

1      Automated Diet Matrix Construction for Marine  
2      Ecosystem Models Using Generative AI

3                Scott Spillias<sup>1,2\*</sup>      Beth Fulton<sup>1,2</sup>      Fabio Boschetti<sup>2,3</sup>  
4                Cathy Bulman<sup>1</sup>      Rodrigo Bustamante<sup>4</sup>      Javier Porobic Garate<sup>1,2</sup>  
5                Joanna Strzelecki<sup>3</sup>      Roshni Subramaniam<sup>1,2</sup>      Rowan Trebilco<sup>1,2</sup>

6                January 28, 2025

7      <sup>1</sup>CSIRO Environment, Hobart, Australia

8      <sup>2</sup>Centre for Marine Socio-Ecology, University of Tasmania, Hobart, Australia

9      <sup>3</sup>CSIRO Environment, IOMRC Crawley, Australia

10     <sup>4</sup>CSIRO Environment, St. Lucia, Australia

11                **Abstract**

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

32     **1 Introduction**

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

---

\*Corresponding author: scott.spillias@csiro.au

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

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

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

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

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

72 The primary objectives of this study are to:

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

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

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

## 90 2 Methods

### 91 2.1 AI-Assisted Framework Overview

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

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

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

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

115 Diet composition analysis merges  quantitative records from databases with processed  
116 literature data. Food items are extracted  from FishBase and SeaLifeBase (Froese et al.,  
117 2010) databases, supplemented with interaction data from the Global Biotic Interactions  
118 (GLOBI) database (Poelen et al., 2014). Each predator-prey interaction includes source  
119 documentation and confidence scores based on data quality metrics.

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

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

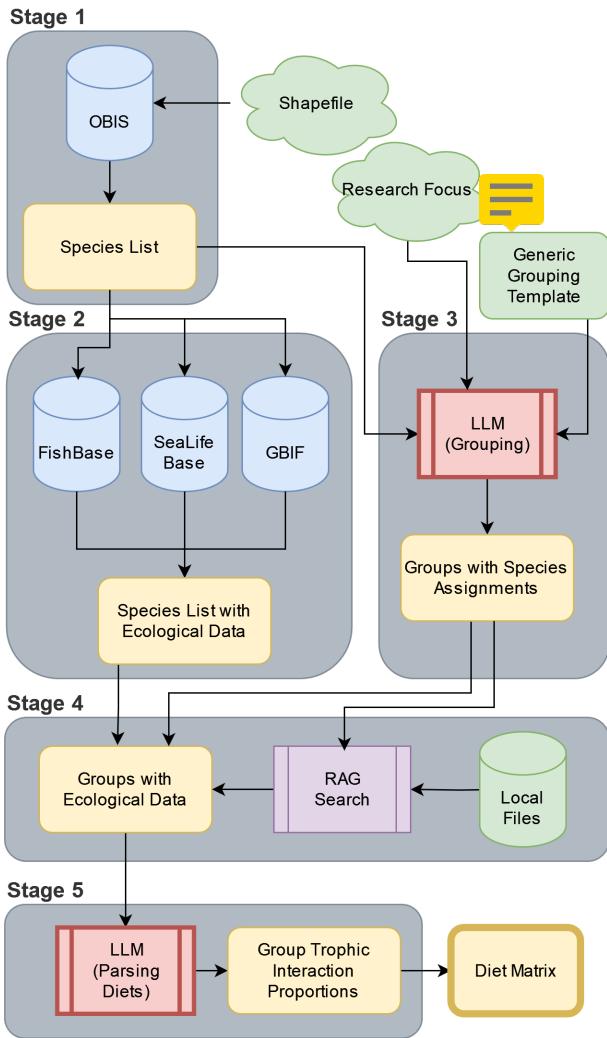


Figure 1: Overview of the AI-assisted framework for ecosystem model development. The process consists of four main steps: species identification, biological data collection, functional group organization, and interaction determination. **Highlighted stages** (functional group organization and interaction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

127 mates where direct parameters are unavailable. The framework validates all estimates  
128 against biological constraints and ecological theory, identifying anomalies for expert re-  
129 view.

130 Section S1 of the supplementary material contains detailed documentation of all pro-  
131 cessing steps, including database queries, literature search criteria, and ecological classifi-  
132 cation rules. The complete codebase and configuration files reside at [GitHub repository  
133 URL].

### 134 2.1.1 Species Identification

135 We begin species identification by accepting a  JSON file that defines the study re-  
136 gion boundaries. We selected the Ocean Biodiversity Information System ([Grassle and](#)  
137 [Stocks, 1999](#)) as our primary data source due to its extensive marine species coverage and  
138 standardized taxonomic classifications. We access OBIS through the `robis` R package  
139 ([Chamberlain, 2020](#)), which enables automated querying and data processing.

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

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

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

### 159 2.1.2 Data Harvesting

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

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

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

176 Our data cleaning protocol consists of three stages. First, we standardize types by con-  
177 verting numerical values to consistent formats and timestamps to ISO format. Second, we  
178 handle null values by removing empty values, ‘NA’ strings, and null entries while preserv-  
179 ing data structure. Third, we track sources by maintaining database origin information  
180 for all data points.

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

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

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

### 194 2.1.3 Species Grouping

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

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

205 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.

206

207 We implement several reliability mechanisms in the classification process:

208

- 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

209

210

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

212

213 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

215

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

#### 221 2.1.4 Diet Matrix Construction

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

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

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

237

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.

238

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

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

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

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

## 255 2.1.5 Parameter Estimation (Likely to be removed - no validation)

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

261 (t/km<sup>2</sup>), production/biomass ratios (P/B, year<sup>-1</sup>), consumption/biomass ratios (Q/B, year<sup>-1</sup>), and ecotrophic efficiency (EE).

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

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

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

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

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

282 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  
283 value while enabling automated validation and mass-balance calculations in Ecopath.

## 285 2.2 Validation Framework

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

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

### 294 2.2.1 Study Regions

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

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

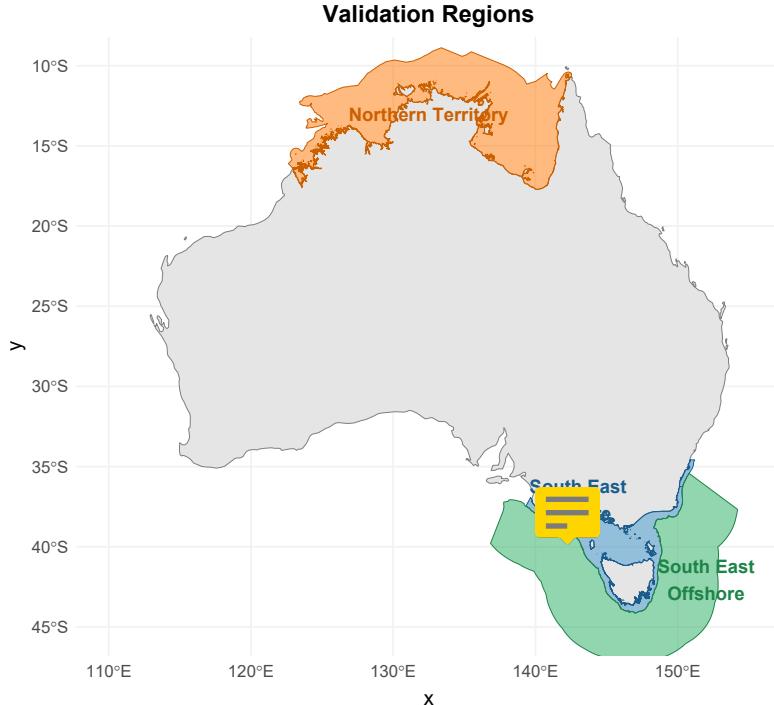


Figure 2: Map of the three validation regions used in this study: Northern Territory (orange), South East Inshore (blue), and South East Offshore (green).

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

### 312    2.2.2 Group Consistency Analysis

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

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

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

317    We assessed group stability using the Jaccard similarity coefficient to measure consis-  
 318 tency of group membership between consecutive iterations:

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

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

322 **2.2.3 Diet Matrix Analysis**

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

- 326 1. Presence ratio across iterations:

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

327 where  $n$  was the total number of iterations.

- 328 2. Mean diet proportion:

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

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

- 330 3. Coefficient of variation:

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

331 where  $\sigma_{ij}$  was the standard deviation of the diet proportion across iterations.

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

335 **2.3 Statistical Analysis**

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

344 **3 Results**

345 **3.1 Framework Validation**

346 **3.1.1 Processing Scale and Performance**

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

### 352 3.1.2 Computational Efficiency

353 The computational requirements of the AI framework varied across regions. Total pro-  
 354 cessing time ranged from 17.5 to 73.0 hours across regions. The most time-intensive stage  
 355 was the harvesting of biological data from online databases, accounting for approximately  
 356 65% of the total processing time. Species identification typically required 0.01 hours,  
 357 while the AI-driven species grouping process averaged 0.12 hours. Diet data collection  
 358 and matrix construction required 13.5 and 0.03 hours respectively, with final parameter  
 359 estimation taking 0.24 hours. On average, the framework required 6.8 seconds per species  
 360 for data harvesting and 3.5 seconds per species for diet data collection, though these rates  
 361 varied considerably between regions due to differences in data availability and species  
 362 complexity.

Table 1: Computational requirements by region and processing stage

Region	Species Count	Processing Time (hours)					Parameter Estimation
		Identification	Data Harvest	Grouping	Diet Collection	Matrix Construction	
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

362 Processing times varied by region and stage (Table 1). Data harvesting required 9.6  
 363 hours for the Northern Territory (3.0 seconds per species) and 53.8 hours for the South  
 364 East Inshore (13.9 seconds per species). Diet data collection took 7.6 hours for the  
 365 Northern Territory (2.4 seconds per species) and 18.7 hours for the South East Inshore  
 366 (4.8 seconds per species). Species identification (0.01 hours), grouping (0.1 hours), and  
 367 parameter estimation (0.1-0.3 hours) remained constant across regions.  
 368

### 369 3.1.3 Classification Consistency

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

378 Figure 3 presents the quantitative analysis across regions. Panel A shows median  
 379 group sizes of 150-200 species. Panel B displays consistency scores ranging from 0.4-1  
 380 in the Northern Territory and 0.8-1.0 in South East regions. Panel C indicates Jaccard  
 381 similarity indices of 0.995 in South East Offshore and 0.985 in Northern Territory. Panel  
 382 D shows group size standard deviations of 5-10 species in South East regions and 15-35  
 383 species in Northern Territory.

384 The stability heatmap (Figure 4) reveals near-perfect stability (Jaccard similarity  
 385 0.99) for most functional groups, with specific exceptions documented in Table 2. Clas-  
 386 sification inconsistencies followed systematic patterns, primarily occurring between eco-  
 387 logically similar groups. In the Northern Territory, anemones alternated between benthic  
 388 infaunal carnivores and benthic filter feeder classifications, while flatfishes varied between

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

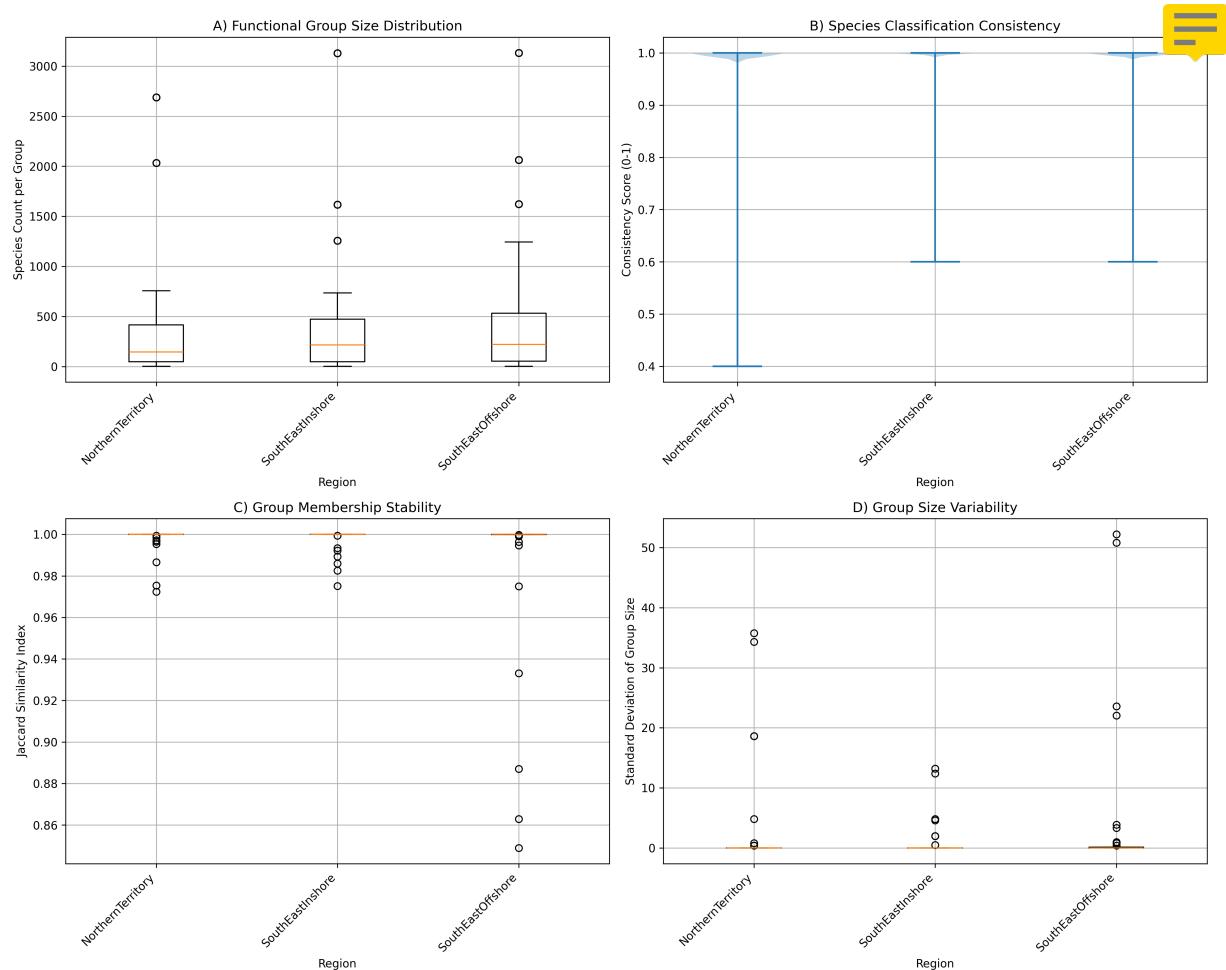


Figure 3: Multi-panel analysis of framework performance across regions. (A) Box plots of functional group sizes (0-3000 species), showing similar median sizes but varying distributions across regions, with outliers indicating some very large groups. (B) Violin plots of species classification consistency (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 (0.975-1.00), 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.

### 392 3.1.4 Diet Matrix Validation

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

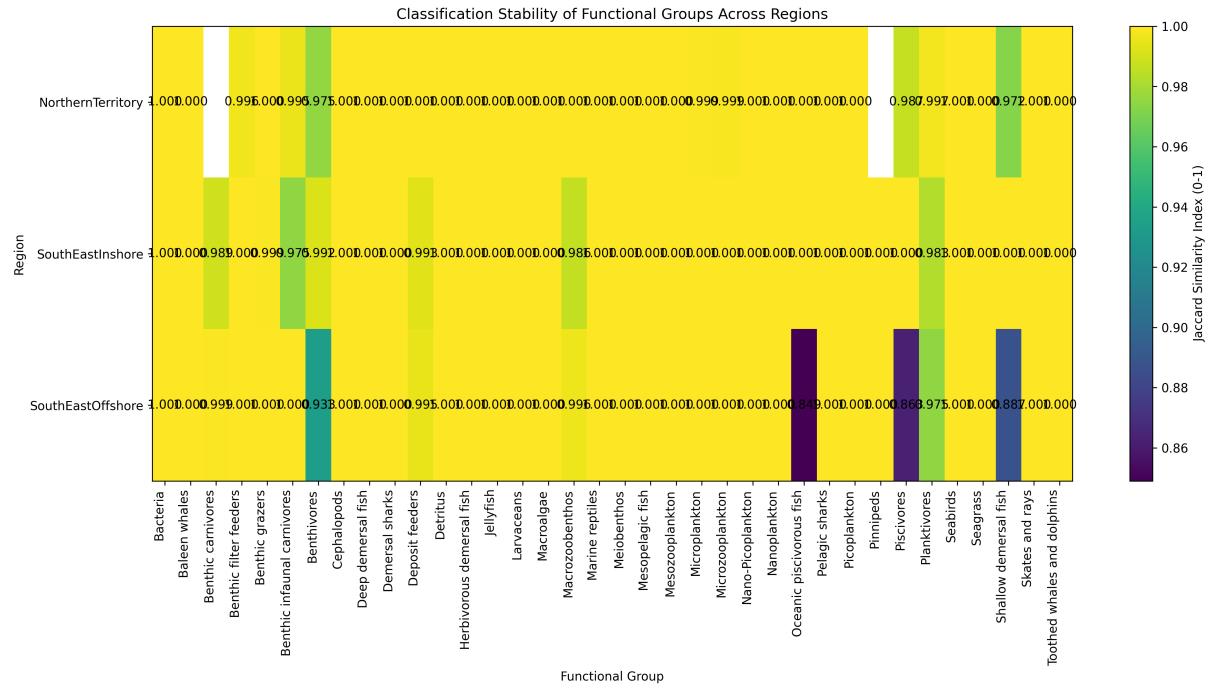


Figure 4: Heatmap showing the stability of functional group classifications across regions. Each cell displays the Jaccard similarity score (ranging from 0.975 to 1.000) between consecutive framework iterations, where 1.000 indicates perfect consistency in species assignments. Darker red colors represent 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	Cirrhitabrus & <i>Antigonia</i> 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%)

<sup>398</sup> maintained consistent ecological patterns. The Northern Territory analysis identified 127  
<sup>399</sup> significant interactions with 89.8% stability (13 unstable interactions). The South East  
<sup>400</sup> Inshore region produced 133 interactions with 91.0% stability (12 unstable interactions),  
<sup>401</sup> while the South East Offshore achieved 92.5% stability across 147 significant interactions  
<sup>402</sup> (11 unstable interactions).

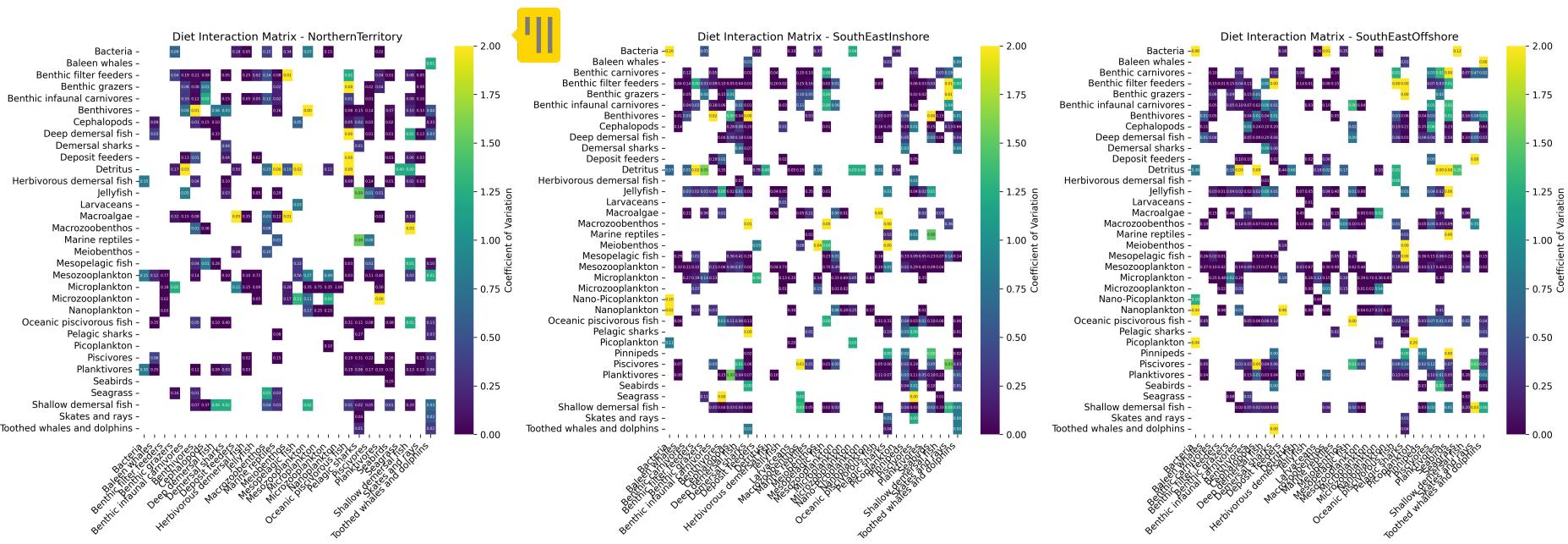


Figure 5: Diet matrix consistency across five iterations for each geographic region. Column names are predator groups and row names are their prey groups. Numbers in each cell indicate the mean diet proportions across five iterations, cell colors indicate the level of variation between iterations, and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

403 **3.2 Ecological Findings**

404 **3.2.1 Trophic Level Patterns**

405 Analysis of trophic structure revealed significant differences in both the Northern Territory  
406 and South East Inshore regions (Kruskal-Wallis H-test,  $H = 164.0$  and  $172.0$  respectively,  $p$   
407  $< 0.001$  for both). Higher trophic level species, particularly apex predators and specialized  
408 feeders, maintained consistent classifications across all regions. In contrast, lower trophic  
409 levels showed greater variability, particularly among planktonic and benthic invertebrate  
410 groups. **Benthivores exhibited the highest variation, ranging from 715 to 788 species**  
411 **between runs, while shallow demersal fish showed moderate variation between 586 and**  
412 **659 species.**

413 **4 Discussion**



414 **4.1 AI Framework Consistency**

415 Our framework demonstrates robust performance in automating the construction of com-  
416 plex ecosystem models across diverse marine environments. The successful processing  
417 of over 41,000 species across three distinct regions validates the framework's scalability  
418 and broad applicability. The framework's ability to maintain consistent species classifi-  
419 cations while adapting to regional ecological differences suggests it effectively captures  
420 fundamental ecological relationships.

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

428 The framework's classification consistency merits particular attention. The high stabil-  
429 ity scores (98.8-99.6%) across regions demonstrate reliable species-to-group assignments.  
430 The systematic nature of classification inconsistencies provides valuable ecological in-  
431 sights. Species with ambiguous classifications often represent organisms that naturally  
432 span multiple ecological niches. For instance, the alternating classifications of anemones  
433 between benthic infaunal carnivores and filter feeders reflect their complex feeding strate-  
434 gies. Similarly, the variable classification of flatfishes between benthivore and demer-  
435 sal categories aligns with their known ecological plasticity. These classification patterns  
436 suggest the framework captures meaningful ecological uncertainty rather than arbitrary  
437 assignment errors.

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

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

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

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

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

## 464 4.2 Limitations and Uncertainties

465 Our framework faces several AI-specific limitations identified in recent ecological mod-  
466eling research. The framework's performance depends on the quality and completeness  
467 of available training data. The framework's classification patterns showed regional varia-  
468 tions in stability, though further research is needed to determine the relationship between  
469 data availability and classification performance. This consideration aligns with findings  
470 from [Kuhn et al. \(2024\)](#) regarding machine learning applications in fisheries, where data  
471 quality significantly impacts model reliability.

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

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

488 Technical limitations include computational resource requirements and processing time  
489 constraints, particularly evident in data harvesting operations. These limitations align  
490 with implementation barriers identified by [Fernandes and D'Mello \(2024\)](#), including ac-  
491 quisition costs and technical expertise requirements. The framework's sensitivity to data  
492 availability varies across ecological roles and regions, affecting both classification stability

493 and diet matrix reliability.

#### 494 4.3 Applications for EBFM

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

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

#### 508 4.4 Future Directions

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

#### 518 4.5 Conclusion

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

525 The observed trade-offs between consistency and complexity reflect fundamental chal-  
526 lenges in ecosystem modeling rather than simple methodological limitations. By quan-  
527 tifying these trade-offs and their regional variations, our analysis provides a foundation  
528 for more informed application of ecosystem models in fisheries management. As marine  
529 ecosystems face increasing pressures from climate change and human activities, this un-  
530 derstanding of model behavior and limitations becomes increasingly crucial for effective  
531 ecosystem-based management ([Geary et al., 2020](#)).

## **532 Acknowledgements**

**533 [Add acknowledgements here]**

## **534 Data Availability**

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

## **539 Author Contributions**

**540** SS: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation,  
**541** Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visual-  
**542** ization, Project administration. BF: Validation, Writing - Review & Editing, Supervision,  
**543** Funding acquisition. FB: Methodology, Software, Validation, Writing - Review & Editing,  
**544** Supervision. CB: Investigation, Validation, Writing - Review & Editing. RB: Investiga-  
**545** tion, Validation, Writing - Review & Editing. JPG: Investigation, Validation, Writing -  
**546** Review & Editing. JS: Conceptualization, Validation, Investigation, Writing - Review &  
**547** Editing. RS: Investigation, Validation, Writing - Review & Editing. RT: Methodology,  
**548** Software, Validation, Investigation, Writing - Review & Editing, Supervision, Funding  
**549** acquisition.

## **550 Statement on the Use of Generative AI**

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

## **555 References**

- 556** Anthropic (2024). Claude 3 model card. Technical documentation, Anthropic.
- 557** Chamberlain, S. (2020). *robis: An R Client for the Ocean Biodiversity Information Sys-*  
**558** *tem*. R package version 0.2.0.
- 559** Chen, M. and Xu, Z. (2024). A deep learning classification framework for research methods  
**560** of marine protected area management. *Journal of Environmental Management*.
- 561** Christensen, V. and Walters, C. J. (2004). Ecopath with Ecosim: methods, capabilities  
**562** and limitations. *Ecological Modelling*, 172(2-4):109–139.
- 563** Chroma (2024). Chroma - the ai-native open-source embedding database. Accessed: 2024.

- 564 Coll, M., Akoglu, E., Arreguín-Sánchez, F., Fulton, E. A., Gascuel, D., Heymans, J. J.,  
565 Libralato, S., Mackinson, S., Palomera, I., Piroddi, C., et al. (2015). Modelling dynamic  
566 ecosystems: venturing beyond boundaries with the Ecopath approach. *Reviews in Fish  
567 Biology and Fisheries*, 25(2):413–424.
- 568 Colléter, M., Valls, A., Guittot, J., Gascuel, D., Pauly, D., and Christensen, V. (2015).  
569 Global overview of the applications of the Ecopath with Ecosim modeling approach  
570 using the EcoBase models repository. *Ecological Modelling*, 302:42–53.
- 571 Fernandes, S. and D'Mello, A. (2024). Artificial intelligence in the aquaculture industry:  
572 Current state, challenges and future directions. *Aquaculture*, page 742048.
- 573 Froese, R., Pauly, D., et al. (2010). FishBase.
- 574 GBIF (2024). GBIF: The global biodiversity information facility. *Global Biodiversity  
575 Information Facility*. What is GBIF?
- 576 Geary, W. L., Bode, M., Doherty, T. S., Fulton, E. A., Nimmo, D. G., Tulloch, A. I.,  
577 Tulloch, V. J., and Ritchie, E. G. (2020). A guide to ecosystem models and their  
578 environmental applications. *Nature Ecology & Evolution*, 4:1459–1471.
- 579 Grassle, J. F. and Stocks, K. (1999). A global ocean biogeographic information system  
580 (OBIS) for the census of marine life. *Oceanography*, 12(3):12–14.
- 581 Guo, J., Li, N., and Xu, M. (2025). Environmental large language model evalua-  
582 tion (ELLE) dataset: A benchmark for evaluating generative AI applications in eco-  
583 environment domain. *arXiv preprint arXiv:2501.06277*.
- 584 Holden, M. H., Akinlotan, M. D., Binley, A. D., Cho, F. H., Helmstedt, K. J., and  
585 Chadès, I. (2024a). Why shouldn't I collect more data? Reconciling disagreements  
586 between intuition and value of information analyses. *Methods in Ecology and Evolution*,  
587 15:1580–1592.
- 588 Holden, M. H., Plagányi, É. E., Fulton, E. A., Campbell, A. B., Janes, R., Lovett, R. A.,  
589 Wickens, M., Adams, M. P., Botelho, L. L., Dichmont, C. M., et al. (2024b). Cost-  
590 benefit analysis of ecosystem modeling to support fisheries management. *Journal of  
591 Fish Biology*, 104:1667–1674.
- 592 Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.  
593 (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.
- 594 Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., and  
595 Hooker, G. (2009). Data-intensive science: a new paradigm for biodiversity studies.  
596 *BioScience*, 59(7):613–620.
- 597 Kommineni, V. K., König-Ries, B., and Samuel, S. (2024). Harnessing multiple LLMs  
598 for information retrieval: A case study on deep learning methodologies in biodiversity  
599 publications. *arXiv preprint arXiv:2411.09269*.
- 600 Kuhn, B., Cayetano, A., and Fincham, J. (2024). Machine learning applications for  
601 fisheries—at scales from genomics to ecosystems. *Reviews in Fisheries Science*.

- 602 Lapeyrolerie, M., Chapman, M. S., Norman, K. E., and Boettiger, C. (2022). Deep  
603 reinforcement learning for conservation decisions. *Methods in Ecology and Evolution*,  
604 13:2649–2662.
- 605 Michener, W. K. and Jones, M. B. (2012). Ecoinformatics: supporting ecology as a  
606 data-intensive science. *Trends in Ecology & Evolution*, 27(2):85–93.
- 607 Poelen, J. H., Simons, J. D., and Mungall, C. J. (2014). Global biotic interactions:  
608 An open infrastructure to share and analyze species-interaction datasets. *Ecological  
609 Informatics*, 24:148–159.
- 610 Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A.,  
611 Mathis, M. W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in  
612 machine learning for wildlife conservation. *Nature Communications*, 13:1–15.
- 613 Villasante, S., Arreguín-Sánchez, F., Heymans, J. J., Libralato, S., Piroddi, C., Chris-  
614 tensen, V., and Coll, M. (2016). Modelling complex systems of multiple species for  
615 ecosystem based management. *Ecological Modelling*, 326:68–76.
- 616 Zheng, Z., Zhang, J., Vu, T. A., Diao, S., Tim, Y. H. W., and Yeung, S. K. (2023).  
617 MarineGPT: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*.

<sub>618</sub> **Supplementary Material**

<sub>619</sub> **S1 Technical Implementation**

<sub>620</sub> **S1.1 Default Grouping with Descriptions**

<sub>621</sub> Table S1 presents the complete template of potential functional groups used by the system.  
<sub>622</sub> This template serves as a reference for group classification, though the system can create  
<sub>623</sub> new groups or modify existing ones based on specific ecosystem characteristics.

Table S1: Complete Functional Group Template

<b>Group Name</b>	<b>Description</b>
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 (microphyto-benthos)	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	Carrión and freshly discarded material from fisheries activities
Detritus	Labile components of natural death and waste

## 624 S1.2 Retrieval-Augmented Generation Implementation

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

631 Document processing follows a two-phase approach. The initial phase generates em-  
 632 beddings for each document chunk using Azure OpenAI’s text-embedding-3-small model,  
 633 storing them in ChromaDB’s PersistentClient. The system maintains an indexed\_files.json  
 634 registry to track processed documents. The second phase handles incremental updates,  
 635 identifying and processing only new content when documents are added to the source  
 636 directory.

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

What do [group] eat?

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

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

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