Automated Diet Matrix Construction for Marine Ecosystem Models Using Generative AI

Scott Spillias[1,2*] Beth Fulton[1,2] Fabio Boschetti[2,3]
Cathy Bulman[1] Rodrigo Bustamante[4] Javier Porobic Garate[1,2]
Joanna Strzelecki[3] Roshni Subramaniam[1,2] Rowan Trebilco[1,2]

January 28, 2025

[1]CSIRO Environment, Hobart, Australia
[2]Centre for Marine Socio-Ecology, University of Tasmania, Hobart, Australia
[3]CSIRO Environment, IOMRC Crawley, Australia
[4]CSIRO Environment, St. Lucia, Australia

Abstract

This study introduces and validates a novel AI-driven framework for automated species grouping in Ecopath with Ecosim (EwE) ecosystem models, addressing a critical bottleneck in model development. We evaluate the framework across three contrasting Australian marine regions, processing over 41,000 species through multiple validation iterations. The framework successfully condensed 63 potential functional groups into 34-36 region-specific groups, achieving high classification stability (>99% consistency) for most species. Notably, the framework demonstrated robust performance in the South East Offshore region with only 0.03% inconsistent classifications, while showing greater variability in complex tropical systems (1% inconsistent classifications). Higher trophic level species maintained consistent classifications across all runs, with the framework identifying 235-327 significant predator-prey interactions per region at >70% consistency. This systematic validation reveals that while the framework can reliably automate species grouping, its performance varies predictably with ecosystem complexity and data availability. These findings provide quantitative evidence for the framework's capability to accelerate ecosystem model development while highlighting specific conditions where additional validation may be necessary. These results demonstrate the potential for AI to ~~significantly reduce model development time while maintaining ecological validity~~, offering a practical pathway to expand the application of ecosystem-based management across diverse marine environments.

1 Introduction

Ecosystem modeling is a critical tool for understanding and managing complex marine environments, with Ecopath with Ecosim (EwE) emerging as one of the most widely used frameworks (Christensen and Walters, 2004; Colléter et al., 2015). EwE models

*Corresponding author: scott.spillias@csiro.au

provide quantitative insights into ecosystem structure and function, enabling researchers to assess cumulative impacts of multiple stressors and support ecosystem-based fisheries management (EBFM) decisions (Coll et al., 2015; Villasante et al., 2016). However, constructing these models presents significant challenges, particularly in parameterizing diet matrices that capture the complex web of trophic interactions within an ecosystem.

Traditional approaches to EwE model development rely heavily on extensive field data collection and expert knowledge, which are time-consuming and resource-intensive (Holden et al., 2024a). The process of assembling diet matrices is particularly challenging, requiring synthesis of diverse data sources including field studies, literature reviews, and expert opinion. This creates a significant bottleneck in model development, especially when applying models to new geographical contexts (Holden et al., 2024b).

Recent advances in artificial intelligence (AI) offer new opportunities to streamline the model development process. AI tools have demonstrated success in tasks such as species distribution modeling and remote sensing applications (Lapeyrolerie et al., 2022; Tuia et al., 2022), but their application to process-based ecosystem modeling remains limited (Karniadakis et al., 2021). Recent developments in marine science demonstrate significant potential, with specialized AI systems achieving high accuracy in critical areas such as water quality prediction and biomass estimation (Fernandes and D'Mello, 2024). However, implementation barriers including technical expertise requirements and system reliability concerns must be carefully addressed (Fernandes and D'Mello, 2024). The key challenge lies in ensuring that AI-driven approaches can effectively synthesize available information while maintaining ecological validity.

This study presents a novel framework for assembling and synthesizing local and online resources to parameterize EwE diet matrices using AI. Our approach integrates multiple data sources, including global biodiversity databases, species interaction repositories, and literature-derived information, to automate key steps in model development. Building on recent advances in marine-specific AI applications (Zheng et al., 2023), the framework employs natural language processing and machine learning techniques to understand regional ecosystem characteristics, group species into functional units, and estimate trophic interactions. This approach aligns with emerging work demonstrating how AI systems can effectively analyze and classify marine conservation approaches (Chen and Xu, 2024).

We validate this framework through three case studies representing distinct Australian marine ecosystems: the Northern Territory, South East Inshore, and South East Offshore regions. These regions offer contrasting environmental conditions, species assemblages, and ecological dynamics, providing a robust test of the framework's adaptability and reliability.

The primary objectives of this study are to:

1. Present a systematic, AI-assisted framework for assembling and parameterizing EwE diet matrices
2. Validate key steps in the AI decision-making process, including:
    - Species grouping decisions and their ecological validity
    - Resulting diet matrix values and their reliability
3. Assess the framework's applicability across different marine ecosystems

By rigorously validating this AI-assisted approach across multiple regions, we aim to demonstrate its potential for accelerating ecosystem model development while maintaining scientific rigor. Recent work by Kuhn et al. (2024) emphasizes how machine learning

synthesise information on ?

Let's say Northern Australia as the Gulf of Carpentaria is technically Queensland

applications in fisheries must span multiple scales, from genomics to ecosystem-level analyses, while maintaining interpretability and transparency in automated decision-making systems. This work contributes to the growing need for rapid, data-driven methodologies in ecology (Kelling et al., 2009; Michener and Jones, 2012), while ensuring their outputs align with established ecological principles and can effectively support ecosystem-based management decisions. Our validation approach is informed by emerging standardization efforts in environmental AI applications (Guo et al., 2025), addressing the critical need for rigorous evaluation frameworks in ecological modeling.

2 Methods

2.1 AI-Assisted Framework Overview

The development of ecosystem models requires substantial time organizing species into functional groups and determining their interactions. This framework automates these tasks through integration of artificial intelligence with ecological databases. The framework executes four sequential steps: species identification within a region, biological data collection, functional group organization, and interaction determination (Figure 1).

The framework utilizes the Claude-3.5 large language model (Anthropic, 2024) for data synthesis and ecological analysis. Species identification begins with queries to the Ocean Biodiversity Information System (Grassle and Stocks, 1999) and Global Biodiversity Information Facility (GBIF, 2024). These queries extract occurrence records, temporal distributions, and abundance data within defined geographical boundaries.

The biological data collection phase integrates information from multiple sources. FishBase and SeaLifeBase (Froese et al., 2010) provide life history traits and ecological parameters. The Ecobase repository (Colléter et al., 2015) supplies parameters from existing ecosystem models. Additional data comes from systematic literature searches of regional fisheries reports and peer-reviewed publications, using standardized search terms (e.g., "[species name] AND diet OR feeding OR prey"). Natural language processing extracts relevant ecological information from these documents.

Species grouping employs a vector database (Chroma (Chroma, 2024)) for characteristic storage and retrieval. The database maintains embedding vectors derived from ecological descriptions, life history traits, and habitat preferences. Analysis of these vectors, combined with ecological rules regarding size classes, feeding guilds, and habitat use, determines functional group assignments. Trophic level estimates from FishBase and SeaLifeBase (Froese et al., 2010) validate the ecological coherence of these classifications.

Diet composition analysis merges quantitative records from databases with processed literature data. Food items are extracted from FishBase and SeaLifeBase (Froese et al., 2010) databases, supplemented with interaction data from the Global Biotic Interactions (GLOBI) database (Poelen et al., 2014). Each predator-prey interaction includes source documentation and confidence scores based on data quality metrics.

Diet matrix construction implements weighted averaging for prey proportions. Weights derive from data quality metrics including sample size, study recency, and geographical relevance. In cases of sparse direct diet data, the framework estimates trophic interactions based on similar species' preferences and established ecological principles. Each interaction maintains metadata documenting evidence sources and confidence levels.

Parameter estimation prioritizes values from comparable ecosystems and species groups in Ecobase (Colléter et al., 2015). Empirical relationships from literature provide esti-

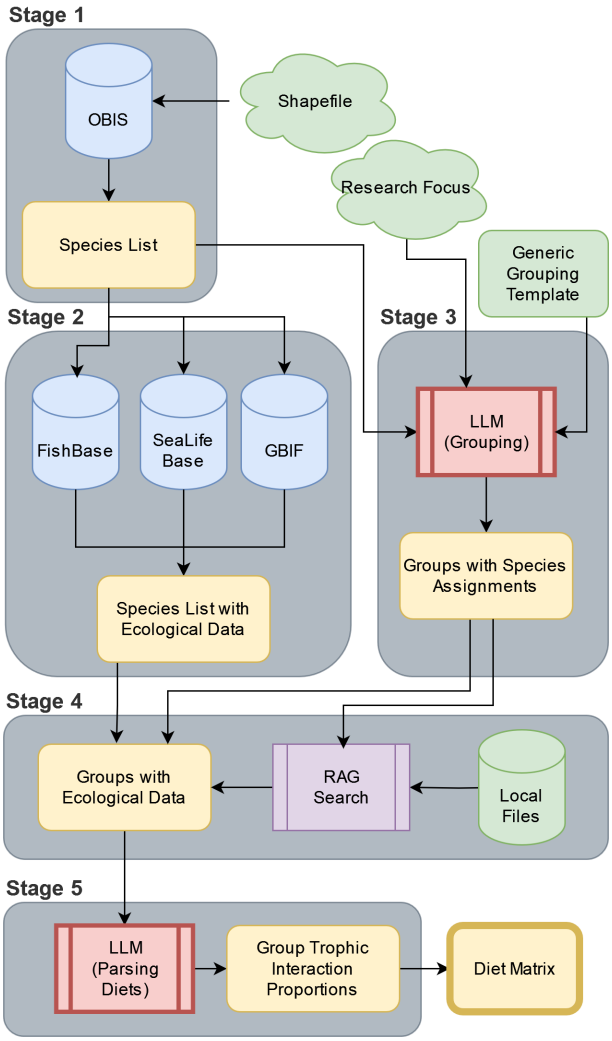

Figure 1: Overview of the AI-assisted framework for ecosystem model development. The process consists of four main steps: species identification, biological data collection, functional group organization, and interaction determination. Highlighted stages (functional group organization and interaction determination) undergo systematic validation across multiple iterations. Each step integrates multiple data sources and analytical approaches.

Does the colouring have a meaning?
Key needed?

mates where direct parameters are unavailable. The framework validates all estimates against biological constraints and ecological theory, identifying anomalies for expert review.

Section S1 of the supplementary material contains detailed documentation of all processing steps, including database queries, literature search criteria, and ecological classification rules. The complete codebase and configuration files reside at [GitHub repository URL].

2.1.1 Species Identification

We begin species identification by accepting a GeoJSON file that defines the study region boundaries. We selected the Ocean Biodiversity Information System (Grassle and Stocks, 1999) as our primary data source due to its extensive marine species coverage and standardized taxonomic classifications. We access OBIS through the `robis` R package (Chamberlain, 2020), which enables automated querying and data processing.

We process the GeoJSON file in two steps to ensure precise geographic filtering. First, we extract a bounding box from the GeoJSON geometry. We then convert this box into a polygon string for database querying. Through the OBIS checklist function, we retrieve scientific names and complete taxonomic classifications from kingdom to species level for all recorded occurrences within these boundaries.

We clean the raw occurrence data through three sequential steps. First, we filter the dataset using OBIS's `is_marine` flag to eliminate terrestrial species that may occur in coastal records. Second, we remove taxonomic redundancy using a rank-based approach that retains only the most specific classification level available. Our algorithm processes taxonomic ranks from most specific (scientific name) to most general (kingdom), keeping only the entry with the highest taxonomic resolution for each organism. Third, we aggregate occurrences by unique species while preserving their complete taxonomic hierarchies, calculating occurrence frequencies for each species.

We store the final species list in a structured CSV file containing verified marine species, their complete taxonomic hierarchies, and occurrence frequencies. This standardized format facilitates efficient data transfer to subsequent framework stages while maintaining provenance information. The complete R implementation, including rank-based filtering algorithms and geographic processing functions, is available in the project repository.

2.1.2 Data Harvesting

Following species identification, we gather ecological and life history information for each identified species through an incremental processing system. We access SeaLifeBase and FishBase (Froese et al., 2010) through their publicly available PARQUET files using DuckDB, which enables efficient querying of large datasets. We selected DuckDB for its ability to handle complex joins and aggregations on PARQUET files directly, without requiring full data loading into memory.

We construct temporary database tables from our species list and execute structured queries to match species across the databases. Our queries join on concatenated genus and species names to ensure accurate matching. We process SeaLifeBase and FishBase data separately to maintain distinct source attribution, as each database may contain different parameters or conflicting values for the same species.

Good thought, though I thought they handled different taxa (fish/ chondricthyans vs all other marine life)

We supplement the base biological data with interaction information from the Global Biotic Interactions (GLOBI) database (Poelen et al., 2014). For each species, we query the GLOBI API using `curl` commands with URL-encoded species names. This approach provides direct access to documented ecological interactions while managing potential network interruptions through a retry mechanism.

Our data cleaning protocol consists of three stages. First, we standardize types by converting numerical values to consistent formats and timestamps to ISO format. Second, we handle null values by removing empty values, 'NA' strings, and null entries while preserving data structure. Third, we track sources by maintaining database origin information for all data points.

We store the processed data incrementally in a JSON document, implementing file locking mechanisms to handle concurrent access. Each species entry contains complete taxonomic hierarchy from the species identification stage, matched records from SeaLifeBase and FishBase, GLOBI interaction data, and source attribution for each data point.

This incremental approach serves two purposes: it enables processing recovery after interruptions and allows parallel processing of different species batches. We maintain a completion check system that verifies the presence of all required data fields before marking a species as fully processed.

The final output consists of a structured JSON file containing standardized ecological and life history information. This format facilitates efficient data transfer to subsequent framework stages while maintaining complete data provenance. Our implementation, including database queries and processing scripts, is available in the project repository.

2.1.3 Species Grouping

We implement species grouping through a flexible system that adapts to regional ecological contexts. The framework supports three approaches for defining functional groups: using templates from existing EcoBase models, generating region-specific groups through AI analysis of geographic characteristics, or applying a predefined template (provided in Section **??** of the supplementary material). For geographic regions, we analyze oceanographic conditions, habitat types, and ecological characteristics to inform group definitions.

We process taxa hierarchically from kingdom to species level, implementing an incremental system with progress tracking to handle large datasets reliably. At each taxonomic level, Claude evaluates taxa against the selected grouping template using the following prompt:

> You are classifying marine organisms into functional groups for an Ecopath with

Ecosim (EwE) model. Functional groups can be individual species or groups of species that perform a similar function in the ecosystem, i.e. have approximately the same growth rates, consumption rates, diets, habitats, and predators. They should be based on species that occupy similar niches, rather than of similar taxonomic groups.

Examine these taxa at the [rank] level and assign each to an ecological functional group.

Rules for assignment:

- If a taxon contains members with different feeding strategies or trophic levels, assign it to 'RESOLVE'
- Examples requiring 'RESOLVE':
  - A phylum containing both filter feeders and predators
  - An order with both herbivores and carnivores
  - A class with species across multiple trophic levels
- If all members of a taxon share similar ecological roles, assign to an appropriate group
- Only consider the adult phase of the organisms, larvae and juveniles will be organized separately
- Only assign a definite group if you are confident ALL members of that taxon belong to that group

Taxa to classify: [List of taxa]

Available ecological groups (name: description): [List of available groups and their descriptions]

Return only a JSON object with taxa as keys and assigned groups as values.

We implement several reliability mechanisms in the classification process:

- Exponential backoff retry system for handling temporary AI service interruptions
- Incremental progress tracking with file locking for concurrent processing
- Validation of group assignments against template definitions
- Separate handling of AI-suggested groups not in the original template

When the research focus indicates areas requiring higher resolution (e.g., commercial fisheries species), we modify the classification process with additional guidance:

Special consideration for research focus: The model's research focus is: [research

focus]

When classifying taxa that are related to this research focus:

- Consider creating more detailed, finer resolution groupings
- Keep species of particular interest as individual functional groups
- For taxa that interact significantly with the focal species/groups, maintain higher resolution groupings
- For other taxa, broader functional groups may be appropriate

We process each taxon through this classification framework, generating structured JSON documents that map species to functional groups. Taxa marked as `RESOLVE` undergo evaluation at finer taxonomic levels until reaching a definitive group assignment or the species level. We maintain complete provenance information, including the source of group definitions and any AI-suggested modifications.

2.1.4 Diet Matrix Construction

We construct diet matrices through a two-stage process designed for efficiency and reliability. The first stage gathers comprehensive diet data using parallel processing and caching mechanisms. We query SeaLifeBase and FishBase (Froese et al., 2010) food items databases through DuckDB, enabling efficient processing of PARQUET files without full memory loading. For each species, we extract food items using specific codes that link to standardized diet categories.

We supplement database records with interaction data from the Global Biotic Interactions (GLOBI) database (Poelen et al., 2014). Our GLOBI processing differentiates between direct observations (`eats`, `preysOn`) and inverse relationships (`eatenBy`, `preyedUponBy`), maintaining separate interaction counts for each type. We further enrich this data through retrieval-augmented generation (RAG) searches of regional literature (detailed in Section S1.2 of the supplementary material), focusing on specific feeding relationships and dietary preferences.

For each functional group, we combine these data sources into a structured profile. Claude (Anthropic, 2024) then analyzes this profile using the following prompt:

Based on the following information about the diet composition of [group], pro-

vide a summary of their diet. Include the prey items and their estimated proportions in the diet.

Available functional groups and their details: [List of groups with descriptions and top species]

Here is the diet data for [group]: [Combined data including RAG search results, compressed food categories, and GLOBI interactions]

Format your response as a list, with each item on a new line in the following format:

Prey Item: Percentage

For example:

```
Small fish: 40%
Zooplankton: 30%
Algae: 20%
Detritus: 10%
```

If exact percentages are not available, estimate percentages based on the information you have been provided. Ensure that all percentages add up to approximately 100%. Consider the RAG search results, compressed food categories, and GLOBI data when creating your summary. Pay special attention to the GLOBI interaction counts, which indicate frequency of observed feeding relationships. Note that some species may feed on juvenile or larval forms of other species, which are often classified in different functional groups than the adults.

We implement an incremental processing system with file locking mechanisms to handle large datasets reliably. The system maintains caches for species-level diet data from databases, combined diet information including literature results, GLOBI interaction networks, and intermediate AI analyses. This caching system enables recovery from interruptions and facilitates parallel processing of different functional groups.

The matrix assembly stage processes the AI's standardized diet descriptions through automated parsing. We convert percentage strings to decimal values and implement a validation system. Our validation ensures that all proportions sum to 1 for each predator, prey items match defined functional groups, and mass-balance requirements are maintained.

When prey items do not exactly match functional group names, we employ a hierarchical matching system. The system first attempts exact matches, then falls back to partial matching using taxonomic information, and logs unmatched items for expert review.

The final output consists of a CSV file containing the complete diet matrix, with predators as rows and prey as columns. Each cell contains the proportion of predator diet comprised by that prey item.

2.1.5 Parameter Estimation (Likely to be removed - no validation)

The final stage estimates the required Ecopath (Christensen and Walters, 2004) parameters for each functional group. We query the EcoBase repository (Colléter et al., 2015) using a structured search strategy that combines group names with regional context (e.g., "Western Australian shelf species in [group]"). For each functional group, we retrieve five key parameters from comparable models: habitat area fractions, biomass densities

Useful if have a gap in another parameter and need to use EE to force estimation of he other parameter in EwE but typically you let this be the unknown to be solved for (as it is in effect a made up property not directly observable)

(t/km^2), production/biomass ratios (P/B, $year^{-1}$), consumption/biomass ratios (Q/B, $year^{-1}$), and ecotrophic efficiency (EE).

We implement parameter-specific validation checks. For biomass densities, we verify values fall within expected ranges for each functional group type. For P/B and Q/B ratios, we check consistency with known allometric relationships and life history characteristics. EE values must fall between 0 and 1, with additional validation against typical ranges for similar species groups.

Claude (Anthropic, 2024) analyzes these parameter sets using standardized criteria. The analysis considers the relevance of each model to the group in question, the consistency of values across different models, trends or patterns in the data that might indicate suitable values, and the ecological context provided by model metadata.

For each parameter assignment, Claude provides a structured response containing the suggested parameter value, detailed reasoning for the selection, references to source models, and any assumptions or caveats.

When EcoBase data is insufficient or parameter values show high variability, we employ a hierarchical fallback system. First, we search for parameters from taxonomically similar groups. Next, we apply empirical relationships where available. Finally, if automated methods are unsuitable, we flag the parameter for manual parameterization.

The system maintains comprehensive logging of parameter assignments. This includes search queries and results, parameter validation checks, AI reasoning and decisions, source model references, and data quality assessments.

The final output consists of a structured JSON file containing the complete parameter set for each functional group. This format preserves the full provenance of each parameter value while enabling automated validation and mass-balance calculations in Ecopath.

2.2 Validation Framework

We executed model generation across three distinct phases. In phase one, we established baseline configurations for each study region by processing species occurrence data and research parameters. We terminated this phase prior to species grouping, creating standardized input states for subsequent validation iterations.

We executed five independent iterations per region in phase two. Each iteration began with species grouping and proceeded through the complete model construction sequence. We maintained fixed input parameters across iterations while allowing the AI's stochastic decision processes to generate natural variation in outputs.

2.2.1 Study Regions

wide shelves, islands, extensive mangroves,

We validate our framework using three Australian marine regions that present distinct ecological characteristics and modeling challenges (Figure 2). The Northern Territory region represents a tropical ecosystem characterized by complex reef systems, seasonal monsoon influences, and high biodiversity. This region tests the framework's ability to handle diverse species assemblages and complex trophic interactions in a dynamic environment.

Australia

The South East Inshore region represents a temperate coastal system with extensive ecological records. This region has comprehensive diet information in established databases, well-documented EwE models spanning multiple years, and active research programs. The rich availability of expert knowledge and historical data makes this region ideal for validating the framework's data integration capabilities.

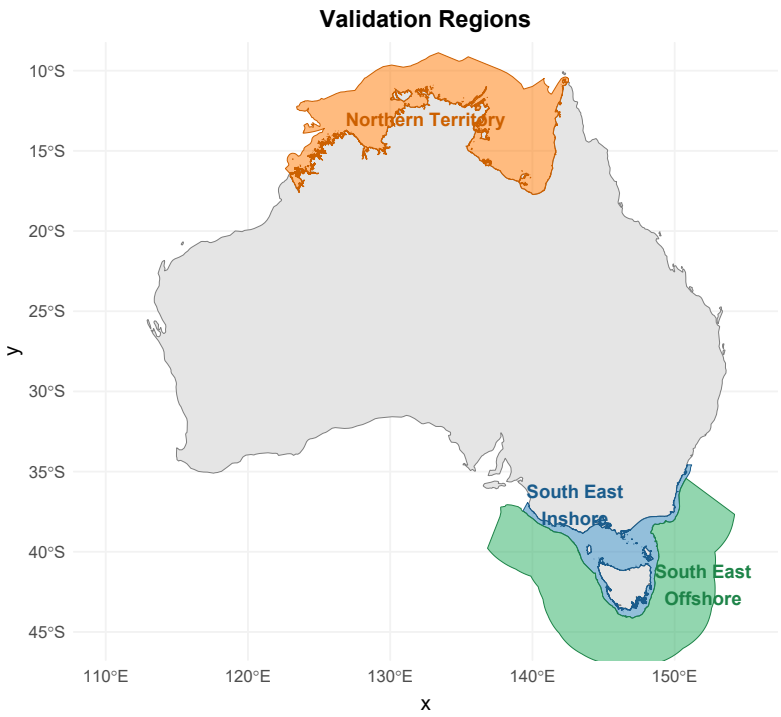

Figure 2: Map of the three validation regions used in this study: Northern Territory (orange), South East Inshore (blue), and South East Offshore (green).

The South East Offshore region presents a deep-water ecosystem that challenges the framework with data-limited conditions and unique ecological patterns. This region tests the framework's capacity to handle situations where direct observational data may be sparse and where species interactions may be less well understood. The contrasting characteristics of these three regions provide a robust test of the framework's adaptability across different ecological contexts.

and where there is the possibility that the model boundary spans a strong ecological break (east/west of Tasmania) in at least some species groups

2.2.2 Group Consistency Analysis

We tracked each species' group assignments across iterations and calculated a consistency score:

$$\text{Consistency Score} = \frac{\text{Number of occurrences in most common group}}{\text{Total number of iterations}}$$

We classified species with consistency scores below 0.95 as unstable, indicating variable group assignments across iterations.

We assessed group stability using the Jaccard similarity coefficient to measure consistency of group membership between consecutive iterations:

$$J(i,j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j|}$$

where M_i and M_j represented species members in iterations i and j. We calculated the overall stability score by averaging Jaccard similarities across consecutive iteration pairs.

Not to state the obvious but is there any test/ check (either in the process or by a human after the fact) that the species has ended up in the right group? ie guarding against being consistently wrong?

2.2.3 Diet Matrix Analysis

We analyzed diet matrix reliability by focusing on significant predator-prey interactions, which we defined as those comprising more than 5% of a predator's diet. For each interaction, we calculated:

1. Presence ratio across iterations:

$$P_{ij} = \frac{\text{Number of iterations with interaction}}{n}$$

where n was the total number of iterations.

2. Mean diet proportion:

$$\mu_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{ijk}$$

where x_{ijk} represents diet proportion for predator i consuming prey j in iteration k.

3. Coefficient of variation:

$$CV_{ij} = \frac{\sigma_{ij}}{\mu_{ij}}$$

where σ_{ij} was the standard deviation of the diet proportion across iterations.

We classified interactions as unstable when their presence ratio fell below 0.95, indicating inconsistent prediction across iterations. For elements where $\mu_{ij} = 0$, we set $CV_{ij} = 0$ to avoid undefined values.

Why this value? And how sensitive is it to this threshold (oils we run it with lower values)? A link that is much less than this in EwE (eg. <0.p1% for instance) can still be a major pressure on a less common group

2.3 Statistical Analysis

We conducted statistical analyses to evaluate framework performance across regions. We assessed group consistency using chi-square tests and coefficients of variation, examining the stability of species assignments across iterations. We analyzed regional differences in group characteristics using one-way ANOVA with Cohen's f effect size calculations. We evaluated trophic level patterns using Kruskal-Wallis tests to identify significant differences in trophic structure. We quantified diet matrix reliability through pairwise Spearman correlations between iterations, focusing on significant interactions (>0.05 proportion).

Also touch on it in discussion - even if passing these tests was considered desirable, would one that failed these tests still be an ok starting point for building a model (as a human we often have to start with really rubbery stuff). More important that you get something" good enough" to be a reliable pie ce of info than have to be perfect and.thus end up with nothing considered suitable

3 Results

3.1 Framework Validation

3.1.1 Processing Scale and Performance

We evaluated our framework through five independent runs across three distinct Australian regions, processing a total of 41,085 species. The framework handled 11,362 species in the Northern Territory's tropical reef ecosystem, 13,901 in the South East Inshore's coastal and pelagic environments, and 15,822 in the South East Offshore's deep-water systems.

Australia

3.1.2 Computational Efficiency

The computational requirements of the AI framework varied across regions. Total processing time ranged from 17.5 to 73.0 hours across regions. The most time-intensive stage was the harvesting of biological data from online databases, accounting for approximately 65% of the total processing time. Species identification typically required 0.01 hours, while the AI-driven species grouping process averaged 0.12 hours. Diet data collection and matrix construction required 13.5 and 0.03 hours respectively, with final parameter estimation taking 0.24 hours. On average, the framework required 6.8 seconds per species for data harvesting and 3.5 seconds per species for diet data collection, though these rates varied considerably between regions due to differences in data availability and species complexity.

Table 1: Computational requirements by region and processing stage

Region	Species Count	Processing Time (hours)					
		Identification	Data Harvest	Grouping	Diet Collection	Matrix Construction	Parameter Estimation
NorthernTerritory	11,362	0.01	9.6	0.1	7.6	0.03	0.1
SouthEastInshore	13,901	0.01	53.8*	0.1	18.7	0.04	0.3
SouthEastOffshore	15,821	0.01	15.3	0.1	14.1	0.03	0.3

*Includes periods of system inactivity between processing batches

Australia

Processing times varied by region and stage (Table 1). Data harvesting required 9.6 hours for the Northern Territory (3.0 seconds per species) and 53.8 hours for the South East Inshore (13.9 seconds per species). Diet data collection took 7.6 hours for the Northern Territory (2.4 seconds per species) and 18.7 hours for the South East Inshore (4.8 seconds per species). Species identification (0.01 hours), grouping (0.1 hours), and parameter estimation (0.1-0.3 hours) remained constant across regions.

3.1.3 Classification Consistency

Needs more explanation earlier I think

The framework reduced 63 potential functional groups to 34-36 region-specific groups. Chi-square tests indicated consistency across regions ($p > 0.85$). Coefficient of variation measurements were: South East Inshore (mean = 0.002, SD = 0.004), Northern Territory (mean = 0.004, SD = 0.012), and South East Offshore (mean = 0.019, SD = 0.052). ANOVA revealed significant regional differences in group sizes ($F = 8279010.7$, $p < 0.001$, Cohen's $f = 2877.3$). Classification stability reached 99% in the Northern Territory (114 inconsistent species), 99.6% in the South East Inshore (52 inconsistent species), and 98.8% in the South East Offshore (191 inconsistent species).

Figure 3 presents the quantitative analysis across regions. Panel A shows median group sizes of 150-200 species. Panel B displays consistency scores ranging from 0.4-1.0 in the Northern Territory and 0.8-1.0 in South East regions. Panel C indicates Jaccard similarity indices of 0.995 in South East Offshore and 0.985 in Northern Territory. Panel D shows group size standard deviations of 5-10 species in South East regions and 15-35 species in Northern Territory.

The stability heatmap (Figure 4) reveals near-perfect stability (Jaccard similarity > 0.99) for most functional groups, with specific exceptions documented in Table 2. Classification inconsistencies followed systematic patterns, primarily occurring between ecologically similar groups. In the Northern Territory, anemones alternated between benthic infaunal carnivores and benthic filter feeder classifications, while flatfishes varied between

benthivores and shallow demersal fish categories. The South East Inshore region showed similar patterns, with *Antigonia* species alternating between planktivore and benthivore classifications.

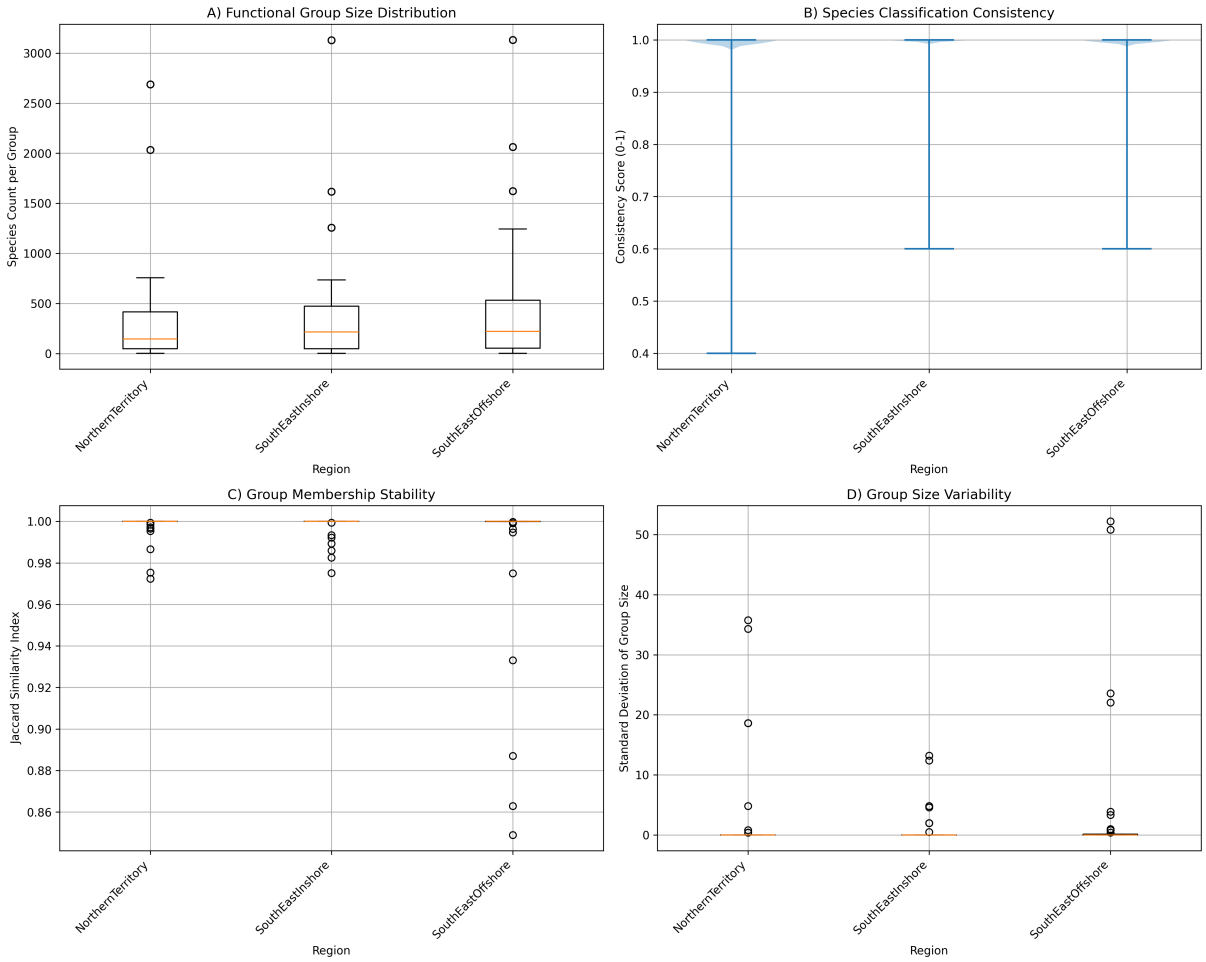

Figure 3: Multi-panel analysis of framework performance across regions. (A) Box plots of functional group sizes (0-3000 species), showing similar median sizes but varying distributions across regions, with outliers indicating some very large groups. (B) Violin plots of species classification consistency (0.4-1.0), where wider sections indicate more species with that consistency score; most species show high consistency (>0.9) with slightly more variation in the Northern Territory. (C) Box plots of group stability measured by Jaccard similarity (0.975-1.000), showing highest stability in South East Offshore and more variable stability in Northern Territory. (D) Box plots of group size variation (standard deviation 0-35), demonstrating larger fluctuations in group membership in the Northern Territory compared to South East regions.

3.1.4 Diet Matrix Validation

The framework demonstrated complex patterns in diet matrix construction. Analysis of significant interactions (>0.05 proportion) revealed moderate negative correlations between iterations in both the Northern Territory (mean r = -0.671, 95% CI [-0.698, -0.645]) and South East Inshore regions (mean r = -0.690, 95% CI [-0.728, -0.651]). Despite these correlations suggesting some variability in fine-scale interactions, the framework

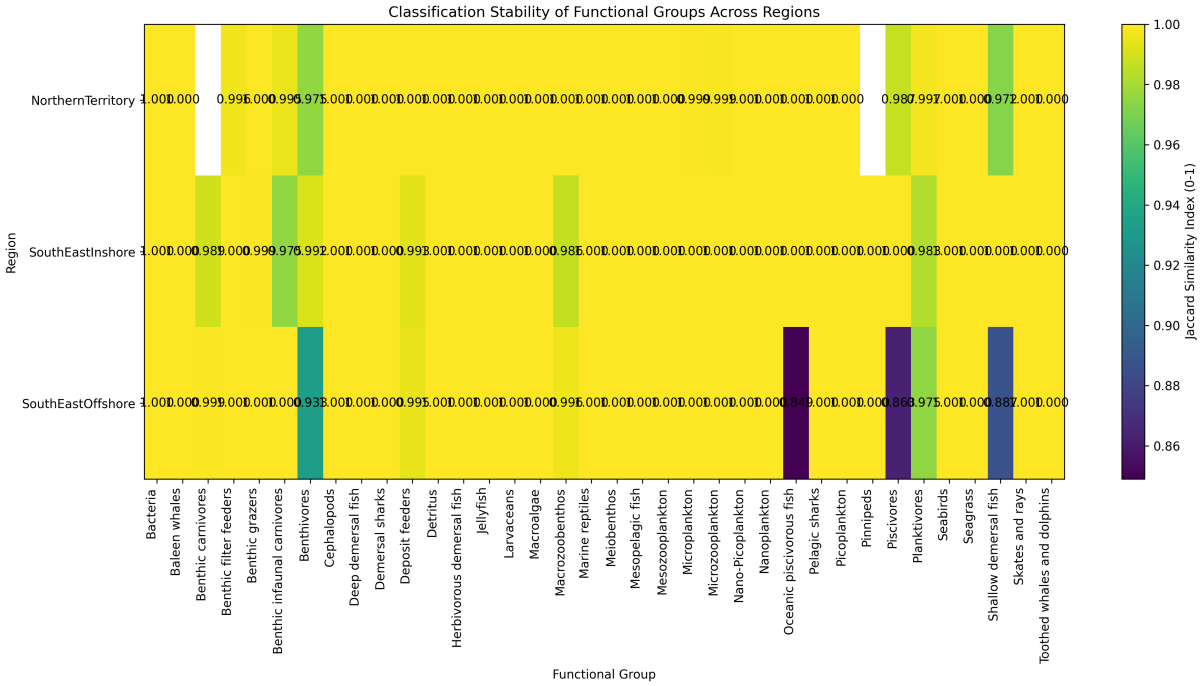

Figure 4: Heatmap showing the stability of functional group classifications across regions. Each cell displays the Jaccard similarity score (ranging from 0.975 to 1.000) between consecutive framework iterations, where 1.000 indicates perfect consistency in species assignments. Darker red colors represent higher stability (scores near 1.000), while lighter colors indicate more variable classifications (scores closer to 0.975). Most functional groups show high stability (>0.99) across all regions, with occasional variations in groups like benthic grazers and deposit feeders, particularly in the Northern Territory region.

Table 2: Species with unstable group assignments across validation iterations. Species are considered unstable if their consistency score is below 0.95, indicating they were assigned to different functional groups in different iterations. The consistency score represents the proportion of iterations where the species was assigned to its most common group.

Region	Species Pattern	Count	Primary Group	Alternative Groups
Northern Territory	Uranoscopidae	12	Shallow demersal fish (40%)	Benthivores (40%), Piscivores (20%)
	Actinodendron & allies	42	Benthic infaunal carnivores (60%)	Benthic filter feeders (40%)
	Platycephalidae	60	Benthivores (60%)	Shallow demersal fish (40%)
South East Inshore	Cirrhilabrus & *Antigonia* spp.	4	Planktivores (60%)	Benthivores (40%)
	Leptostraca	13	Deposit feeders (60%)	Macrozoobenthos (40%)
	Polycladida	35	Benthic infaunal carnivores (80%)	Benthic carnivores (20%)
South East Offshore	Carangidae	34	Oceanic piscivorous fish (60%)	Piscivores (40%)
	Pleuronectiformes	42	Shallow demersal fish (80%)	Benthivores (20%)
	Clinidae	115	Shallow demersal fish (80%)	Benthivores (20%)

maintained consistent ecological patterns. The Northern Territory analysis identified 127 significant interactions with 89.8% stability (13 unstable interactions). The South East Inshore region produced 133 interactions with 91.0% stability (12 unstable interactions), while the South East Offshore achieved 92.5% stability across 147 significant interactions (11 unstable interactions).

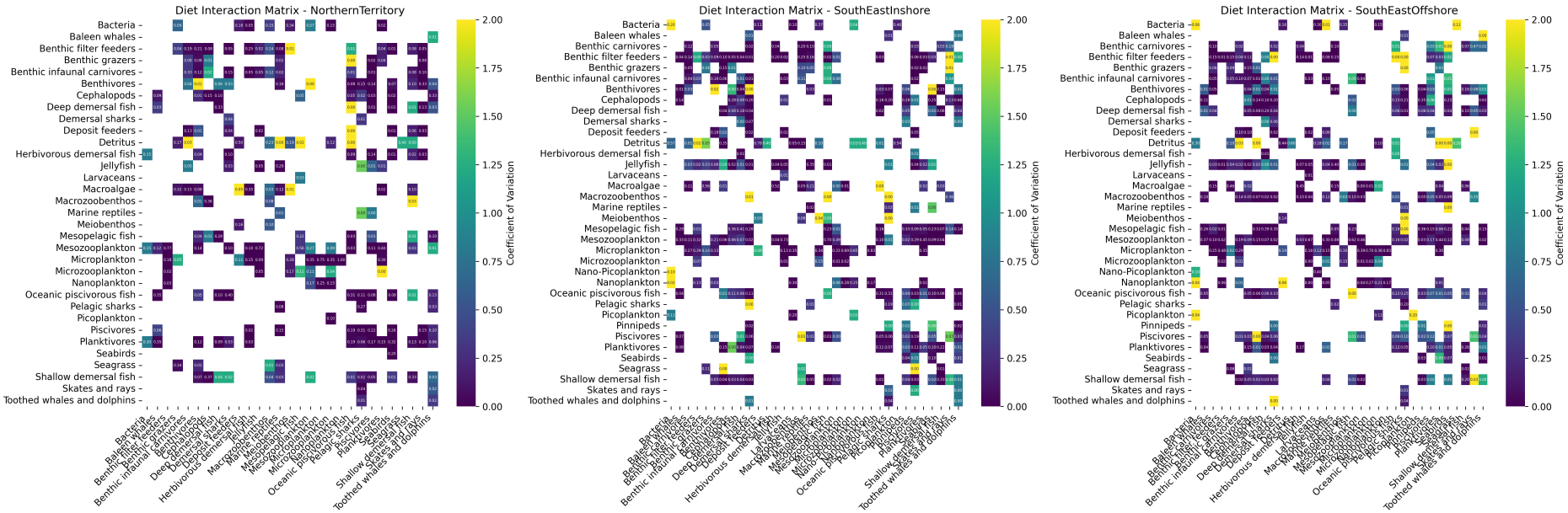

Figure 5: Diet matrix consistency across five iterations for each geographic region. Column names are predator groups and row names are their prey groups. Numbers in each cell indicate the mean diet proportions across five iterations, cell colors indicate the level of variation between iterations, and white cells represent absent feeding relationships.

Plot of diet consistency or magnitude? Caption currently unclear.

3.2 Ecological Findings

3.2.1 Trophic Level Patterns

Analysis of trophic structure revealed significant differences in both the Northern Territory and South East Inshore regions (Kruskal-Wallis H-test, H = 164.0 and 172.0 respectively, $p < 0.001$ for both). Higher trophic level species, particularly apex predators and specialized feeders, maintained consistent classifications across all regions. In contrast, lower trophic levels showed greater variability, particularly among planktonic and benthic invertebrate groups. Benthivores exhibited the highest variation, ranging from 715 to 788 species between runs, while shallow demersal fish showed moderate variation between 586 and 659 species.

4 Discussion

I do think we need to mention the human steps too.

4.1 AI Framework Consistency

Our framework demonstrates robust performance in automating the construction of complex ecosystem models across diverse marine environments. The successful processing of over 41,000 species across three distinct regions validates the framework's scalability and broad applicability. The framework's ability to maintain consistent species classifications while adapting to regional ecological differences suggests it effectively captures fundamental ecological relationships.

The computational efficiency analysis reveals important insights about framework scalability. Data harvesting and diet collection emerge as the primary computational bottlenecks, particularly evident in the South East Inshore region's extended processing times. These bottlenecks likely stem from API rate limitations and the complexity of extracting ecological information from diverse data sources. The relatively constant processing times for species identification, grouping, and parameter estimation across regions indicate these components scale efficiently with increasing species counts.

The framework's classification consistency merits particular attention. The high stability scores (98.8-99.6%) across regions demonstrate reliable species-to-group assignments. The systematic nature of classification inconsistencies provides valuable ecological insights. Species with ambiguous classifications often represent organisms that naturally span multiple ecological niches. For instance, the alternating classifications of anemones between benthic infaunal carnivores and filter feeders reflect their complex feeding strategies. Similarly, the variable classification of flatfishes between benthivore and demersal categories aligns with their known ecological plasticity. These classification patterns suggest the framework captures meaningful ecological uncertainty rather than arbitrary assignment errors.

The diet matrix validation reveals a nuanced picture of trophic relationship stability. The moderate negative correlations between iterations initially appear concerning. However, the high stability of significant feeding interactions (89.8-92.5%) suggests the framework maintains consistent broad-scale trophic structure while allowing flexibility in fine-scale interactions. This pattern aligns with ecological theory, where core trophic relationships remain stable while peripheral feeding interactions may vary with resource availability and environmental conditions.

The observed trophic level patterns provide compelling evidence for the framework's ecological validity. The consistent classification of higher trophic level species reflects the

relatively constrained niches of specialized predators. Conversely, the greater variability in lower trophic level classifications mirrors the natural complexity and adaptability of these groups. The framework's ability to capture this fundamental ecological pattern suggests it successfully incorporates biological realism into its classification decisions.

The regional differences in group size variation and classification stability offer insights into ecosystem complexity. The Northern Territory's higher variation in group sizes and slightly lower classification stability likely reflect the increased ecological complexity of tropical reef systems. The more stable classifications in the South East regions may indicate more clearly defined ecological niches in temperate marine environments. These regional patterns demonstrate the framework's sensitivity to underlying ecological differences while maintaining consistent overall performance.

The reduction from 63 potential functional groups to 34-36 region-specific groups indicates the framework's ability to identify ecologically relevant groupings while avoiding artificial complexity. The statistical consistency across regions suggests these groupings represent fundamental ecological units rather than arbitrary divisions. This optimization of functional group complexity balances model detail with practical utility, a crucial consideration for ecosystem modeling applications.

4.2 Limitations and Uncertainties

Our framework faces several AI-specific limitations identified in recent ecological modeling research. The framework's performance depends on the quality and completeness of available training data. The framework's classification patterns showed regional variations in stability, though further research is needed to determine the relationship between data availability and classification performance. This consideration aligns with findings from Kuhn et al. (2024) regarding machine learning applications in fisheries, where data quality significantly impacts model reliability.

A significant limitation of our approach stems from its reliance on Claude 3.5 Sonnet, a closed-source large language model. The proprietary nature of this model introduces uncertainty regarding the training data used in its development and potential biases that may affect ecological interpretations. While our validation demonstrates consistent performance, the inability to examine the model's training data or internal decision-making processes raises important considerations for scientific reproducibility. Future iterations of the framework may benefit from exploring open-source alternatives or implementing multiple model approaches as demonstrated by Kommineni et al. (2024) in their work with various LLMs for biodiversity research.

The framework's interpretability presents another key challenge. While our validation demonstrates robust performance metrics, the underlying AI decision-making processes, particularly in parameter estimation, require careful scrutiny. This challenge mirrors concerns raised by Fernandes and D'Mello (2024) regarding the "black box" nature of AI systems in aquaculture applications. Our framework partially addresses these concerns through explicit uncertainty quantification and validation metrics, but further work is needed to enhance model transparency.

Technical limitations include computational resource requirements and processing time constraints, particularly evident in data harvesting operations. These limitations align with implementation barriers identified by Fernandes and D'Mello (2024), including acquisition costs and technical expertise requirements. The framework's sensitivity to data availability varies across ecological roles and regions, affecting both classification stability

and diet matrix reliability.

4.3 Applications for EBFM

The framework offers significant practical value for ecosystem-based fisheries management through several key capabilities. Managers can now construct initial EwE models for new regions in days rather than months, enabling faster response to management needs. This addresses a key bottleneck identified by Zheng et al. (2023) in marine resource management. The framework's explicit quantification of uncertainty in species classifications and trophic relationships enables managers to identify areas requiring additional data collection or careful monitoring, following approaches recommended by Kuhn et al. (2024).

The demonstrated adaptability across diverse ecosystems enables customization for specific regions while maintaining methodological consistency, supporting standardized approaches to ecosystem management across jurisdictions. Through systematic processing of available data, the framework also reveals specific areas where additional research or monitoring would most improve model reliability, aligning with recent work by Chen and Xu (2024) on marine protected area management.

4.4 Future Directions

Future development should focus on several key areas to enhance the framework's utility and reliability. Integration with emerging marine AI systems, such as specialized models like MarineGPT (Zheng et al., 2023), could enhance the framework's capabilities in processing region-specific information and visual data. Implementation of standardized evaluation frameworks, such as those proposed in the ELLE dataset (Guo et al., 2025), would enable more rigorous assessment of the framework's performance across different ecological contexts. Development of improved visualization and explanation tools for AI decision-making processes would address current limitations in model transparency and support broader adoption in management contexts.

4.5 Conclusion

Our validation analysis demonstrates both the capabilities and limitations of AI-assisted ecosystem modeling. The framework shows remarkable stability in many aspects while highlighting areas of ecological uncertainty that deserve attention. The clear regional patterns in performance suggest that the approach can adapt to different ecological contexts while maintaining scientific rigor. These findings support the framework's utility for ecosystem-based management while providing clear directions for future improvement.

The observed trade-offs between consistency and complexity reflect fundamental challenges in ecosystem modeling rather than simple methodological limitations. By quantifying these trade-offs and their regional variations, our analysis provides a foundation for more informed application of ecosystem models in fisheries management. As marine ecosystems face increasing pressures from climate change and human activities, this understanding of model behavior and limitations becomes increasingly crucial for effective ecosystem-based management (Geary et al., 2020).

Acknowledgements

[Add acknowledgements here]

Data Availability

The complete codebase, including all scripts, configuration files, and analysis tools, is available at [GitHub repository URL]. The validation framework, including reference group definitions and classification rules, is documented in the project repository to ensure reproducibility.

Author Contributions

SS: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. BF: Validation, Writing - Review & Editing, Supervision, Funding acquisition. FB: Methodology, Software, Validation, Writing - Review & Editing, Supervision. CB: Investigation, Validation, Writing - Review & Editing. RB: Investigation, Validation, Writing - Review & Editing. JPG: Investigation, Validation, Writing - Review & Editing. JS: Conceptualization, Validation, Investigation, Writing - Review & Editing. RS: Investigation, Validation, Writing - Review & Editing. RT: Methodology, Software, Validation, Investigation, Writing - Review & Editing, Supervision, Funding acquisition.

Statement on the Use of Generative AI

Generative AI tools, specifically Claude Sonnet 3.5, were utilized in the preparation of this manuscript to assist with tasks such as language refinement, text structuring, and summarization. All scientific content, data interpretation, and conclusions were independently developed and verified by the authors to ensure accuracy and integrity.

References

Anthropic (2024). Claude 3 model card. Technical documentation, Anthropic.

Chamberlain, S. (2020). *robis: An R Client for the Ocean Biodiversity Information System*. R package version 0.2.0.

Chen, M. and Xu, Z. (2024). A deep learning classification framework for research methods of marine protected area management. *Journal of Environmental Management*.

Christensen, V. and Walters, C. J. (2004). Ecopath with Ecosim: methods, capabilities and limitations. *Ecological Modelling*, 172(2-4):109–139.

Chroma (2024). Chroma - the ai-native open-source embedding database. Accessed: 2024.

Coll, M., Akoglu, E., Arreguín-Sánchez, F., Fulton, E. A., Gascuel, D., Heymans, J. J., Libralato, S., Mackinson, S., Palomera, I., Piroddi, C., et al. (2015). Modelling dynamic ecosystems: venturing beyond boundaries with the Ecopath approach. *Reviews in Fish Biology and Fisheries*, 25(2):413–424.

Colléter, M., Valls, A., Guitton, J., Gascuel, D., Pauly, D., and Christensen, V. (2015). Global overview of the applications of the Ecopath with Ecosim modeling approach using the EcoBase models repository. *Ecological Modelling*, 302:42–53.

Fernandes, S. and D'Mello, A. (2024). Artificial intelligence in the aquaculture industry: Current state, challenges and future directions. *Aquaculture*, page 742048.

Froese, R., Pauly, D., et al. (2010). FishBase.

GBIF (2024). GBIF: The global biodiversity information facility. *Global Biodiversity Information Facility*. What is GBIF?

Geary, W. L., Bode, M., Doherty, T. S., Fulton, E. A., Nimmo, D. G., Tulloch, A. I., Tulloch, V. J., and Ritchie, E. G. (2020). A guide to ecosystem models and their environmental applications. *Nature Ecology & Evolution*, 4:1459–1471.

Grassle, J. F. and Stocks, K. (1999). A global ocean biogeographic information system (OBIS) for the census of marine life. *Oceanography*, 12(3):12–14.

Guo, J., Li, N., and Xu, M. (2025). Environmental large language model evaluation (ELLE) dataset: A benchmark for evaluating generative AI applications in eco-environment domain. *arXiv preprint arXiv:2501.06277*.

Holden, M. H., Akinlotan, M. D., Binley, A. D., Cho, F. H., Helmstedt, K. J., and Chadès, I. (2024a). Why shouldn't I collect more data? Reconciling disagreements between intuition and value of information analyses. *Methods in Ecology and Evolution*, 15:1580–1592.

Holden, M. H., Plagányi, É. E., Fulton, E. A., Campbell, A. B., Janes, R., Lovett, R. A., Wickens, M., Adams, M. P., Botelho, L. L., Dichmont, C. M., et al. (2024b). Cost–benefit analysis of ecosystem modeling to support fisheries management. *Journal of Fish Biology*, 104:1667–1674.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies. *BioScience*, 59(7):613–620.

Kommineni, V. K., König-Ries, B., and Samuel, S. (2024). Harnessing multiple LLMs for information retrieval: A case study on deep learning methodologies in biodiversity publications. *arXiv preprint arXiv:2411.09269*.

Kuhn, B., Cayetano, A., and Fincham, J. (2024). Machine learning applications for fisheries—at scales from genomics to ecosystems. *Reviews in Fisheries Science*.

Lapeyrolerie, M., Chapman, M. S., Norman, K. E., and Boettiger, C. (2022). Deep reinforcement learning for conservation decisions. *Methods in Ecology and Evolution*, 13:2649–2662.

Michener, W. K. and Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, 27(2):85–93.

Poelen, J. H., Simons, J. D., and Mungall, C. J. (2014). Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics*, 24:148–159.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13:1–15.

Villasante, S., Arreguín-Sánchez, F., Heymans, J. J., Libralato, S., Piroddi, C., Christensen, V., and Coll, M. (2016). Modelling complex systems of multiple species for ecosystem based management. *Ecological Modelling*, 326:68–76.

Zheng, Z., Zhang, J., Vu, T. A., Diao, S., Tim, Y. H. W., and Yeung, S. K. (2023). MarineGPT: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*.

Supplementary Material

S1 Technical Implementation

S1.1 Default Grouping with Descriptions

Table S1 presents the complete template of potential functional groups used by the system. This template serves as a reference for group classification, though the system can create new groups or modify existing ones based on specific ecosystem characteristics.

Table S1: Complete Functional Group Template

Group Name	Description
Skates and rays	Bottom-dwelling cartilaginous fish that play a role in controlling benthic prey populations
Nearshore and smaller seabirds	Small gulls, terns etc that feed near shore (possibly include penguins here too) - avian predators that link marine and terrestrial ecosystems
Albatrosses	Large seabirds that forage exclusively at sea, feeding on marine prey (fishes, squids, gelatinous organisms)
Skuas and giant petrels	Large predatory seabirds that feed both at sea and on land, including predation on other birds
Fish-eating pinnipeds	Marine mammals (seals, sea lions) that primarily prey on fish in coastal and pelagic ecosystems
Invertebrate-eating pinnipeds	Marine mammals (particularly Antarctic seals) that primarily feed on krill and other invertebrates
Baleen whales	Large filter-feeding marine mammals that regulate zooplankton populations and contribute to nutrient cycling
Orcas	Apex predators that uniquely prey upon other top predators including marine mammals, sharks, and large fish
Sperm whales	Deep-diving cetaceans that primarily feed on deep-water squid and fish
Small toothed whales and dolphins	Smaller cetaceans that primarily feed on fish and squid in surface and mid-waters
Sea snakes	Marine reptiles that prey primarily on fish, particularly eels and fish eggs
Crocodiles	Large predatory reptiles in coastal and estuarine waters that prey on fish, birds, and mammals
Turtles	Herbivores and omnivores that breed on land
Planktivores	Small fishes that feed on plankton, crucial in transferring energy from plankton to larger predators
Flying fish	Epipelagic fish capable of gliding above the water surface, important prey for many predators
Remoras	Fish that form commensal relationships with larger marine animals, feeding on parasites and food scraps
Large oceanic piscivorous fish	Fish-eating predators in open ocean environments, mid-sized non-migratory species (e.g. barracuda)

Continued on next page

Table S1 – Continued

Group Name	Description
Tuna and Billfish	Large oceanic predatory fish, highly mobile, often dive to feed deeper into the water column
Shelf small benthivores	Small bodied fish that feed on benthic organisms, playing a key role in benthic-pelagic coupling, live in shelf waters
Shelf demersal omnivorous fish	Medium sized demersal fish that feed on invertebrates as well as smaller fish, live in shelf waters
Shelf medium demersal piscivores	Medium sized demersal fish living near the bottom in shallow waters, often important in benthic food webs, feed on other fish primarily, live in shelf waters
Shelf large piscivores	Fish-eating predatory fishes found in various marine habitats, important in controlling prey fish populations
Herbivorous demersal fish	Bottom-associated fish that primarily feed on plants, important in controlling algal growth
Slope/deep water benthivores	Small to mid sized fish that feed on benthic organisms and live on the shelf or seamounts
Slope/deep demersal omnivorous fish	Medium sized demersal fish that feed on invertebrates as well as smaller fish, live in slope or seamount waters
Slope/deep medium demersal piscivores	Medium sized demersal fish that feed on other fish primarily, live in slope or seamount waters
Slope/deep large piscivores	Fish-eating predatory fishes found in various marine habitats in deeper water, live in slope or seamount waters
Migratory mesopelagic fish	Fish living in the mesopelagic zone, undertake diel vertical migration, important in energy transfer between depths
Non-migratory mesopelagic fish	Fish living in the mesopelagic zone, non-migratory species, important in energy transfer between depths
Reef sharks	Top predators in coral reef ecosystems, controlling fish populations and maintaining reef health
Pelagic sharks	Open-ocean predators that help regulate populations of fishes and squids
Demersal sharks	Bottom-dwelling sharks, including dogfishes, that control populations of fishes and invertebrates on and near the seafloor
Cephalopods	Intelligent mollusks like squid and octopus, important predators in many marine ecosystems
Hard corals	Reef-building colonial animals that create complex habitat structure through calcium carbonate deposition
Soft corals	Colonial animals that contribute to reef habitat complexity without building calcium carbonate structures
Sea anemones	Predatory anthozoans that can form symbiotic relationships with fish and crustaceans
Hydrothermal vent communities	Specialized organisms living around deep-sea vents, including chemosynthetic bacteria and associated fauna
Cold seep communities	Organisms adapted to methane and sulfide-rich environments on the seafloor
Deep-sea glass sponges	Filter-feeding animals that create complex deep-water habitats and are important in silicon cycling

Continued on next page

Table S1 – Continued

Group Name	Description
Sea cucumbers	Deposit-feeding echinoderms important in sediment processing and bioturbation
Sea urchins	Herbivorous echinoderms that can control macroalgal abundance and affect reef structure
Crown-of-thorns starfish	Coral-eating sea stars that can significantly impact reef health during population outbreaks
Benthic filter feeders	Bottom-dwelling organisms that filter water for food, important in nutrient cycling and regulating water quality in various depths - bivalves, crinoids, sponges
Macrozoobenthos	Mobile large bottom-dwelling invertebrates in both shallow and deep waters, important in benthic food webs and bioturbation (predatory or omnivorous)
Benthic grazers	Bottom-dwelling organisms that graze on algae and detritus, influencing benthic community structure
Prawns	Small crustaceans that are important in benthic and pelagic food webs
Meiobenthos	Tiny bottom-dwelling organisms, important in sediment processes and as food for larger animals
Deposit feeders	Animals that feed on organic matter in sediments, important in nutrient cycling
Benthic infaunal carnivores	Predatory animals living within the seafloor sediments
Sedimentary Bacteria	Microscopic organisms crucial in nutrient cycling and the microbial loop in marine ecosystems
Large carnivorous zooplankton	Fish larvae, arrow worms and other large predatory zooplankton
Antarctic krill	Key species in Antarctic food webs, particularly important as prey for whales, seals, and seabirds
Ice-associated algae	Microalgae living within and on the underside of sea ice, important primary producers in polar regions
Ice-associated fauna	Specialized invertebrates living in association with sea ice, important in polar food webs
Mesozooplankton	Medium-sized zooplankton (200 µm to 2 cm) that feed on smaller plankton and serve as food for larger animals
Microzooplankton	Tiny zooplankton (20 µm to 200 µm) that graze on phytoplankton and bacteria, forming a crucial link in the microbial food web
Pelagic tunicates	Including larvaceans, salps, and pyrosomes, important in marine snow formation and carbon cycling
Jellyfish	Predatory gelatinous species
Diatoms	Larger phytoplankton (20 µm to 200 µm), silica dependent important primary producers in marine ecosystems
Dinoflagellates	Mixotrophic species (20 µm to 200 µm) that can switch between primary production and consumption as needed
Nanoplankton	Plankton ranging from 2 µm to 20 µm in size, including small algae and protozoans

Continued on next page

Table S1 – Continued

Group Name	Description
Picoplankton	Plankton ranging from 0.2 µm to 2 µm in size, including both photosynthetic and heterotrophic organisms
Microalgae (microphytobenthos)	Microscopic algae that live on the seafloor or attached to other organisms
Pelagic bacteria	Watercolumn dwelling bacteria, consume marine snow amongst other things
Seagrass	Marine flowering plants that form important coastal habitats and nursery areas
Mangroves	Salt-tolerant trees forming critical coastal nursery habitats and protecting shorelines
Salt marsh plants	Coastal vegetation adapted to periodic flooding, important in nutrient cycling and shoreline protection
Macroalgae	Seaweeds of various sizes that provide habitat and food for many species, including both canopy and understory forms
Symbiotic zooxanthellae	Photosynthetic dinoflagellates living within coral and other marine invertebrates
Cleaner fish and shrimp	Species that remove parasites from other marine animals, important in reef health
Discards	Carrion and freshly discarded material from fisheries activities
Detritus	Labile components of natural death and waste

S1.2 Retrieval-Augmented Generation Implementation

We implement a retrieval-augmented generation system using ChromaDB for vector storage and document management. Document processing begins with LlamaParse conversion of source materials to markdown format, preserving structural elements while enabling consistent text extraction across document types. We segment documents using a token-aware chunking strategy with a 2000-token maximum size, determined through empirical testing to balance context preservation with model limitations.

Document processing follows a two-phase approach. The initial phase generates embeddings for each document chunk using Azure OpenAI's text-embedding-3-small model, storing them in ChromaDB's PersistentClient. The system maintains an indexed_files.json registry to track processed documents. The second phase handles incremental updates, identifying and processing only new content when documents are added to the source directory.

For diet composition analysis, we implement a two-stage query process. The first stage employs a simple query to retrieve relevant document chunks:

> What do [group] eat?

The system embeds this query using the same Azure OpenAI model and performs vector similarity search to identify relevant document chunks. These results combine with structured data sources including species occurrence frequencies, food category classifications, and GLOBI interaction data to form a comprehensive input for the second stage.

We implement comprehensive error handling throughout the pipeline. The system employs exponential backoff retry logic for API interactions, with configurable parameters including maximum retries (10), initial delay (1 second), and maximum delay (300 seconds). For model interactions, we utilize LlamaIndex's query engine with zero-temperature sampling to ensure deterministic responses. The system supports multiple language model backends including Claude-3 Sonnet (200k token context), GPT-4, and AWS Claude, enabling flexible deployment based on availability and performance requirements.

The system maintains separate storage contexts for different document collections through ChromaDB's collection management. This separation prevents cross-contamination between knowledge bases while enabling efficient parallel processing. We track document citations throughout the retrieval process, maintaining provenance information for all retrieved content. The complete implementation, including embedding generation, chunking algorithms, and query processing functions, is available in the project repository.