools for taxonomy, biodiversity informatics, and NHCs. Across multiple taxa and contexts, DL models were able to classify specimens with accuracy that matched or exceeded expert performance (Hollister et al., 2023), while attention visualisations confirmed that these predictions were grounded in morphological features rather than opaque artefacts (Simonyan et al., 2013; Selvaraju et al., 2020). By applying interpretable pipelines to mislabelling detection, cryptic species, and human–machine comparison, this work has demonstrated that AI can not only help accelerate large-scale identifications but also surface new, reproducible insights into the morphology of organisms.

As digitisation expands into the tens of millions of specimens worldwide (Blagoderov et al., 2012; Nelson and Ellis, 2019), the bottleneck is no longer data generation but data interpretation. Automated pipelines that can audit, annotate, and explain images will be essential for

maximising the value of these resources. At the same time, integrative biology increasingly depends on bridging the gap between molecular datasets and observable form. By revealing subtle, spatially consistent morphological variation that aligns with genetic structure (Pfenninger and Schwenk, 2007; Padial et al., 2010), XAI provides a route towards uniting phenotype with genotype at a scale that is otherwise impossible. Yet the broader lesson is that speed alone is not enough. Trust, transparency, and reproducibility are critical for adoption in biodiversity science. The work here reinforces the argument that AI must be interpretable if it is to be integrated into ecological and taxonomic workflows (Doshi-Velez and Kim, 2017; Rudin, 2019). By making explicit which traits underpin decisions, heatmap-guided models can act as a shared language between algorithms and experts, ensuring that automation remains accountable to biological reality. In this sense, AI does not displace expertise but amplifies it, highlighting patterns invisible to the human eye while allowing taxonomists to validate, question, and extend those patterns within existing frameworks (Wäldchen et al., 2022; Behrens et al., 2023).

In conclusion, this thesis demonstrates that interpretable AI is not only a technical advance but a new way of seeing morphology, one that extends the reach of NHCs, sharpens the practice of taxonomy, and provides reproducible baselines for the stewardship of biodiversity. At the same time, it is important to recognise that CV will not provide universal solutions, and molecular or anatomical evidence will remain essential. Rather than replacing such approaches, interpretable CV should be viewed as a complementary tool, one that accelerates large-scale identifications, uncovers hidden patterns, and integrates seamlessly with other lines of evidence to strengthen the study and conservation of biodiversity.

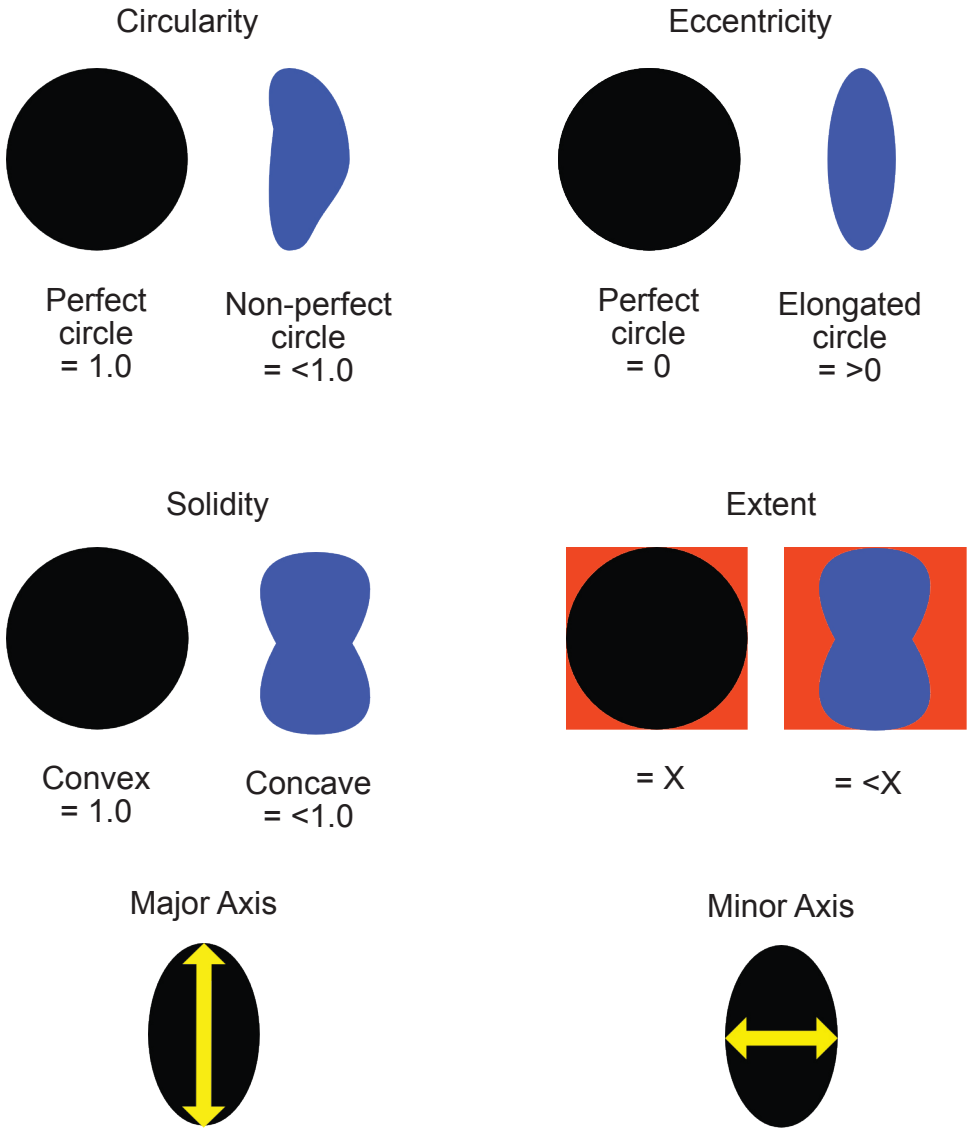
Appendix A Funding

Genome skimming was done within Biodiversity Genomics Europe (Grant no.101059492) which is funded by Horizon Europe under the Biodiversity, Circular Economy and Environment (REA.B.3); co-funded by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00173; and by the UK Research and Innovation (UKRI) under the Department for Business, Energy and Industrial Strategy's Horizon Europe Guarantee Scheme.

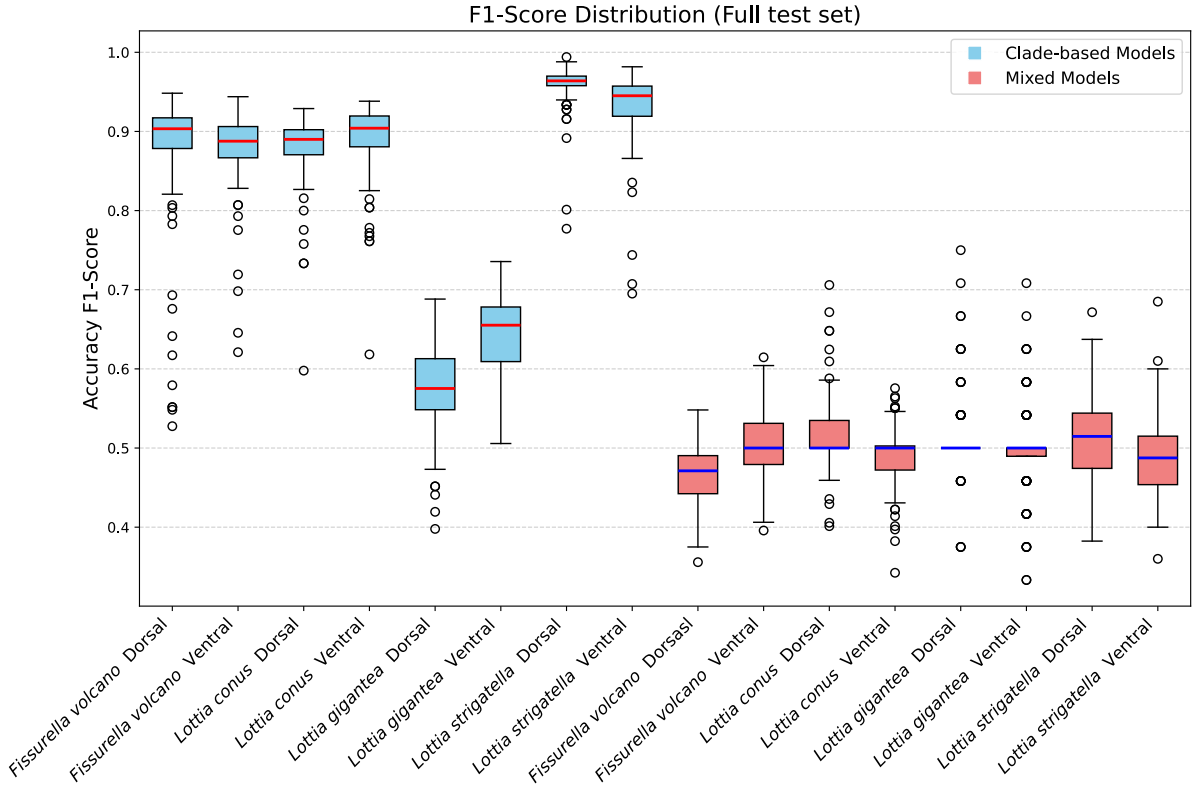
PBF acknowledges support from NERC grant NE/X011518/1.

JDH would like to thank NERC and the INSPIRE doctoral training programme for funding portions of this research.

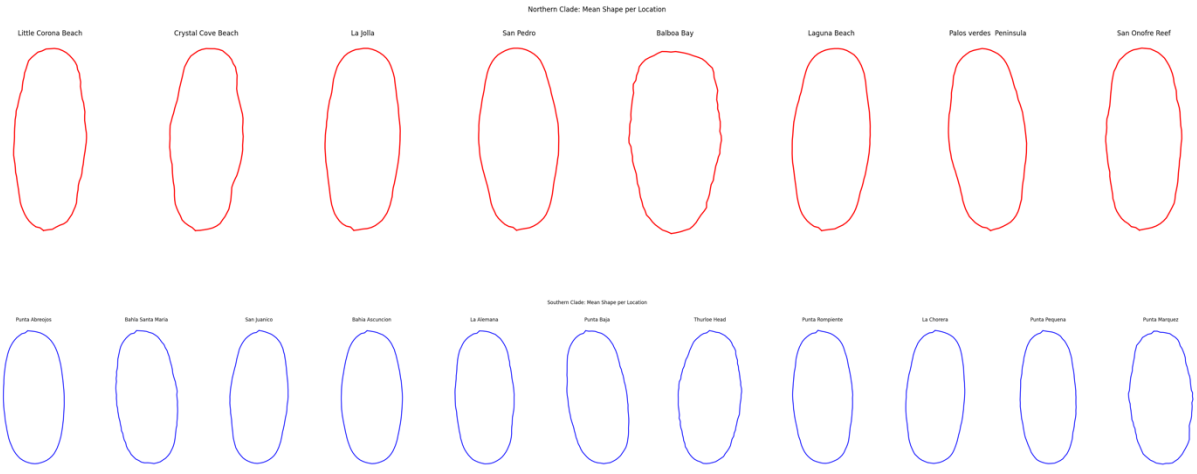
Appendix B Supplementary Data



Supplementary figure 1. Visual representations of the shape metrics.



Supplementary figure 2. box plots for model combination F1-scores across 100 runs for the full-test datasets. Blue boxes are the clade-based models per species and orientation, and the red boxes are the mixed-group controls. For each species and orientation, the F1-scores are significantly greater for the clade-based models compared to the mixed-group controls.



Supplementary figure 3. Karcher-means shape for all locations of Fissurella volcano specimens.

Number of limpets sampled per location and clade allocations per species (FV = Fissurella volcano; LC = Lottia conus; LG = Lottia gigantea; LS = Lottia strigatella). All specimens sampled in the field were collected in 2023. LACM is the Los Angeles County Museum of Natural History.

Species	Location	Sampled	Number	Clade	Latitude	Longitude
FV	La Chorera	Field	39	South	30.47	-116.06
FV	Balboa Bay	LACM	5	North	33.58	-117.86
FV	Crystal Cove Beach	LACM	7	North	33.57	-117.84
FV	La Jolla	LACM	34	North	32.83	-117.28
FV	Laguna Beach	LACM	42	North	33.54	-117.79
FV	Little Corona Beach	LACM	6	North	33.59	-117.87
FV	Palos verdes Peninsula	LACM	27	North	33.74	-118.42
FV	San Onofre Reef	LACM	10	North	33.37	-117.56
FV	San Pedro	LACM	70	North	33.70	-118.29
FV	Bahia Ascuncion	Field	97	South	27.12	-114.26
FV	Bahla Santa Maria	LACM	8	South	24.76	-112.27
FV	La Alemana	Field	16	South	29.95	-115.75
FV	Punta Abreojos	Field	114	South	26.70	-113.55
FV	Punta Baja	Field	20	South	29.95	-115.81
FV	Punta Marquez	LACM	3	South	23.96	-110.87
FV	Punta Pequena	LACM	23	South	26.23	-112.50
FV	Punta Rompiente	LACM	38	South	27.68	-114.93
FV	San Juanico	Field	36	South	26.23	-112.50
FV	Thurloe Head	LACM	4	South	27.62	-114.85
LC	Encinitas	LACM	4	North	33.04	-117.30
LC	San Pedro	LACM	51	North	33.70	-118.29
LC	La Jolla	LACM	9	North	32.83	-117.28
LC	Lunada Bay	LACM	7	North	33.77	-118.43
LC	Venice	LACM	30	North	33.97	-118.45
LC	Santa Monica	LACM	32	North	34.00	-118.49
LC	Campo Kennedy	Field	92	North	31.70	-116.68
LC	La Chorera	Field	42	North	30.47	-116.06
LC	UABC	Field	59	North	31.86	-116.67
LC	Punta Baja	Field	50	North	29.95	-115.81
LC	Bahia Ascuncion	Field	117	South	27.12	-114.26
LC	Punta Abreojos	Field	167	South	26.70	-113.55
LC	San Juanico	Field	400	South	26.23	-112.50
LC	Millers Landing	LACM	2	South	28.47	-114.05
LC	Punta Abreojos	LACM	14	South	26.70	-113.55
LC	Punta Pequena	LACM	3	South	26.23	-112.50
LC	Punta Rompiente	LACM	26	South	27.72	-115.02
LG	Bahia Ascuncion	Field	95	South	27.12	-114.26
LG	Punta Abreojos	Field	54	South	26.70	-113.55
LG	Cabo Thurloe	LACM	1	South	27.62	-109.92
LG	Pta Pequena	LACM	4	South	26.23	-112.50

LG	Shale Reef_Carpenteria	LACM	3	North	34.39	-119.52
LG	Topanga	LACM	6	North	34.03	-118.61
LG	Hazard Canyon	LACM	3	North	35.29	-120.88
LG	Pacific Grove	LACM	16	North	36.62	-121.92
LG	Thurloe Bay	LACM	15	South	27.62	-109.92
LG	Crescent City	LACM	10	North	41.75	-124.20
LG	Granite Creek	LACM	6	North	36.43	-121.91
LG	Waddell Beach	LACM	2	North	37.12	-122.28
LG	Santa Monica	LACM	13	North	34.00	-118.49
LG	Franklin Point	LACM	1	North	37.65	-86.87
LG	Venice	LACM	36	North	33.97	-118.45
LG	Santa Barbara	LACM	8	North	34.42	-119.70
LG	Morro Bay	LACM	27	North	35.33	-120.85
LG	Punta Eugenia	LACM	3	South	27.85	-115.08
LG	Punta Rompiente	LACM	7	South	27.68	-115.02
LG	Morro Rock	LACM	3	North	35.33	-120.87
LG	Pt Dume	LACM	27	North	34.00	-118.81
LG	Pt Reyes	LACM	1	North	38.00	-122.99
LG	Punta San Hipolito	LACM	3	South	26.98	*-113.98*
LG	Thurloe Head	LACM	1	South	27.62	-114.85
LG	Venice Breakwater	LACM	3	North	33.97	-118.45
LG	San Mateo	LACM	1	North	37.65	-122.33
LG	Elwood Pier	LACM	3	North	34.42	-119.92
LG	Anchorage isla Ascuncion	LACM	4	South	27.10	-114.29
LS	Bahia Ascuncion	Field	65	North	27.12	-114.26
LS	La Alemana	Field	33	North	29.95	-115.75
LS	Punta Abreojos	Field	79	North	26.70	-113.55
LS	Campo Kennedy	Field	25	North	31.70	-116.68
LS	Punta Baja	Field	42	North	29.95	-115.81
LS	San Juanico	Field	65	North	26.23	-112.50
LS	UABC	Field	15	North	31.86	-116.67
LS	Pozo De Cota	Field	115	South	23.02	-110.10
LS	Punta Marquez	Field	470	South	23.96	-110.87

Supplementary table 1. Number of limpets sampled per location and clade allocations per species (FV = Fissurella volcano; LC = Lottia conus; LG = Lottia gigantea; LS = Lottia strigatella). All specimens sampled in the field were collected in 2023. LACM is the Los Angeles County Museum of Natural History.

List of References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. And Kim, B. 2018. Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31, pp.9505–9515.

Advani, M. S., Saxe, A. M. & Sompolinsky, H. 2020. High-dimensional dynamics of generalization error in neural networks. Neural Networks, 132, 428-446.

Agnarsson, I. & Kuntner, M. 2007. Taxonomy in a changing world: seeking solutions for a science in crisis. Systematic biology, 56, 531-539.

Ahmed, W., Kommineni, V. K., König-Ries, B., Gaikwad, J., Gadelha, L. & Samuel, S. 2024. Evaluating the method reproducibility of deep learning models in the biodiversity domain. arXiv preprint arXiv:2407.07550.

Akpan, E. B. & Farrow, G. E. 1985. Shell bioerosion in high-latitude low-energy environments: Firths of Clyde and Lorne, Scotland. Marine Geology, 67, 139-150.

Allan, E. L., Livermore, L., Price, B. W., Shchedrina, O. & Smith, V. S. 2019. A novel automated mass digitisation workflow for natural history microscope slides. Biodiversity Data Journal, 7, e32342.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E. & Berthouze, N. Evaluating Saliency Map Explanations For Convolutional Neural Networks: A User Study. Proceedings Of The 25Th International Conference On Intelligent User Interfaces, 2020. 275-285.

Alsallakh, B., Kokhlikyan, N., Schaub-Meyer, S. And Merten, T. 2017. Do convolutional neural networks learn class hierarchy? arXiv preprint arXiv:1710.06501.

Anthropic, 2025. Claude Sonnet 4 model card. Available at: Https://Docs.Anthropic.Com/En/Docs/About-Claude/Models/Overview

Ariño, A. H. 2010. Approaches to estimating the universe of natural history collections data. Biodiversity Informatics, 7.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. & Bengio, Y. A Closer Look At Memorization In Deep Networks. International Conference On Machine Learning, 2017. PMLR, 233-242.

Austen, G. E., Bindemann, M., Griffiths, R. A. & Roberts, D. L. 2016. Species identification by experts and non-experts: comparing images from field guides. Scientific Reports, 6, 33634.

Avise, J. C. 2009. Phylogeography: retrospect and prospect. Journal of biogeography, 36, 3-15.

Aysel, H. I., Cai, X. & Prugel-Bennett, A. 2023. Multilevel explainable artificial intelligence: Visual and linguistic bonded explanations. IEEE Transactions on Artificial Intelligence, 5, 2055-2066.

Aysel, H. I., Cai, x. & Prugel-Bennett, A. 2025. Explainable artificial intelligence: advancements and limitations. applied sciences, 15, 7261.

Baird, C. N., Ernst, M., Waurick, I., Blom, M. P. & Bibi, F. 2024. Integrative taxonomy using historical specimens provides evidence for a single species of bushbuck, Tragelaphus scriptus (Mammalia: Bovidae). Zoological Journal of the Linnean Society, 200, 532-546.

Ballesteros, R., Intrigliolo, D. S., Ortega, J. F., Ramírez-Cuesta, J. M., Buesa, I. & Moreno, M. A. 2020. Vineyard yield estimation by combining remote sensing, computer vision and artificial neural network techniques. Precision Agriculture, 21, 1242-1262.

Balme, D. M. & Gotthelf, A. 2002. Aristotle:'Historia Animalium': Volume 1, Books IX

Barbedo, J.G.A. 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture, 153, pp.46–53.

Barton, K. E. 2024. The ontogenetic dimension of plant functional ecology. Functional Ecology, 38, 98-113.

Bayoudh, K. 2024. A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges. Information Fusion, 105, 102217.

Bayoudh, K., Knani, R., Hamdaoui, F. & Mtibaa, A. 2022. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer, 38, 2939-2970.

Beaman, R. S. & Cellinese, N. 2012. Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. ZooKeys, 7.

Behrens, M., Gube, M., Chaabene, H., Prieske, O., Zenon, A., Broscheid, K.-C., Schega, L., Husmann, F. & Weippert, M. 2023. Fatigue and human performance: an updated framework. Sports medicine, 53, 7-31.

Behrensmeyer, A. K., Kidwell, S. M. & Gastaldo, R. A. 2000. Taphonomy and paleobiology. Paleobiology, 26, 103-147.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K., Meier, R., Winker, K., Ingram, K. K. & Das, I. 2007. Cryptic species as a window on diversity and conservation. Trends in ecology & evolution, 22, 148-155.

Bik, H. M. 2017. Let's rise up to unite taxonomy and technology. PLoS Biol, 15, e2002231.

Bird, C.E. 2011. Morphological and behavioral evidence for adaptive diversification of sympatric Hawaiian limpets (Cellana spp.). Integrative and comparative biology 51, 466-473.

Blagoderov, V., Kitching, I. J., Livermore, L., Simonsen, T. J. & Smith, V. S. 2012. No specimen left behind: industrial scale digitization of natural history collections. ZooKeys, 133.

Blagoderov, V., Penn, M., Sadka, M., Hine, A., Brooks, S., Siebert, D. J., Sleep, C., Cafferty, S., Cane, E. & Martin, G. 2017. iCollections methodology: workflow, results and lessons learned. Biodiversity Data Journal.

Blair, J.D., Khidas, K. And Marshall, K.E. 2024. A gentle introduction to computer vision-based specimen classification in ecology and evolution. Journal of Animal Ecology, 93(3), pp.405–421.

Boaventura, D., Da Fonseca, L. S. C. & Hawkins, S. J. 2002. Analysis of competitive interactions between the limpets Patella depressa Pennant and Patella vulgata L. on the northern coast of Portugal. Journal of Experimental Marine Biology and Ecology, 271, 171-188.

Boiko, D. A., Macknight, R. & Gomes, G. 2023. Emergent autonomous scientific research capabilities of large language models. arXiv preprint arXiv:2304.05332.

Borowiec, M. L., Dikow, R. B., Frandsen, P. B., Mckeeken, A., Valentini, G. & White, A. E. 2022. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution, 13, 1640-1660.

Bortolus, A. 2008. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. AMBIO: A journal of the human environment, 37, 114-118.

Brasseur, M. V., Astrin, J. J., Geiger, M. F. & Mayer, C. 2023. MitoGeneExtractor: Efficient extraction of mitochondrial genes from next-generation sequencing libraries. Methods in Ecology and Evolution, 14, 1017-1024.

BRECKO, J. & MATHYS, A. 2020. Handbook of best practice and standards for 2D+ and 3D imaging of natural history collections. European Journal of Taxonomy.

Brito, D. 2010. Overcoming the Linnean shortfall: data deficiency and biological survey priorities. Basic and Applied Ecology, 11, 709-713.

Burdi, C. 2015. A test of diagnostic shell differences of the limpets Lottia conus and Lottia scabra identified with PCR-based assay. Masters, California state university.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M. & Kalinin, A. A. 2020. Albumentations: fast and flexible image augmentations. Information, 11, 125.

Cabon, L., Varma, M., Winter, G., Ebeling, A. & Schielzeth, H. 2025. Phenotypic plasticity, heritability, and genotype-by-environment interactions in an insect dispersal polymorphism. bioRxiv, 2025.03. 18.643873.

Carpenter, P.P., 1864. Diagnoses of new forms of Mollusca collected at Cape St. Lucas, Lower California, by Mr. Xantus. Annals and Magazine of Natural History, ser. 3, vol. 13, pp. 311–315 & 474–479.

Carranza-Rojas, J., Goeau, H., Bonnet, P., Mata-Montero, E. & Joly, A. 2017. Going deeper in the automated identification of Herbarium specimens. BMC evolutionary biology, 17, 181.

Castelvecchi, D. 2016. Can we open the black box of AI? Nature News, 538, 20.

Cawley, G. C. & Talbot, N. L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, 2079-2107.

Cerrejón, C., Noualhaguet, M., Fenton, N. J., Indorf, M.-F. & Feldman, M. J. 2025. Inconspicuous taxa in citizen science-based botanical research: actual contribution, limitations, and new opportunities for non-vascular cryptogams. Frontiers in Environmental Science, 12, 1448512.

Cheng, Z., Wu, Y., Li, Y., Cai, L. & Ihnaini, B. 2025. A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision. Sensors, 25, 4166.

Christin, S., Hervet, É. & Lecomte, N. 2019. Applications for deep learning in ecology. Methods in Ecology and Evolution, 10, 1632-1644.

Consortium, T. G. S., Aflitos, S., Schijlen, E., De Jong, H., De Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G. & Li, N. 2014. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. The Plant Journal, 80, 136-148.

Cook, J. A., Edwards, S. V., Lacey, E. A., Guralnick, R. P., Soltis, P. S., Soltis, D. E., Welch, C. K., Bell, K. C., Galbreath, K. E. & Himes, C. 2014. Natural history collections as emerging resources for innovative education. BioScience, 64, 725-734.

Cristescu, M. E. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in ecology & evolution, 29, 566-571.

Crummett, L.T., And Eernisse, D.J. 2007. Genetic evidence for the cryptic species pair, Lottia digitalis and Lottia austrodigitalis and microhabitat partitioning in sympatry. Marine Biology 152, 1-13.

Dale, J., Dey, C. J., Delhey, K., Kempenaers, B. & Valcu, M. 2015. The effects of life history and sexual selection on male and female plumage colouration. Nature, 527, 367-370.

Dawson, M.N., Hays, C.G., Grosberg, R.K., And Raimondi, P.T. 2014. Dispersal potential and population genetic structure in the marine intertidal of the eastern North Pacific. Ecological Monographs 84, 435-456.

Day, E.G., Branch, G.M., And Viljoen, C. 2000. How costly is molluscan shell erosion? A comparison of two patellid limpets with contrasting shell structures. Journal of Experimental Marine Biology and Ecology 243, 185-208.

Dayrat, B. 2005. Towards integrative taxonomy. Biological journal of the Linnean society, 85, 407-417.

Del Moral, P., Nowaczyk, S. & Pashami, S. 2022. Why is multiclass classification hard? IEEE Access, 10, 80448-80462.

Deng, C. H., Naithani, S., Kumari, S., Cobo-Simón, I., Quezada-Rodríguez, E. H., Skrabisova, M., Gladman, N., Correll, M. J., Sikiru, A. B. & Afuwape, O. O. 2023. Genotype and phenotype data standardization, utilization and integration in the big data era for agricultural sciences. Database, 2023, baad088.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. Imagenet: A Large-Scale Hierarchical Image Database. 2009 ieee conference on computer vision and pattern recognition, 2009. Ieee, 248-255.

Denis, M. & Schiffermüller, I. 1775. Ankündung eines systematischen Werkes von den Schmetterlingen der Wienergegend. verelgts Augustin Bernardi Buchhändler. Wien.

Dhal, P. & Azad, C. 2022. A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence, 52, 4543-4581.

Dharmaraaj, B. & Kunte, K. 2025. Natural and sexual selection and functional roles influence colouration but not the amount of variation in butterfly wing colour patterns. BMC Ecology and Evolution, 25, 11.

Doshi-Velez, F. & Kim, B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Ebach, M. C., Valdecasas, A. G. & Wheeler, Q. D. 2011. Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. Cladistics, 27, 550-557.

Eshghi, S., Rajabi, H., Poser, J. & Gorb, S. N. 2024. WingSegment: A Computer Vision-Based Hybrid Approach for Insect Wing Image Segmentation and 3D Printing. Advanced Intelligent Systems, 6, 2300712.

European Butterflies, 2024: available at http://www.european-butterflies.org.uk/species.html. (Accessed, 2 July, 2024)

Fabricius, J. C. 1793. Entomologia systematica emendata et aucta, Christ. Gottl. Proft.

Falkingham, P.L. 2012. Acquisition of high resolution three-dimensional models using free, open-source, photogrammetric software. Palaeontologia Electronica, 15(1), 1T.

Fenberg, P. B., Hellberg, M. E., Mullen, L. & Roy, K. 2010. Genetic diversity and population structure of the size-selectively harvested owl limpet, Lottia gigantea. Marine Ecology, 31, 574-583.

Fenberg, P. B., Self, A., Stewart, J. R., Wilson, R. J. & Brooks, S. J. 2016. Exploring the universal ecological responses to climate change in a univoltine butterfly. Journal of Animal Ecology, 85, 739-748.

Fenberg, P.B., And Roy, K. 2008. Ecological and evolutionary consequences of size-selective harvesting: how much do we know? Molecular ecology 17, 209-220.

Fenberg, P.B., And Roy, K. 2012. Anthropogenic harvesting pressure and changes in life history: insights from a rocky intertidal limpet. The American Naturalist 180, 200-210.

Firth, L.B. 2021. What have limpets ever done for us?: On the past and present provisioning and cultural services of limpets. International Review of Environmental History 7, 5-45.

Folmsbee, J., Johnson, S., Liu, X., Brandwein-Weber, M. & Doyle, S. Fragile Neural Networks: The Importance Of Image Standardization For Deep Learning In Digital Pathology. Medical Imaging 2019: digital pathology, 2019. SPIE, 222-228.

Foster, J. T., Price, L. B., Beckstrom-Sternberg, S. M., Pearson, T., Brown, W. D., Kiesling, D. M., Allen, C. A., Liu, C. M., Beckstrom-Sternberg, J. & Roberto, F. F. 2012. Genotyping of Brucella species using clade specific SNPs. BmC microbiology, 12, 1-8.

Garner, B., Allan, L., Crowther, R., Devenish, L., Kokkini, P., Livermore, L., Lowndes, N., Lohonya, K., Price, B. & Wing, P. 2024. The taxonomic and chronological composition of a museum collection of Coleoptera revealed through large-scale digitisation. Frontiers in Ecology and Evolution, 12, 1305931.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. & Wichmann, F. A. 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2, 665-673.

Giesel, J.T. 1970. On the maintenance of a shell pattern and behavior polymorphism in Acmaea digitalis, a limpet. Evolution 24, 98-119.

Glučina, M., Lorencin, A., Anđelić, N. & Lorencin, I. 2023. Cervical cancer diagnostics using machine learning algorithms and class balancing techniques. Applied Sciences, 13, 1061.

Godfray, H. C. J. 2002. Challenges for taxonomy. Nature, 417, 17-19.

Gong, Y., Liu, G., Xue, Y., Li, R. & Meng, L. 2023. A survey on dataset quality in machine learning. Information and Software Technology, 162, 107268.

Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R. & Scotland, R. W. 2015. Widespread mistaken identity in tropical plant collections. Current biology, 25, R1066-R1067.

Govaert, L., Altermatt, F., De Meester, L., Leibold, M. A., Mcpeek, M. A., Pantel, J. H. & Urban, M. C. 2021. Integrating fundamental processes to understand eco-evolutionary community dynamics and patterns. Functional Ecology, 35, 2138-2155.

Greeff, M., Caspers, M., Kalkman, V., Willemse, L., Sunderland, B.D., Bánki, O., And Hogeweg, L. 2022. Sharing taxonomic expertise between natural history collections using image recognition. Research Ideas and Outcomes 8, e79187.

Groom, Q., Dillen, M., Addink, W., Ariño, A. H., Bölling, C., Bonnet, P., Cecchi, L., Ellwood, E. R., Figueira, R., Gagnier, P.-Y., Grace, O., Güntsch, A., Hardy, H., Huybrechts, P., Hyam, R., Joly, A., Kommineni, V. K., Larridon, I., Livermore, L., Lopes, R. J., Meeus, S., Miller, J., Milleville, K., Panda, R., Pignal, M., Poelen, J., Ristevski, B., Robertson, T., Rufino, A., Santos, J., Schermer, M., Scott, B., Seltmann, K., Teixeira, H., Trekels, M. & Gaikwad, J. 2023. Envisaging a global infrastructure to exploit the potential of digitised collections. Biodiversity Data Journal, 11.

Grupstra, C. G., Gómez-Corrales, M., Fifer, J. E., Aichelman, H. E., Meyer-Kaiser, K. S., Prada, C. & Davies, S. W. 2024. Integrating cryptic diversity into coral evolution, symbiosis and conservation. Nature Ecology & Evolution, 8, 622-636.

GUIRAUD, M., GROOM, Q., BOGAERTS, A., DE SMEDT, S., DILLEN, M., SAARENMAA, H., WIJKAMP, N., VAN DER MIJE, S., WIJERS, A. & WU, Z. 2019. Best practice guidelines for imaging of herbarium specimens. ICEDIG, 41 pp.

Haahtela, T. 2019. Butterflies of Britain and Europe: A photographic guide, Bloomsbury Publishing.

Hajibabaei, M., Singer, G. A., Hebert, P. D. & Hickey, D. A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. TRENDS in Genetics, 23, 167-172.

Hall, A.C., Powell, O., Cubar, P. & Price, B.. 2023 ‘low-cost museum dna extraction using magnetic beads v2’, protocols.io [preprint]. doi:10.17504/protocols.io.4r3l27ebxg1y/v2.

Hamilton, A.M., Selwyn, J.D., Hamner, R.M., Johnson, H.K., Brown, T., Springer, S.K., And Bird, C.E. 2020. Biogeography of shell morphology in over-exploited shellfish reveals adaptive trade-offs on human-inhabited islands and incipient selectively driven lineage bifurcation. Journal of Biogeography 47, 1494-1509.

Hand, D. J. & Yu, K. 2001. Idiot's Bayes—not so stupid after all? International statistical review, 69, 385-398.

Hanly, J. J., Francescutti, C. M., Loh, L. S., Corning, O. B., Long, D. J., Nakatani, M. A., Porter, A. H. & Martin, A. 2023. Genetics of yellow-orange color variation in a pair of sympatric sulphur butterflies. Cell Reports, 42.

Hansen, O.L., Svenning, J.C., Olsen, K., Dupont, S., Garner, B.H., Iosifidis, A., Price, B.W., And Høye, T.T. 2020. Species-level image classification with convolutional neural network enables insect identification from habitus images. Ecology and Evolution 10, 737-747.

Hardisty, A., Saarenmaa, H., Casino, A., Dillen, M., Görddérz, K., Groom, Q., Hardy, H., Koureas, D., Nieva De La Hidalga, A. & Paul, D. L. 2020. Conceptual design blueprint for the DiSSCo digitization infrastructure-DELIVERABLE D8. 1. Research Ideas and Outcomes, 6.

Hardy, H., Livermore, L., Kersey, P., Norris, K. & Smith, V. 2023. Understanding the users and uses of UK Natural History Collections. Research Ideas and Outcomes, 9, e113378.

He, Y., Mulqueeney, J. M., Watt, E. C., Salili-James, A., Barber, N. S., Camaiti, M., Hunt, E. S., Kippax-Chui, O., Knapp, A. & Lanzetti, A. 2024. Opportunities and challenges in applying AI to evolutionary morphology. Integrative Organismal Biology, 6, obae036.

Heberling, J. M. & Isaac, B. L. 2018. iN aturalist as a tool to expand the research value of museum specimens. Applications in Plant Sciences, 6, e01193.

Hebert, P. D., Cywinska, A., Ball, S. L. & Dewaard, J. R. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270, 313-321.

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy of Sciences, 101, 14812-14817.

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A. & Blackburn, D. C. 2020. Digitization and the future of natural history collections. BioScience, 70, 243-251.

Hending, D. 2025. Cryptic species conservation: a review. Biological Reviews, 100, 258-274.

Hendry, D., Carter, D. & Walker, A. 2016. Care and conservation of natural history collections, Oxford, Butterwoth Heinemann

Hennig, W. 1966. Phylogenetic Systematics. University of Illinois Press.

Hof, A. E. V. T., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C. & Saccheri, I. J. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature, 534, 102-105.

Hollister, J. D., Cai, X., Horton, T., Price, B. W., Zarzyczny, K. M. & Fenberg, P. B. 2023. Using computer vision to identify limpets from their shells: a case study using four species from the Baja California peninsula. Frontiers in Marine Science, 10.

Hollister, J., Vega, R. & Azhar, M. A. H. B. 2022. Automatic Identification of Non-biting Midges (Chironomidae) using Object Detection and Deep Learning Techniques. Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods.

Hollister, J.D., Martin, G., Cai, X., Horton, T., Powell, O., Sterling, M., Turnbull, G., Price, B.W. And Fenberg, P.B., 2025. A Computer Vision Method for Finding Mislabelled Specimens Within Natural History Collections. Ecology and Evolution, 15(7), p.e71648.

Holmes, M.W., Hammond, T.T., Wogan, G.O.U., Walsh, R.E., Labarbera, K., Wommack, E.A., Martins, F.M., Crawford, J.C., Mack, K.L., Bloch, L.M. And Nachman, M.W. (2016). Natural history collections provide an immense record of biodiversity on Earth. Proceedings of the National Academy of Sciences, 113(4), pp.1203–1208.

Høye, T.T., Ärje, J., Bjerge, K., Hansen, O.L., Iosifidis, A., Leese, F., Mann, H.M., Meissner, K., Melvad, C., And Raitoharju, J. 2021. Deep learning and computer vision will transform entomology. Proceedings of the National Academy of Sciences 118, e2002545117.

Isaac, N. J., Mallet, J. & Mace, G. M. 2004. Taxonomic inflation: its influence on macroecology and conservation. Trends in ecology & evolution, 19, 464-469.

Jansen, M., Beukes, M., Weiland, C., Blumer, M., Rudolfi, M., Poerting, J., Meißner, R., Weiß, M., Condori, Y. & Aramayo-Ledezma, G. 2024. Engaging citizen scientists in biodiversity monitoring: insights from the WildLIVE! Project. Citizen Science: Theory and Practice, 9.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E. & Fung, P. Towards Mitigating Llm Hallucination Via Self Reflection. Findings Of The Association For Computational Linguistics: Emnlp 2023, 2023. 1827-1843.

Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M. & Wei, Y. 2021. Layercam: Exploring hierarchical class activation maps for localization. IEEE Transactions on Image Processing, 30, 5875-5888.

Jiang, Q., Gao, Z. & Karniadakis, G. E. 2025. DeepSeek vs. ChatGPT vs. Claude: A comparative study for scientific computing and scientific machine learning tasks. Theoretical and Applied Mechanics Letters, 15, 100583.

Johannes, H., Verónica, S. & Miguel, V. 2024. Generative AI to Understand Complex Ecological Interactions. Applications of Generative AI. Springer.

Joshi, S., Owens, J.A., Shah, S., And Munasinghe, T. 2021. "Analysis Of Preprocessing Techniques, Keras Tuner, And Transfer Learning On Cloud Street Image Data", In: 2021IEEE International Conference on Big Data (Big Data): IEEE), 4165-4168.

Kapp, J. D., Green, R. E. & Shapiro, B. 2021. A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. Journal of Heredity, 112, 241-249.

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., Wagner, N. D., Emerson, B. C., Albach, D. C. & Scheu, S. 2024A. Species delimitation 4.0: integrative taxonomy meets artificial intelligence. Trends in Ecology & Evolution.

Keen, A.M. 1971. Sea Shells of Tropical West America: Marine Mollusks from Baja California to Peru. Stanford University Press.

Khalid, S., Khalil, T. & Nasreen, S. A Survey Of Feature Selection And Feature Extraction Techniques In Machine Learning. 2014 science and information conference, 2014. IEEE, 372-378.

Khalifa, N. E., Loey, M. & Mirjalili, S. 2022. A comprehensive survey of recent trends in deep learning for digital images augmentation. Artificial Intelligence Review, 55, 2351-2377.

Kido, J. S. & Murray, S. N. 2003. Variation in owl limpet Lottia gigantea population structures, growth rates, and gonadal production on southern California rocky shores. Marine Ecology Progress Series, 257, 111-124.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Wright, R., Ganguli, D., Yastrow, A., Spilsbury, C., Zitnick, C.L., Dollár, P. And Girshick, R., 2023. Segment Anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. pp.4015–4026

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C. & Lo, W.-Y. Segment Anything. Proceedings Of The Ieee/Cvf International Conference On Computer Vision, 2023. 4015-4026.

Kordas, R.L., Donohue, I., And Harley, C.D.G. 2017. Herbivory enables marine communities to resist warming. Science Advances 3, e1701349.

Kuo, E.S., And Sanford, E. (2013). Northern Distribution of the Seaweed Limpet Lottia insessa (Mollusca: Gastropoda) along the Pacific Coast. Pacific Science 67, 303-313.

Kürzel, K., Kaiser, S., Lörz, A.-N., Rossel, S., Paulus, E., Peters, J., Schwentner, M., Martinez Arbizu, P., Coleman, C. O. & Svavarsson, J. 2022. Correct species identification and its implications for conservation using Haploniscidae (Crustacea, Isopoda) in Icelandic waters as a proxy. Frontiers in Marine Science, 8, 795196.

Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. 2020. Explainable ai: A review of machine learning interpretability methods. Entropy, 23, 18.

Linnaeus, C. 1758. Systema Naturae, 10th edit. Vol. I, 823.

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16, 31-57.

Lister, A. M. 2011. Natural history collections as sources of long-term datasets. Trends in ecology & evolution, 26, 153-154.

Liu, K.-H., Yang, M.-H., Huang, S.-T. & Lin, C. 2022. Plant species classification based on hyperspectral imaging via a lightweight convolutional neural network model. Frontiers in Plant Science, 13, 855660.

Liu, L., Meng, J. & Yang, Y. 2024A. LLM technologies and information search. Journal of Economy and Technology, 2, 269-277.

Liu, Y., Cao, J., Liu, C., Ding, K. & Jin, L. 2024B. Datasets for large language models: A comprehensive survey. arXiv preprint arXiv:2402.18041.

Livraghi, L., Hanly, J. J., Loh, L. S., Henry, A., Keck, C., Shirey, V. M., Tsai, C.-C., Yu, N., Van Belleghem, S. M. & Roberts, W. M. 2025. Genetic basis of an adaptive polymorphism controlling butterfly silver iridescence. Current Biology, 35, 2154-2163. e7.

Lu, S., Liu, L., Lei, W., Wang, D., Zhu, H., Lai, Q., Ma, L. & Ru, D. 2024. Cryptic divergence in and evolutionary dynamics of endangered hybrid Picea brachytyla sensu stricto in the Qinghai-Tibet Plateau. BMC Plant Biology, 24, 1202.

Lughadha, E. N. 2004. Towards a working list of all known plant species. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359, 681-687.

Luikart, G., Kardos, M., Hand, B. K., Rajora, O. P., Aitken, S. N. & Hohenlohe, P. A. 2018. Population genomics: advancing understanding of nature. Population genomics: Concepts, approaches and applications. Springer.

Lundberg, S. M. & Lee, S.-I. 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Lürig, M. D., Donoughe, S., Svensson, E. I., Porto, A. & Tsuboi, M. 2021. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. Frontiers in Ecology and Evolution, 9, 642774.

Mace, G. M. 2004. The role of taxonomy in species conservation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359, 711-719.

Mallet, J. & Joron, M. 1999. Evolution of diversity in warning color and mimicry: polymorphisms, shifting balance, and speciation. Annual review of ecology and systematics, 30, 201-233.

Mallet, J. 2005. Hybridization as an invasion of the genome. Trends in ecology & evolution, 20, 229-237.

Mallis, M.M., Mejdal, S., Nguyen, T.T., And Dinges, D.F. 2004. Summary of the key features of seven biomathematical models of human fatigue and performance. Aviation, space, and environmental medicine 75, 4-14.

Mann, K., Edsinger-Gonzales, E. & Mann, M. 2012. In-depth proteomic analysis of a mollusc shell: acid-soluble and acid-insoluble matrix of the limpet Lottia gigantea. Proteome science, 10, 1-18.

Marinček, P., Wagner, N. D. & Tomasello, S. 2022. Ancient DNA extraction methods for herbarium specimens: When is it worth the effort? Applications in Plant Sciences, 10, e11477.

Mclean, J. H. 1984. Systematics of Fissurella in the Peruvian and Magellanic faunal provinces (Gastropoda: Prosobranchia), Natural History Museum of Los Angeles County.

Meineke, E. K., Davis, C. C. & Davies, T. J. 2018. The unrealized potential of herbaria for global change biology. Ecological Monographs, 88, 505-525.

Merrill, R. M., Dasmahapatra, K. K., Davey, J., Dell'Aglio, D., Hanly, J., Huber, B., Jiggins, C. D., Joron, M., Kozak, K. & Llaurens, V. 2015. The diversification of Heliconius butterflies: what have we learned in 150 years? Journal of Evolutionary Biology, 28, 1417-1438.

Mesnick, S. & Ralls, K. 2018. Sexual dimorphism. Encyclopedia of marine mammals. Elsevier.

Meyer, C. P. & Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS biology, 3, e422.

Miller, B., Conway, W., Reading, R. P., Wemmer, C., Wildt, D., Kleiman, D., Monfort, S., Rabinowitz, A., Armstrong, B. & Hutchins, M. 2004. Evaluating the conservation mission of zoos, aquariums, botanical gardens, and natural history museums. Conservation Biology, 18, 86-93.

Molbert, N., Ghanavi, H. R., Johansson, T., Mostadius, M. & Hansson, M. C. 2023. An evaluation of DNA extraction methods on historical and roadkill mammalian specimen. Scientific Reports, 13, 13080.

Morris, R.H., Abbott, D.P., And Haderlie, E.C. 1980. Intertidal invertebrates of California. Stanford University Press Stanford.

Moura, M. R. & Jetz, W. 2021. Shortfalls and opportunities in terrestrial vertebrate species discovery. Nature Ecology & Evolution, 5, 631-639.

Mujtaba, T., Lawrence, M., Oliver, M. & Reiss, M. J. 2018. Learning and engagement through natural history museums. Studies in science education, 54, 41-67.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116, 22071-22080.

Nakano, T., And Spencer, H.G. 2007. Simultaneous polyphenism and cryptic species in an intertidal limpet from New Zealand. Molecular Phylogenetics and Evolution 45, 470-479.

Nazari, V. & Evans, L. 2015. Butterflies of ancient Egypt. The Journal of the Lepidopterists' Society, 69, 242-267.

Nelson, G. & Ellis, S. 2019. The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B, 374, 20170391.

Nguyen, C., Lovell, D., Adcock, M. And La Salle, J. 2017. Towards high-throughput 3D insect capture for species discovery and diagnostics. arXiv preprint arXiv:1709.02033.

Nielsen, E. E., Hemmer-Hansen, J., Larsen, P. F. & Bekkevold, D. 2009. Population genomics of marine fishes: identifying adaptive variation in space and time. Molecular ecology, 18, 3128-3150.

Nielsen, E. S., Walkes, S., Sones, J. L., Fenberg, P. B., Paz-García, D. A., Cameron, B. B., Grosberg, R. K., Sanford, E. & Bay, R. A. 2024. Pushed waves, trailing edges, and extreme events: Eco-evolutionary dynamics of a geographic range shift in the owl limpet, Lottia gigantea. Global Change Biology, 30, e17414.

Nijhout, H. F. 1986. Pattern and pattern diversity on Lepidopteran wings. Bioscience, 36, 527-533.

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C. & Clune, J. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences, 115, E5716-E5725.

Northcutt, C. G., Athalye, A. & Mueller, J. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749.

Nugent, J. 2018. iNaturalist: Citizen science for 21st-century naturalists. Science Scope, 41, 12-15.

Oliver, J. C., Robertson, K. A. & Monteiro, A. 2009. Accommodating natural and sexual selection in butterfly wing pattern evolution. Proceedings of the Royal Society B: Biological Sciences, 276, 2369-2375.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H. & Belinkov, Y. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. arXiv preprint arXiv:2410.02707.

Oróstica, M. H., Richardson, C. A., Estrella-Martínez, J., Jenkins, S. R. & Hawkins, S. J. 2021. Shell growth and age determined from annual lines in the southern warm-water limpet Patella depressa at its poleward geographic boundaries. Journal of the Marine Biological Association of the United Kingdom, 101, 707-716.

Oyedotun, O. K., Olaniyi, E. O. & Khashman, A. 2017. A simple and practical review of over-fitting in neural network learning. International Journal of Applied Pattern Recognition, 4, 307-328.

Padial, J. M., Miralles, A., De La Riva, I. & Vences, M. 2010. The integrative future of taxonomy. Frontiers in zoology, 7, 16.

Paknia, O., Rajaei Sh, H. & Koch, A. 2015. Lack of well-maintained natural history collections and taxonomists in megadiverse developing countries hampers global biodiversity exploration. Organisms Diversity & Evolution, 15, 619-629.

Parsons, D. J., Pelletier, T. A., Wieringa, J. G., Duckett, D. J. & Carstens, B. C. 2022. Analysis of biodiversity data suggests that mammal species are hidden in predictable places. Proc Natl Acad Sci U S A, 119, e2103400119.

Paterson, G., Albuquerque, S., Blagoderov, V., Brooks, S., Cafferty, S., Cane, E., Carter, V., Chainey, J., Crowther, R. & Douglas, L. 2016. iCollections–Digitising the British and Irish Butterflies in the Natural History Museum, London. Biodiversity Data Journal.

Paul, D. L., Thompson, C. W., Arroyo, L., Nuñez, G. B., Castro-Arellano, I., Colella, J. P., Cook, J. A., Cove, M. V., Dearborn, J. & De La Sancha, N. U. 2025. Harnessing natural history collections for collaborative pandemic preparedness. BioScience, biaf035.

Paulo Cabral, J. 2007. Shape and growth in European Atlantic Patella limpets (Gastropoda, Mollusca). Ecological implications for survival. Web Ecology, 7, 11-21.

Perez, L., And Wang, J. 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Pfenninger, M. & Schwenk, K. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC evolutionary biology, 7, 1-6.

Picard, S., Chapdelaine, C., Cappi, C., Gardes, L., Jenn, E., Lefèvre, B. & Soumarmon, T. Ensuring Dataset Quality For Machine Learning Certification. 2020 ieee international symposium on software reliability engineering workshops (issrew), 2020. IEEE, 275-282.

Pichler, M. & Hartig, F. 2023. Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution, 14, 994-1016.

Pilsbry, H. 1890. Manual of conchology, structural and systematic, with illustrations of the species. 1(12): Stomatellidae, Scissurellidae, Pleurotomariidae, Haliotidae, Scutellinidae, Addisoniidae, Cocculinidae, Fissurellidae. Philadelphia: Conchological Section, Academy of Natural Sciences.

Pinho, C., Kaliontzopoulou, A., Ferreira, C. A. & Gama, J. 2023. Identification of morphologically cryptic species with computer vision models: wall lizards (Squamata: Lacertidae: Podarcis) as a case study. Zoological Journal of the Linnean Society, 198, 184-201.

Piva, A., Raimondi, L., Rasca, E., Kozmanyan, A. & De Matteis, M. 2024. A machine learning application for the automatic recognition of planktonic foraminifera in thin sections. Marine and Petroleum Geology, 166, 106911.

Poon, S. T., Leong, A. M., Fogerty, T., Twitchett, R., Salili-James, A., Stukins, S., Scott, B. & Smith, V. S. 2024. Automatic Detection and Identification of Calcareous Nannofossils in Chalk Using Deep Learning: A Proof-of-Concept Study for Biostratigraphy and Climate Research. Biodiversity Information Science and Standards, 8, e138673.

Popov, D., Roychoudhury, P., Hardy, H., Livermore, L., And Norris, K. 2021. The Value of Digitising Natural History Collections. Research Ideas and Outcomes 7, e78844.

Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular ecology, 21, 1864-1877.

Raxworthy, C. J. & Smith, B. T. 2021. Mining museums for historical DNA: advances and challenges in museomics. Trends in ecology & evolution, 36, 1049-1060.

Rayo, E., Ulrich, G. F., Zemp, N., Greeff, M., Schuenemann, V. J., Widmer, A. & Fischer, M. C. 2024. Minimally destructive hDNA extraction method for retrospective genetics of pinned historical Lepidoptera specimens. Scientific Reports, 14, 12875.

Reeve, L. 1849. Monograph of the genus Fissurella. In: Conchologia Iconica, or, illustrations of the shells of molluscous animals. London: Reeve & Co.

Ribeiro, M. T., Singh, S. & Guestrin, C. " Why Should I Trust You?" Explaining The Predictions Of Any Classifier. Proceedings Of The 22Nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining, 2016. 1135-1144.

Riddle, B. R., Hafner, D. J., Alexander, L. F. & Jaeger, J. R. 2000. Cryptic vicariance in the historical assembly of a Baja California Peninsular Desert biota. Proceedings of the National Academy of Sciences, 97, 14438-14443.

Rogers, A.J., And Weisler, M.I. 2020A. Assessing the efficacy of genus-level data in archaeomalacology: A case study of the Hawaiian limpet (Cellana spp.), Moloka ‘i, Hawaiian islands. The Journal of Island and Coastal Archaeology 15, 28-56.

Rogers, A.J., And Weisler, M.I. 2020B. Limpet (Cellana spp.) shape is correlated with basalt or eolianite coastlines: Insights into prehistoric marine shellfish foraging and mobility in the Hawaiian Islands. Journal of Archaeological Science: Reports 34, 102561.

Ross, E. (2022). Phylogeography of the cryptic intertidal gastropod Lottia conus along the Pacific coast from Southern California to Central Mexico. Masters, University of Southampton.

Rowe, K. C., Singhal, S., Macmanes, M. D., Ayroles, J. F., Morelli, T. L., Rubidge, E. M., Bi, K. & Moritz, C. C. 2011. Museum genomics: low-cost and high-accuracy genetic data from historical specimens. Molecular Ecology Resources, 11, 1082-1092.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1, 206-215.

Rädsch, T., Reinke, A., Weru, V., Tizabi, M.D., Schreck, N., Kavur, A.E., Pekdemir, B., Roß, T., Kopp-Schneider, A., And Maier-Hein, L. 2023. Labelling instructions matter in biomedical image analysis. Nature Machine Intelligence 5, 273-283.

Salmon, M. A. 2000. ‘The Grand Panacea’A Short History of Butterfly Collecting in Britain. The Aurelian Legacy–a History of British Butterflies and their Collectors. Brill.

Sauer, F. G., Werny, M., Nolte, K., Villacañas De Castro, C., Becker, N., Kiel, E. & Lühken, R. 2024. A convolutional neural network to identify mosquito species (Diptera: Culicidae) of the genus Aedes by wing images. Scientific Reports, 14, 3094.

Savage, N. 2022. Breaking into the black box of artificial intelligence. Nature. https://www.nature.com/articles/d41586-022-00858-1

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C. & Brännström, Å. 2014. Genomics and the origin of species. Nature Reviews Genetics, 15, 176-192.

Seeland, M., Rzanny, M., Boho, D., Wäldchen, J. & Mäder, P. 2019. Image-based classification of plant genus and family for trained and untrained plant species. BMC bioinformatics, 20, 1-13.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. International journal of computer vision, 128, 336-359.

Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D. & Batra, D. 2016. Grad-CAM: Why did you say that? arXiv preprint arXiv:1611.07450.

Sempere-Valverde, J., Ostalé-Valriberas, E., Espinosa, F. & Márquez, F. 2024. Morphometric variations of two patellid limpets between artificial breakwaters and natural reefs. Estuarine, Coastal and Shelf Science, 297, 108617.

Shaffer, H. B., Fisher, R. N. & Davidson, C. 1998. The role of natural history collections in documenting species declines. Trends in ecology & evolution, 13, 27-30.

Sham, A.H., Aktas, K., Rizhinashvili, D., Kuklianov, D., Alisinanoglu, F., Ofodile, I., Ozcinar, C., And Anbarjafari, G. 2022. Ethical AI in facial expression analysis: racial bias. Signal, Image and Video Processing 17, 1-8.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A. & Waterston, R. H. 2017. DNA sequencing at 40: past, present and future. Nature, 550, 345-353.

Shi, X., Liu, J., Liu, Y., Cheng, Q. & Lu, W. 2025. Know where to go: Make LLM a relevant, responsible, and trustworthy searchers. Decision Support Systems, 188, 114354.

Shorten, C. & Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. Journal of big data, 6, 1-48.

Simison, W.B., And Lindberg, D.R. 1999. Morphological and molecular resolution of a putative cryptic species complex: a case study of Notoacmea fascicularis (Menke, 1851)(Gastropoda: Patellogastropoda). Journal of Molluscan Studies 65, 99-109.

Simison, W.B., And Lindberg, D.R. 2003. On the identity of Lottia strigatella (Carpenter, 1864)(Patellogastropoda: Lottiidae). Veliger 46, 1-19.

Simonyan, K. & Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Simonyan, K., Vedaldi, A. & Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Slater, G. S. C. & Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. BMC bioinformatics, 6, 1-11.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.

Smith, J., Wycherley, A., Mulvaney, J., Lennane, N., Reynolds, E., Monks, C.-A., Evans, T., Mooney, T. & Fancourt, B. 2024. Man versus machine: cost and carbon emission savings of 4G-connected Artificial Intelligence technology for classifying species in camera trap images. Scientific Reports, 14, 14530.

Sohan, M., Sai Ram, T., Reddy, R. & Venkata, C. A Review On Yolov8 And Its Advancements. International Conference On Data Intelligence And Cognitive Informatics, 2024. Springer, 529-545.

Sowerby, G.B., 1834. The genera of recent and fossil shells, for the use of students, in conchology and geology. Vol. 2. London: G.B. Sowerby.

Sultan, S. E. 2000. Phenotypic plasticity for plant development, function and life history. Trends in plant science, 5, 537-542.

Szczepankiewicz, B., Markiewicz, M., Podolak, I. And Spurek, P. 2023. Ground-truth based comparison of saliency map algorithms. Scientific Reports, 13, 14786.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology, 21, 2045-2050.

Tamura, K., Stecher, G., And Kumar, S. 2021. MEGA11: molecular evolutionary genetics analysis version 11. Molecular biology and evolution 38, 3022-3027.

Tang, X.-T., Cai, L., Shen, Y. & Du, Y.-Z. 2018. Diversity and evolution of the endosymbionts of Bemisia tabaci in China. PeerJ, 6, e5516.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., And Vogler, A.P. 2003. A plea for DNA taxonomy. Trends in ecology & evolution 18, 70-74.

Test, A. 1945. Description of new species of Acmaea. Nautilus, 58, 92-96.

Thomas, J. 2020. The butterflies of Britain and Ireland, Bloomsbury Publishing.

Toews, D. P. & Brelsford, A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. Molecular ecology, 21, 3907-3930.

Tolman, T. 2008. Collins butterfly guide, HarperCollins UK

Trail, P. W. 2021. Morphological analysis: A powerful tool in wildlife forensic biology. Forensic Science International: Animals and Environments, 1, 100025.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. 2017. Taxonomic bias in biodiversity data and societal preferences. Scientific reports, 7, 9132.

Trussell, G. C. 1996. Phenotypic plasticity in an intertidal snail: the role of a common crab predator. Evolution, 448-454.

Urdy, S., Goudemand, N., Bucher, H. & Chirat, R. 2010. Growth-dependent phenotypic variation of molluscan shells: implications for allometric data interpretation. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 314, 303-326.

Van Belleghem, S. M., Lewis, J. J., Rivera, E. S. & Papa, R. 2021. Heliconius butterflies: a window into the evolution and development of diversity. Current opinion in genetics & development, 69, 72-81.

Vasconcelos, J., Caamaño, D., Tuset, V. M., Sousa, R. & Riera, R. 2021. The shell phenotypic variability of the keyhole limpet Fissurella latimarginata: insights from an experimental approach using a water flow flume. Journal of Molluscan Studies, 87, eyab043.

Vermeij, G. J. 1973. Morphological patterns in high-intertidal gastropods: adaptive strategies and their limitations. Marine Biology, 20, 319-346.

Vermeij, G. J. 2002. Characters in context: molluscan shells and the forces that mold them. Paleobiology, 28, 41-54.

Vollmar, A., Macklin, J. A. & Ford, L. 2010. Natural history specimen digitization: challenges and concerns. Biodiversity informatics, 7.

Von Eschenbach, W. J. 2021. Transparency and the black box problem: Why we do not trust AI. Philosophy & technology, 34, 1607-1622.

Walch, J.E.I., 1775. Papilio fulgerator description. Der Naturforscher, 7, p.115, pl. 1, figs. 2a–2b.

Walters, A. D., Cannizzaro, A. G., Trujillo, D. A. & Berg, D. J. 2021. Addressing the Linnean shortfall in a cryptic species complex. Zoological Journal of the Linnean Society, 192, 277-305.

Wan, Z., Wang, Z., Chung, C. & Wang, Z. 2024. A survey of dataset refinement for problems in computer vision datasets. ACM computing surveys, 56, 1-34.

Weisler, M.I., And Rogers, A.J. 2021. Ritual use of limpets in late Hawaiian prehistory. Journal of Field Archaeology 46, 52-61.

Wilcox, B. A. & Ellis, B. 2006. Forests and emerging infectious diseases of humans. UNASYLVA-FAO-, 57, 11.

Will, K. W., Mishler, B. D. & Wheeler, Q. D. 2005. The perils of DNA barcoding and the need for integrative taxonomy. Systematic biology, 54, 844-851.

Williams, S. T. 2017. Molluscan shell colour. Biological Reviews, 92, 1039-1058.

Wilson, R. J., De Siqueira, A. F., Brooks, S. J., Price, B. W., Simon, L. M., Van Der Walt, S. J. & Fenberg, P. B. 2023. Applying computer vision to digitised natural history collections for climate change research: Temperature-size responses in British butterflies. Methods in Ecology and Evolution, 14, 372-384.

Wäldchen, J. & Mäder, P. 2018. Machine learning for image based species identification. Methods in Ecology and Evolution, 9, 2216-2225.

Wright, W. G. 1988. Sex change in the Mollusca. Trends in Ecology & Evolution, 3, 137-140.

Wright, W. G. & Lindberg, D. R. 1982. Direct observation of sex change in the patellacean limpet *Lottia gigantea*. Journal of the Marine Biological Association of the United Kingdom, 62, 737-738.

Xu, M., Yoon, S., Fuentes, A. & Park, D. S. 2023. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognition, 137, 109347.

Yap, J. Q. H., Kamble, Z., Kuah, A. T. & Tolkach, D. 2024. The impact of digitalisation and digitisation in museums on memory-making. Current Issues in Tourism, 27, 2538-2560.

Yaqoob, M., Ishaq, M., Ansari, M. Y., Qaiser, Y., Hussain, R., Rabbani, H. S., Garwood, R. J. & Seers, T. D. 2025. Advancing paleontology: a survey on deep learning methodologies in fossil image analysis. Artificial Intelligence Review, 58, 83.

Ying, X. An Overview Of Overfitting And Its Solutions. Journal Of Physics: Conference Series, 2019. IOP Publishing, 022022.

Younis, S., Schmidt, M., Weiland, C., Dressler, S. And Seeger, B. 2020. Deep learning for plant identification using digitized herbarium specimens. Frontiers in Plant Science, 9, 841.

Younis, S., Schmidt, M., Weiland, C., Dressler, S., Seeger, B. & Hickler, T. 2020. Detection and annotation of plant organs from digitised herbarium scans using deep learning. Biodiversity Data Journal, 8, e57090.

Zarzyczny, K. M., Et Al. (2024). "The ecological and evolutionary consequences of tropicalisation." Trends in Ecology & Evolution 39(3): 267-279.

Zarzyczny, K. M., Hellberg, M. E., Lugli, E. B., Maclean, M., Paz-García, D. A., Rius, M., Ross, E. G., Treviño Balandra, E. X., Vanstone, J. & Williams, S. T. 2024. Opposing genetic patterns of range shifting temperate and tropical gastropods in an area undergoing tropicalisation. Journal of Biogeography, 51, 246-262.

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64, 107-115.

Zhang, S., Wang, Y. & Wu, G. 2022. Earthquake-induced landslide susceptibility assessment using a novel model based on gradient boosting machine learning and class balancing methods. Remote Sensing, 14, 5945.

Zhou, C.-L., Ge, L.-M., Guo, Y.-B., Zhou, D.-M. & Cun, Y.-P. 2021. A Comprehensive Comparison On Current Deep Learning Approaches For Plant Image Classification. Journal Of Physics: Conference Series, 2021. IOP Publishing, 012002.