

The choice of species representation in multi-species aquatic acute-toxicity prediction is constrained by data availability rather than by predictive accuracy

Seonghan Kim

Sangmyung University

Kyungmin Jeon

Sangmyung University

Hyunseo Cho

Sangmyung University

Seongyun Park

Sangmyung University

Daesik Jeong

`jungsoft97@smu.ac.kr`

Sangmyung University

Research Article

Keywords: ecotoxicology, aquatic toxicity, species representation, equivalence testing, pre-registration, data availability, graph neural networks

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10915713/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

1 The choice of species representation in multi-species
2 aquatic acute-toxicity prediction is constrained by
3 data availability rather than by predictive accuracy

4 Seonghan Kim¹, Kyungmin Jeon², Hyunseo Cho²,
5 Seongyun Park², Daesik Jeong^{2*}

6 ¹Department of Electronic Engineering, Sangmyung University, Seoul,
7 Republic of Korea.

8 ²Department of Computer Science, Sangmyung University, Seoul,
9 Republic of Korea.

10 *Corresponding author(s). E-mail(s): jungsoft97@smu.ac.kr;

11 Contributing authors: lucidus814@gmail.com;

12 wjsrudals0910@gmail.com; johyunseo020416@gmail.com;

13 qkrtjsrbs315@gmail.com;

14 **Abstract**

15 Multi-species models of aquatic toxicity must represent the organism as well
16 as the chemical, and that representation is normally chosen by predictive accu-
17 racy. We compared seven representations in two families – five identifier-based
18 (no species information, a species mean offset, out-of-fold residual calibra-
19 tion, one-hot identity, and a learned embedding) and two externally structured
20 (source-authority and NCBI taxonomy) – on 70,670 ADORE records covering
21 1,267 species, under a written pre-registration. The species-using representations
22 were accompanied by permuted, zeroed, dummy, and capacity-matched controls;
23 differences were judged by two one-sided tests against a margin measuring the
24 models’ own re-run noise. Two graph neural networks and two non-neural base-
25 lines were used, on the ADORE benchmark and the ECOTOX-external set,
26 drawn from a later ECOTOX release with records disjoint from the first. Species
27 information improved prediction in every dataset, backbone, and split we exam-
28 ined, lowering the root-mean-square error by 0.07–0.20 $\log_{10}$ units. Choosing

among the informative representations did not reliably improve prediction further: in nine of twelve comparisons against one-hot identity, the difference was smaller than the dataset-specific re-run noise, although estimated too imprecisely to establish equivalence, and the two differences that excluded zero pointed in opposite directions, only one of them surviving multiplicity correction. Across the full ECOTOX aquatic species pool, phylogenetic distances exist for 17.7% of species and energy-budget traits for 7.6%, the latter absent for algae entirely; both nonetheless cover about half of all records. In the ADORE public subset the two resources cover 100% and 77% of retained species, against 30% and 8% of those dropped.

Scientific Contribution In the studies we surveyed, species representations were compared by predictive accuracy alone; none reported a control that permuted the species assignment and refitted the model, and none tested for equivalence against a pre-specified margin. We supply both, under pre-registration and on two datasets, and find that accuracy does not identify a representation. We therefore treat the choice as one of data availability and acquisition cost, and measure that availability across the full ECOTOX aquatic species pool (3,268 species) rather than on a curated subset.

Keywords: ecotoxicology, aquatic toxicity, species representation, equivalence testing, pre-registration, data availability, graph neural networks

1 Introduction

Predicting the acute toxicity of chemicals to aquatic organisms increasingly relies on models that pool data across multiple species, and each model must decide how to represent the organism. The options range from a bare categorical label to a position in a taxonomy, a point on a phylogeny, or a vector of physiological traits, and they differ mainly in the external information required per species [1–4]. The ADORE benchmark was assembled to support the evaluation of multi-species toxicity models by distributing phylogenetic distances, species traits, and predefined splits alongside toxicity records [5].

Across this literature, the species representation is almost always chosen on the basis of predictive accuracy alone. This leaves three gaps. First, comparisons rarely include a control that retains the species block but destroys its assignment; an apparent gain from a richer representation therefore cannot be separated from the benefit

of simply having a species-indexed parameter of the same dimension [6]. Second, differences are tested for superiority but not for equivalence; therefore, “no significant difference” is not distinguished from “the representations are interchangeable” [7, 8]; a practical-equivalence criterion has only recently entered molecular-representation benchmarking [9]. Third, the auxiliary data required by richer representations are treated as a precondition rather than measured; the fraction of species for which a phylogeny or a trait vector actually exists is rarely reported. It has been argued in an adjacent field that complex representations should be held to a deliberately simple baseline they do not always beat [10].

We compared seven species representations under an identical training procedure, repeated across two datasets, four backbones, ten seeds, and four split schemes, with every criterion that determines a verdict fixed in a written specification before any comparison was computed. The representations divide into two families by what each requires per species: five are identifier-based, needing nothing beyond the species label – no species information, a species mean offset, out-of-fold residual calibration, one-hot identity, and a learned embedding – and two are externally structured, requiring a taxonomic record: source-authority and NCBI taxonomy. Each species-using representation was trained alongside permuted, zeroed, dummy, and fixed-projection controls, which separate the contribution of the species assignment from that of the parameters the representation introduces. Differences were judged by two one-sided tests against a margin derived per dataset from the models’ own re-run noise, with hierarchical gatekeeping and Benjamini–Hochberg control within families. The whole procedure was carried out on two datasets: the ADORE benchmark and the ECOTOX-external set, drawn from a later ECOTOX release with records disjoint from the first.

Species information helps: adding it to a no-species model lowers error in every dataset, backbone, and split we examined. But the performance evidence does not

88 identify which representation to use. Most differences between the informative repre-
89 sentations are smaller than the re-run noise, and the few that reach significance do
90 not agree across datasets or evaluation subsets. The equivalence tests did not establish
91 that the representations were interchangeable. We instead measure what does con-
92 strain the choice. At the scale of the full ECOTOX aquatic species pool, phylogenetic
93 distances and energy-budget traits are missing for the large majority of species – the
94 latter entirely for algae – and, where they do exist, are concentrated in the species
95 already richest in data, whereas taxonomy is available for every species. The choice of
96 species representation is therefore governed by the availability and acquisition cost of
97 the data each representation needs, not by predictive performance.

98 2 Related work

99 2.1 How species enters a multi-species toxicity model

100 Multi-species models of acute aquatic toxicity differ less in their chemistry module
101 than in how the species is supplied to the model. Four approaches are commonly
102 used, differing mainly in the external information each requires per species. *(i) Cat-*
103 *egorical identity.* The species is a label. Viljanen et al. encoded 1,506 species as a
104 one-hot indicator and learned interactions with a factorization machine [4]; Posthuma
105 et al. used the same construction for 1,267 species with species, chemical, and dura-
106 tion intercepts [11]. Nothing beyond the species label is needed. *(ii) Taxonomic rank.*
107 The species is replaced or supplemented by its position in a classification. Sheffield
108 and Judson replaced species dummy variables with taxonomic-group variables, cit-
109 ing generalisation to untested species as the motivation [1]; Wu et al. supplied class,
110 order, family, genus, and species as categorical features [12]; Schlender et al. supplied
111 species, class, and phylum as categorical features to a single pooled model over 351
112 species [13]. The required external information is membership in a taxonomy, which
113 is recorded for every species in the datasets examined here. *(iii) Externally derived*

114 *continuous descriptors*. The species is a vector computed outside the toxicity data:
115 dynamic energy budget (DEB) parameters [2, 14, 15], phylogenetic eigenvector maps
116 from sequence data [3, 16, 17], or mitochondrial sequence embeddings alongside tax-
117 onomy [18]. The required external information is substantial and species-specific. (*iv*)
118 *Task structure*. The species indexes tasks – separate heads or meta-learned adaptation
119 [13] – so the architecture, not the input, carries the same information as (i).

120 2.2 Prior results on the same data

121 Three studies have used the same family of datasets. The ADORE benchmark assem-
122 bles fish, crustacean, and algae acute mortality records from ECOTOX and distributes
123 phylogenetic distances, species traits, chemical properties, and predefined splits; it
124 reports no models [5]. Posthuma et al. drew 70,670 records over 3,295 chemicals and
125 1,267 species from the same intermediate processed table that we used, rather than
126 the reduced public modelling set. Their nested sequence of models – global intercept,
127 then species, chemical, and duration intercepts, then pairwise interactions – reached
128 RMSE 1.78, 0.96, and 0.85 against 0.51 for an “ideal” model fitted to replicate mea-
129 surements, which they used as an attainable floor [11]. Gasser et al. modelled the
130 fish-only ADORE challenge, reaching RMSE near 0.90 on the $\log_{10}$ scale [19]. Two
131 of their findings recurred in ours: the results depend only weakly on the molecular
132 representation and strongly on the splitting scheme. They reported a third – insensi-
133 tivity to whether concentrations are expressed by mass or molarity – that we did not
134 separately test.

135 2.3 A note on terminology

136 Several terms in this literature denote different objects in different papers; we fix the
137 usages this study depends on here and give the full disambiguation in Additional file
138 1. We reserve *cross-species* for evaluation on species absent from training. Elsewhere,

139 it also names held-out species [3], unobserved (species, chemical) pairs [4], trait-based
140 prediction under a random split [2], pooling many species in one model [13], or a
141 compound-registry split [20]. We likewise separate the three senses of *species embed-*
142 *ding*: learned from the response, derived from external data without learning, or the
143 output of an encoder over an external descriptor [3, 4, 18]. We also note that an
144 “ablation” of species is sometimes a one-hot identity rather than a removal of species
145 information [20], so the comparison it anchors is between degrees of structure, not
146 presence versus absence.

147 2.4 What these comparisons establish, and what they leave 148 open

149 Taken together, these studies establish that species information improves multi-species
150 toxicity prediction, that structured species representations can substitute for identity,
151 and that the improvement is sensitive to how compounds are split [1, 12, 20]. Three
152 elements are absent from this practice. *No permuted-label control on the species block.*
153 In the studies we surveyed, species representations are compared to one another and,
154 in some cases, to a chemistry-only condition, but none reports a condition in which
155 the species block is retained and its assignment destroyed. The distinction matters:
156 y-randomisation permutes the response [21] and permutation importance permutes
157 at test time without refitting [22], whereas permuting species identity and refitting –
158 *permute-and-relearn* [6] – isolates what the species block contributes beyond its dimen-
159 sionality. *No test of equivalence.* Every comparison in the studies we surveyed is framed
160 as superiority; where a difference is small, it is reported as small rather than tested
161 against a pre-specified margin, so “no difference was found” and “the representations
162 are interchangeable” are not separated. Equivalence testing is established elsewhere
163 [7, 8, 23]. The practical-equivalence criterion that has entered molecular-representation
164 benchmarking declares two embeddings equivalent when the posterior probability that

one beats the other, under a hierarchical Bayesian model, falls inside a region of practical equivalence [9]. We are not aware of an equivalence-testing treatment of species representations. Separately, multiple-comparison control is itself uncommon in this setting, though not absent [24]. It matters here: in our own study, the one comparison between representations of differing information requirement that survives correction does so in a confirmatory family, not in either primary family (Section 4). *Availability is a precondition, not a result.* Each representation in category (iii) is defined only for species that possess the corresponding record, yet the fraction of species that do is not reported as a finding. Where a comparison is run on the subset for which all representations exist – as in Yuan et al., where the species-representation table is computed on 30,039 of 50,603 records [20] – that subset is by construction restricted to species that carry the auxiliary data.

3 Methods

Figure 1 summarises the design: seven species representations with label controls, trained across backbones, seeds, and splits under two evaluation regimes, and judged by pre-registered equivalence testing against a frozen margin.

3.1 Data

We used the intermediate processed output of the ADORE pipeline: 70,670 records covering 1,267 species and 3,295 compounds, derived from the US EPA ECOTOX release of September 2022 [5]. We did not use the reduced public modelling set (203 species) [5]; the intermediate table is distributed with the benchmark under CC-BY 4.0, and has been used previously at the same scale [11]. The reduction to the public set removes species in a way that correlates with the availability of auxiliary species information: the 203 retained species have a phylogenetic-distance coverage of 100% and an energy-budget trait coverage of 77%, against 30% and 8% for the 1,064 species

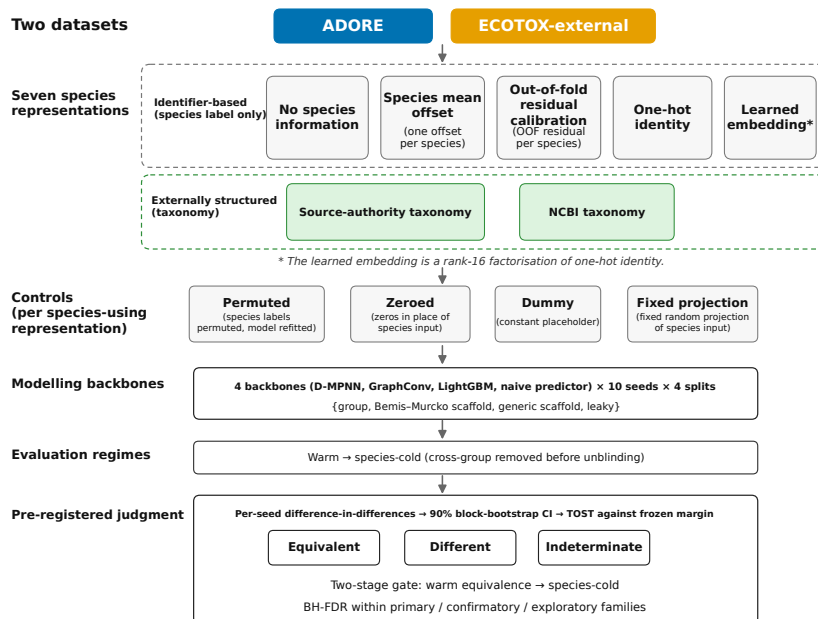


Fig. 1 Study design. Seven species representations divide into two families – identifier-based (label only; the learned embedding is a rank-16 factorisation of one-hot identity) and externally structured taxonomy (source-authority and NCBI) – and every species-using representation is accompanied by permuted, zeroed, dummy, and fixed-projection controls. The representations are trained on two datasets with four backbones, ten seeds, and three evaluation split schemes (plus a deliberately leaky split kept as a leakage positive control), and evaluated warm and species-cold; cross-group evaluation, in which an entire taxonomic group is held out, was removed before unblinding and is left to future work (Section 3.5). Verdicts follow two one-sided tests against a margin frozen before any comparison, gated in two stages with Benjamini–Hochberg control within families.

removed (Table 8). Because that availability is one object of this study, we retained the broader table and applied our own quality screening and splits. We do not assert intent on the part of the benchmark authors; the published selection criteria do not include a minimum-records-per-species rule, and we report only that the criteria are correlated with coverage. Absolute errors reported here are therefore not directly comparable to benchmarks computed on the public set.

Records were aggregated by (SMILES, species, endpoint, duration) and expressed as $-\log_{10}(\text{mol/L})$. Two partitions were distinguished (Table 1): a single-condition partition (LC_{50} at 96 h; 779 species) and a seven-condition partition spanning

Table 1 Data partitions of the ADORE processed source table. The table holds 70,670 records over 1,267 species and 3,295 compounds. Record counts are given after the exclusion of fourteen records by the screening rule (Section 3.1), and therefore sum to 70,656. The modelling units are strata aggregated by (SMILES, species, endpoint, duration); the group split is by compound registry. Warm out-of-vocabulary (OOV) counts are the numbers of test species, with the corresponding numbers of test rows in parentheses, that are absent from training under the group split.

Partition	Records	Strata	Species	Cmpds.	Test str.	Test sp.	Warm OOV
single-condition (LC ₅₀ , 96 h)	26,523	10,149	779	2,081	2,029	396	61 (69)
seven-condition (LC ₅₀ /EC ₅₀ , 24–96 h)	44,133	21,759	1,006	2,760	4,351	479	54 (85)

24/48/72/96 h and EC₅₀ endpoints (1,006 species). A screening rule fixed before training excluded records below a lower concentration bound ($-\log_{10}(\text{mol/L}) > -1.75$) and records whose mass concentration fell below 0.1 ng/L, with borderline cases judged individually and logged. Fourteen records were excluded, all from the seven-condition partition. We did not apply a water-solubility filter, unlike two adjacent studies that excluded records whose effect concentration exceeded five times water solubility [2, 20]. Retaining these records makes the task harder, and our absolute errors are correspondingly higher. It may also inflate residual variance and reduce the power to resolve differences between representations; since our finding is an absence of resolvable differences, we treat this as a threat to that finding rather than as support for it (Section 5.1). The extent to which this matters could not be determined, because the units of the solubility column in the source table are undocumented. Among the ten compounds with the widest concentration spread, eight had a web-verifiable solubility, and records exceeded it for five of them; the largest excess was a factor of 2.97×10^5 , for DDT, where 202 of 276 records exceeded solubility.

3.2 Species representations

The seven representations divide into two families by the information each requires (Table 2). *Identifier-based* (nothing beyond the species label): no species information, a species mean offset, out-of-fold residual calibration, one-hot identity, and a learned embedding. *Externally structured* (a record outside the toxicity data): source-authority

Table 2 The seven species representations, grouped into two families by information requirement, and their availability by backbone. Per-species d.o.f. is the number of degrees of freedom the representation introduces for each species. Every species-using representation was trained alongside three label controls – permuted (permute-and-relearn), zeroed, and dummy – and a capacity-matched frozen random projection. Phylogenetic distance and energy-budget traits are not listed here: they were not evaluated for predictive performance, and their coverage is reported in Table 8. The identifiers in the last column are those used in the frozen pre-registration and in the code repository, and are given here so that the reported comparisons can be matched to those records. ^a An out-of-fold post-hoc correction, not a fitted per-species parameter. ^b Mathematically degenerate for the naive predictor: a one-hot species mean is identical to the species mean offset, and no gradient-trained embedding exists. This is not a scope choice. ^c Evaluated, but reported only in Additional file 1.

Representation	Per-species d.o.f.	D-MPNN	GraphConv	LightGBM	Naive	Pre-reg.
<i>Identifier-based</i> (nothing beyond the species label)						
no species information	0	✓	✓	✓	✓	t0
species mean offset	1	✓	✓	✓	✓	t1
out-of-fold residual calibration	post-hoc ^a	✓	✓	✓	✓	t1'
one-hot identity	unconstrained	✓	✓	✓	– ^b	t2
learned embedding	16	✓	✓	✓	– ^b	t4
<i>Externally structured</i> (a record outside the toxicity data)						
source-authority taxonomy	rank membership	✓	✓	✓	S ^c	t3a
NCBI taxonomy	rank membership	✓	✓	✓	S ^c	t3b

219 taxonomy and National Center for Biotechnology Information (NCBI) taxonomy, each
220 at four ranks. The learned embedding is a 16-dimensional lookup table indexed by
221 species; its input is the species index and nothing else, so it is a rank-16 factorisation
222 of the mapping that one-hot identity learns without a rank constraint. It therefore
223 belongs to the identifier-based family. Phylogenetic distance and energy-budget traits
224 were not evaluated for predictive performance; the reason, availability, is given in
225 Section 4. By phylogenetic distance, we mean specifically the curated pairwise distance
226 matrix distributed with the benchmark, not a taxonomy-derived cophenetic distance,
227 which is available wherever taxonomy is and is in effect the source-authority taxonomy
228 condition. One-hot identity serves as the reference throughout: the contrast from no
229 species information to one-hot identity measures the value of species information, and
230 the contrasts from one-hot identity to the remaining representations measure the value
231 of the choice among them.

3.3 Control conditions

Every species-using representation was accompanied by control arms trained under the identical procedure. *Permuted* – the species index is randomly permuted across rows, independently within train, validation, and test, and the model is refitted; chemistry and response are untouched, which isolates the contribution of the species assignment from that of the dimensionality the species block introduces. This is *permute-and-relearn* [6], distinct from y-randomisation [21] and from permutation importance [22]. *Zeroed* – the species input is set to zero. *Dummy* – the species input is a constant placeholder. *Fixed projection* – the learned embedding is replaced by a frozen random projection whose readout width matches that of the learned embedding: a capacity control that separates the effect of learning from the effect of the rank-16 bottleneck. In the warm regime, this yields eight primary arms (the seven representations plus the capacity control), six permuted, five zeroed, and five dummy, per split and backbone.

3.4 Models and splits

We used four backbones: two message-passing networks – a graph-convolutional model (GraphConv) and a directed message-passing network (D-MPNN) – plus a gradient-boosted tree model (LightGBM) and a naive predictor. Both networks are our own PyTorch implementations within the message-passing framework of Gilmer et al. [25]. The directed message-passing backbone follows Yang et al. [26]; the graph-convolutional backbone is a bond-aware convolution in which each atom is updated from a transform of its own state together with the mean over neighbours of a linear transform of each neighbour’s state concatenated with the connecting bond feature, and is specified in full in the released code. The depth reported below is the number of message-passing steps in both implementations. In both networks, the molecular vector is computed without species information, and the species vector is concatenated immediately before the readout feed-forward network. The readout construction,

species-vector dimensionality, species-path parameter count, and fusion point were identical between the two networks, and no per-backbone hyperparameter tuning was performed. The species vector is 16-dimensional for the learned embedding, for both taxonomy representations (the per-rank embeddings are summed), and for the fixed-projection control; one-hot identity enters as the full n_{species} indicator. A single fixed configuration was applied throughout: hidden width 300, depth 3, dropout 0.1, embedding dimension 16, batch size 256, learning rate 5×10^{-4} , weight decay 10^{-5} , the AdamW optimiser, mean squared error loss, at most 100 epochs with early stopping (patience 15), and a 10% validation carve from training. The network width, depth, and molecular featurisation follow the defaults of Yang et al. [26]; the dropout and optimiser settings were fixed by us and were not tuned. Four split schemes were used per partition: grouping by compound registry number (CASRN), Bemis–Murcko scaffold, generic scaffold, and a deliberately leaky split retained as a positive control for the leakage tripwire (Section 3.8). Ten random seeds were used for the networks and one for the deterministic models; a further 90 seeds were used for the ensemble sensitivity analysis. All runs used Python 3.11, PyTorch 2.12.1 (CUDA 13.0), RDKit 2026.3.3, and LightGBM 4.6.0, with both scaffold assignments computed by RDKit; the full environment specification is archived with the code. The complete run inventory comprises 6,286 prediction files. In the warm regime (Section 3.7), a small fraction of test species is nonetheless absent from training because splitting is by compound: in the CASRN split, 61 of 396 test species (69 of 2,029 rows) do not appear in training and retain an untrained embedding. We report this rate rather than removing the rows.

3.5 Pre-registration and freeze

All criteria that determine a verdict – the margin specification, the significance level, the verdict categories, the gate condition, the definition of the effect measure, and the assignment of comparisons to families – were fixed in a written specification before

any comparison was computed. We refer to that specification as our pre-registration throughout: it is a pre-specified, outcome-blinded analysis plan archived with the code, not a registration lodged with an external registry. Training and prediction preceded the freeze, and the freeze preceded any comparison: no comparison output existed at the time of freezing. Choices that do not determine a verdict – figure placement, section order, wording – were made after the results were seen [27]. A protocol for breaking the freeze was recorded in advance.

The pre-registered gate had three stages – warm, species-cold, and cross-group (an entire taxonomic group held out) – ordered by widening prescription scope. The third stage was removed before unblinding, for a reason fixed by the taxonomy and not by the results. Fish, crustaceans, and algae share no rank at any of the four levels (class, order, family, and genus) in either the source or NCBI resolution. Therefore, outside its training group, every discrete species representation is equally undefined, and a cross-group comparison between representations is empty by construction rather than merely underpowered. Cross-group evaluation, and the continuous phylogenetic and energy-budget representations it would test, are left to future work; we report a two-stage gate.

3.6 Margins

The margin is specified as a procedure and not a constant. Within-condition variation is the standard deviation of per-seed RMSE within a cell (10 seeds, 9 degrees of freedom), and the margin is the pooled standard deviation across cells,

$$\delta = \sqrt{\sum_c (n_c - 1) s_c^2 / \sum_c (n_c - 1)}, \quad (1)$$

where a cell c is one (representation, backbone) combination, n_c is the number of seeds contributing an RMSE in that cell, and s_c is the standard deviation of those per-seed

Table 3 Equivalence margins, frozen before any comparison and re-derived independently on the ECOTOX-external set. δ = pooled within-cell SD of per-seed RMSE (14 cells); δ' = ensemble margin; δ_{det} = block-bootstrap SD for deterministic models. Values are given at full precision here; the main text quotes them rounded to four decimal places.

Margin	ADORE	ECOTOX-external	Ratio
δ (per-seed)	0.019777	0.035641	$1.80\times$
δ' (ensemble)	0.005717	0.015957	$2.79\times$
δ_{det} (deterministic)	0.087189	0.074082	$0.85\times$

307 RMSE values. It was derived over 14 cells (seven representations $\times$ two backbones) in
 308 the single-condition partition, under compound-registry splitting, in the warm regime,
 309 and over primary arms only, giving 126 degrees of freedom and $\delta = 0.0198$ on the $\log_{10}$
 310 scale. Seeds were not all executed on the same graphics processor: two devices of differ-
 311 ent generation and class were used (NVIDIA GeForce RTX 4090 and NVIDIA GeForce
 312 RTX 5060 Ti), the assignment of seeds to processors was fixed in the pre-registration,
 313 and we decomposed the per-seed variance by processor after unblinding. The proces-
 314 sor accounts for 4.5% of that variance on ADORE and 6.2% on the ECOTOX-external
 315 set (Additional file 1), so δ measures re-run noise rather than a hardware artefact.
 316 The margin therefore measures the re-run noise of a model that is learning signal, a
 317 variance component that is smaller than, and distinct from, the between-compound
 318 sampling variation dominating the confidence intervals against which it is compared
 319 (Section 4). Two further margins were derived using the same logic: an ensemble mar-
 320 gin ($\delta' = 0.0057$) and a margin for deterministic models, where re-run noise is zero by
 321 construction and sampling variation was used instead ($\delta_{\text{det}} = 0.0872$, block bootstrap).
 322 All three were re-derived on the ECOTOX-external set ($\delta = 0.0356$, $\delta' = 0.0160$,
 323 $\delta_{\text{det}} = 0.0741$; Table 3) rather than reused as constants. No relevance margin was
 324 specified: our pre-registration contains none, and we did not adopt one after seeing
 325 the results, because no external quantitative acceptance criterion exists in this area
 326 [28]. Effect sizes are reported on both the $\log_{10}$ and concentration-fold scales.

3.7 Effect measure and judgment

The effect measure is a per-seed paired difference-in-differences (dd), (candidate – candidate base) – (reference – reference base), where the base of an arm is the matched no-species condition trained under the same backbone and pipeline, except when a representation is compared with its own permuted twin, in which case the permuted arm is the base. Each side of the difference is therefore an improvement over its own species-free baseline. Within a backbone the two bases coincide and cancel, so the dd reduces to a direct difference in RMSE; the bases remain only in the exploratory cross-backbone comparisons, where each arm is anchored to the no-species condition of its own model family. Confidence intervals were obtained by block bootstrap with 2,000 resamples, blocking on SMILES and resampling seeds jointly across the arms. Verdicts were assigned by two one-sided tests (TOST) at $\alpha = 0.05$, equivalent to a 90% confidence interval, against the appropriate margin: *equivalent* (interval within $\pm\delta$), *different* (interval excluding zero), and *indeterminate* (neither); where both conditions hold, the verdict is *equivalent*. Before unblinding, we recorded our expectation that indeterminate results would be common, on the ground that a margin of this width is narrow relative to the confidence intervals the design can be expected to produce. No formal power calculation was carried out. Comparisons were assigned to three families before results were observed: primary (single-condition partition), confirmatory (seven-condition partition, with identical hypotheses and procedure), and exploratory, with Benjamini–Hochberg control within each. Verdict categories were determined by the confidence interval alone; adjusted p -values are reported alongside, and where a comparison is called *different* with $q > 0.05$ we state that it does not survive correction. A two-stage gate was specified: warm comparisons – those evaluated on the test set as produced by the compound-based split, in which most but not all test species also occur in training (Section 3.4) – formed stage one, and only comparisons returning *equivalent* proceeded to the species-cold stage, in which test species are absent from

354 training by construction, each (backbone, representation pair) forming an indepen-
355 dent chain. One pre-registered stratified analysis decomposed the primary equivalence
356 pairs by per-species training-record bin (1–5, 6–20, 21–100, and more than 100).

357 **3.8 Pipeline safeguards**

358 Three automated guards were used throughout, each of which caught an actual defect
359 before the reported runs: a representation-input guard, which halts a representation
360 whose species input has degenerated – all null, or of cardinality one – as happened
361 when a faulty taxonomy join silently reduced the NCBI taxonomy condition to the no-
362 species condition; a guard requiring stratum and taxonomic-group labels to be derived
363 from the curated dataset rather than from per-run prediction files, since the latter
364 assigned all 396 test species of the single-condition partition to the unresolved stratum,
365 against 2 of 396 when the labels were taken from the dataset; and an atomic-write
366 guard, which replaces shared files in a single operation so that a concurrent reader sees
367 either the complete previous file or the complete new one, never a partially written one.
368 A stopping threshold on implausibly low error acted as a leakage tripwire; it fired only
369 in the deliberately leaky split and nowhere else, which we report as a positive control
370 for the tripwire itself. The learned embedding condition was additionally checked for
371 leakage by a permutation test with a threshold recorded in advance and by an index-
372 level verification that the out-of-fold boundary was never crossed.

373 **3.9 The ECOTOX-external set**

374 The ECOTOX-external set was constructed from the ECOTOX release of June 2026
375 by removing every record present in the ADORE processed table, recovering the
376 remainder through our own pipeline, and excluding records sharing reference, species,
377 compound, and duration with the original set: 22,809 aggregated strata over 1,975
378 species and 1,953 compounds. The rows are disjoint from those of the ADORE data,
379 whereas species and compounds overlap. The NCBI taxonomy resolver used for the

ADORE data was not retained and was reimplemented from the same taxonomy snapshot; agreement with the original output was 100% on resolution status and 99.7%, 98.0%, 99.7%, and 100% on class, order, family, and genus names respectively, every apparent order-level disagreement being a difference between a null and an empty value. All three margins were re-derived on this dataset rather than reused as constants (Section 3.6). The aggregated set is deposited under CC0 1.0 [29]. The availability analysis of Section 4 is conducted at a wider scale than the modelling. Its species pool was obtained by retaining every aquatic record of the June 2026 ECOTOX release whose endpoint, effect, exposure medium, and habitat each take one of the values that occur in the ADORE processed table, without the duration, concentration, and structure filters that further define that table; this gives 3,268 species over 160,869 records, assigned to fish, crustacean, algae, and other groups by the `ecotox.group` field of the source data. A species was recorded as having a phylogenetic distance or an energy-budget trait vector by the corresponding availability flag of the ADORE table where the species is one of the 1,267 modelled here, and by name matching against the two resources otherwise; species names were resolved against the ECOTOX synonym list before matching.

3.10 Literature search

To bound the two comparative claims of this work – that the species-representation comparisons we surveyed do not report a permuted-label control on the species block, and do not report an equivalence test – we searched the primary and preprint literature on 31 August 2026, via scholarly and web search (Google Scholar, PubMed, and the bioRxiv and arXiv servers). We looked for machine-learning models that predict toxicity or species sensitivity for more than one species and that compare two or more ways of representing the species, or a species representation against a no-species baseline. Search terms combined {species representation, species embedding, taxonomic

Table 4 Studies examined in the literature search. These are machine-learning models that predict toxicity for more than one species and compare species representations. partitions of the ADORE processed source table. The table holds 70,670 records over 1,267 species and 3,295 compounds. “Perm. control” = a control that permutes the species assignment and refits the model; “Equiv. test” = a test against a pre-specified equivalence margin. None of the studies reported either.

Study	Species representations compared	Perm. control	Equiv. test
Guénard et al. 2014 [17]	phylogenetic eigenvectors and chemical descriptors	no	no
Sheffield & Judson 2019 [1]	one-hot vs taxonomic group	no	no
Wu et al. 2022 [12]	taxonomic group vs chemistry-only	no	no
Zubrod et al. 2023 [2]	energy-budget traits vs categorical species	no	no
Schlender et al. 2023 [13]	species / class / phylum; species-as-task	no	no
Viljanen et al. 2023 [4]	one-hot vs low-rank latent factor	no	no
Zubrod et al. 2024 [15]	energy-budget traits (extended data)	no	no
Posthuma et al. 2025 [11]	null / species / pairwise interaction	no	no
Yuan et al. 2025 [20]	one-hot / source and NCBI taxonomy / energy-budget traits	no	no
Yang et al. 2026 [18]	taxonomy embedding vs sequence embedding	no	no
Forastiere et al. 2026 [3]	genetic/phylogenetic embedding vs taxonomy	no	no

encoding, cross-species, multi-species} with {toxicity, QSAR, ecotoxicology}, and, in targeted queries, with {permutation control, shuffled label, permute-and-relearn} and {equivalence test, two one-sided tests, TOST, region of practical equivalence}. We excluded single-species models and did not count response permutation (y-randomisation) or test-time permutation without refitting as a species-block control. The search identified 11 such studies (Table 4).

Among them, none reported a control that permutes the species assignment and refits the model, or an equivalence test against a pre-specified margin; a practical-equivalence criterion has been reported for molecular-representation benchmarking [9], but we found none applied to species representations.

3.11 Use of large language models

A large language model (Claude, Anthropic) was used for L^AT_EX formatting, and figure and table editing. It was not part of the modelling or analysis pipeline: it did not design the study, generate or analyse the data, or determine any scientific claim, and all content was verified by the authors.

4 Results

4.1 Species information improves prediction

Adding species information to a no-species model lowered RMSE in all twelve dataset $\times$ backbone $\times$ split cells, counting the three non-leaky splits, and the improvement was significant in the primary family on both datasets (Fig. 2, Table 5). On the ADORE data, one-hot identity beat no species information by $dd = -0.076$ $[-0.124, -0.025]$, $q = 0.012$ under D-MPNN and -0.067 $[-0.114, -0.018]$, $q = 0.039$ under GraphConv; on the ECOTOX-external set, the same contrast was -0.150 and -0.195 , both $q < 0.001$. That this gain comes from the species assignment and not merely from the added parameters is shown by a permute-and-relearn control: each species-using representation was retrained on randomly permuted species labels, holding its per-species degrees of freedom fixed, and every one beat its permuted twin (Table 6). The difference was significant after correction in twenty-three of the twenty-four cells, the sole exception being the species mean offset under D-MPNN on the external set, with the real-minus-permuted difference ranging from -0.048 to -0.403 . Under the scaffold schemes, which belong to the exploratory families, the interval excluded zero in seven of eight cells; after multiplicity correction, these survived on the external set ($q \leq 0.019$) but not on ADORE ($q \geq 0.077$). The deliberately leaky split, a pre-registered positive control for the leakage tripwire, was not used to support this conclusion.

4.2 Performance does not identify a representation

Among the richer representations, the differences from one-hot identity are small (Table 5). Nine of the twelve one-hot-to-richer comparisons had $|dd|$ below the dataset’s own re-run noise δ (four of six on the ADORE data, five of six on the ECOTOX-external set). Not one comparison between species representations returned an equivalence verdict on the trained networks in the primary or confirmatory family on either dataset: the primary equivalence family was 0/1/5

(equivalent/different/indeterminate) on both (Table 7). Equivalence appeared only for the deterministic baselines, against their much wider margin δ_{det} , and for a few exploratory comparisons: rank truncation and the generic-scaffold split on the external set, and the leaky positive control. The two comparisons that reached significance pointed in opposite directions. NCBI taxonomy was better than one-hot identity under GraphConv on ADORE ($dd = -0.043 [-0.079, -0.009]$, though it does not survive correction, $q = 0.060$), whereas the learned embedding was worse than one-hot identity under GraphConv on the external set ($+0.044 [+0.009, +0.080]$, $q = 0.045$). This last comparison, the only one in a primary family to survive correction on the external set, became indeterminate when the 341 test rows (7.5%) whose species are absent from training were removed. The one comparison between representations of differing information requirement that survived correction anywhere did so in the confirmatory family on ADORE, not in a primary family. Where an effect was significant on one dataset and indeterminate on the other, the two confidence intervals overlapped; we therefore say that the effect was not reproduced, not that it reversed.

The interpretation of these verdicts follows from a structural feature of the test, which we examined in a post-hoc precision analysis. The margin δ is the standard deviation of the aggregate RMSE across random seeds – re-run noise measured with the test set held fixed. In contrast, the confidence interval on each difference-in-differences is obtained by resampling compounds (block bootstrap on SMILES) and is dominated by between-compound heterogeneity, a larger and more distinct source of variation. On the trained networks, the two are mismatched: none of the twelve one-hot-to-richer intervals was contained in $[-\delta, +\delta]$. Their half-widths (0.024–0.036 on ADORE and 0.036–0.057 on the external set) were comparable to or larger than δ itself (0.020 and 0.036 respectively), and the intervals that reached significance excluded zero, so no network comparison could be certified equivalent. This reflects the margin being finer than the estimator’s own uncertainty – both fixed before the data were seen –

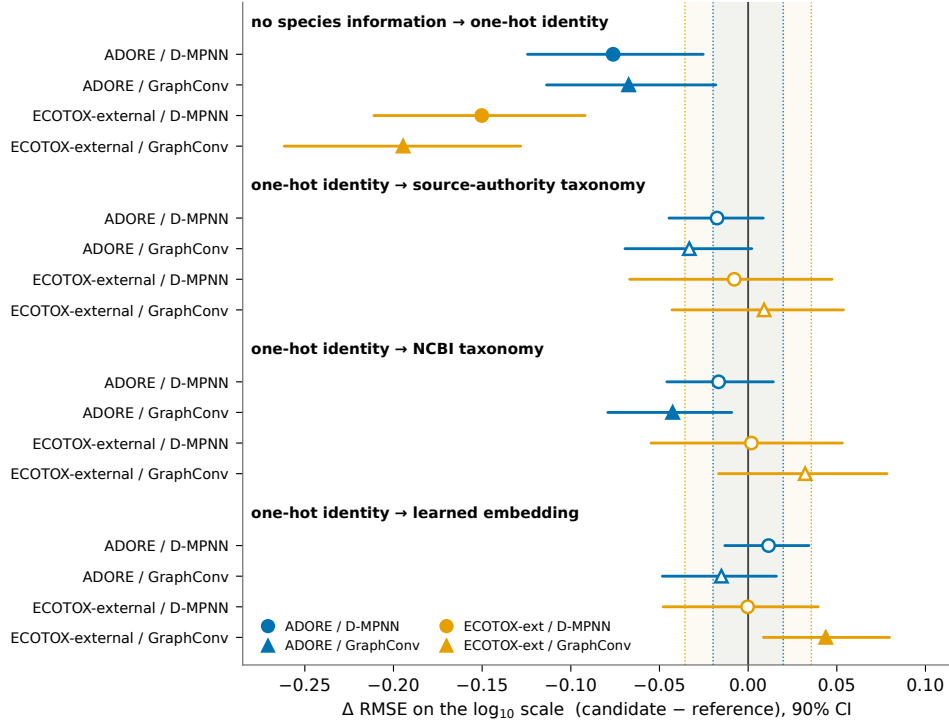


Fig. 2 Headline effect sizes for the sixteen comparisons on the primary partition. Differences-in-differences in RMSE on the $\log_{10}$ scale with 90% confidence intervals. Colour distinguishes the two datasets and marker shape the backbone; a filled marker indicates that the interval excludes zero. Dotted lines mark $\pm\delta$, the re-run noise of the dataset concerned. Adding species information improves prediction in every cell; among the richer representations most estimates fall within the noise band, and the two estimates whose intervals exclude zero – NCBI taxonomy under GraphConv on the ADORE data and the learned embedding under GraphConv on the ECOTOX-external set – point in opposite directions; only the latter survives multiplicity correction.

not the representations resembling one another. The deterministic baselines returned equivalence only because their margin δ_{det} is roughly two to four times wider and can therefore contain their intervals (Table 7). We therefore read the primary result as a statement about where the estimates lie relative to the margin, rather than as certified equivalence: the resolvable differences fell within the models' own re-run noise ($|\text{dd}| < \delta$ in nine of twelve cells), and the few that reached significance disagreed across datasets and evaluation subsets. The two-stage gate, which admits a comparison to the species-cold stage only on an equivalence verdict, was consequently structurally empty.

Table 5 Headline effect sizes: per-seed paired difference-in-differences (dd) in RMSE on the $\log_{10}$ scale, with 90% block-bootstrap confidence intervals and Benjamini–Hochberg q . Negative favours the candidate (the richer representation). The contrast from no species information to one-hot identity is a superiority test in the primary family; the three contrasts from one-hot identity are equivalence (TOST) tests in a separate family, so the two q columns are not comparable. Verdict: sig. (superiority) and diff. (equivalence) both mean the interval excludes 0; equiv. = within $\pm\delta$; indet. = neither. Effect sizes are given to three decimal places; the entry -0.000 is a value of -0.0002 rounded at that precision.

Comparison	Backbone	ADORE				ECOTOX-external			
		dd	90% CI	q	verdict	dd	90% CI	q	verdict
no species information $\rightarrow$ one-hot identity	D-MPNN	-0.076	[-0.124, -0.025]	0.012	sig.	-0.150	[-0.211, -0.092]	<0.001	sig.
	GraphConv	-0.067	[-0.114, -0.018]	0.039	sig.	-0.195	[-0.262, -0.128]	<0.001	sig.
one-hot identity $\rightarrow$ source-authority taxonomy	D-MPNN	-0.018	[-0.045, +0.009]	0.317	indet.	-0.008	[-0.067, +0.047]	0.924	indet.
	GraphConv	-0.033	[-0.069, +0.002]	0.133	indet.	+0.009	[-0.043, +0.054]	0.811	indet.
one-hot identity $\rightarrow$ NCBI taxonomy	D-MPNN	-0.017	[-0.046, +0.014]	0.369	indet.	+0.002	[-0.055, +0.053]	0.924	indet.
	GraphConv	-0.043	[-0.079, -0.009]	0.060	diff.	+0.032	[-0.017, +0.078]	0.302	indet.
one-hot identity $\rightarrow$ learned embedding	D-MPNN	+0.012	[-0.013, +0.034]	0.451	indet.	-0.000	[-0.048, +0.040]	0.926	indet.
	GraphConv	-0.015	[-0.048, +0.016]	0.397	indet.	+0.044	[+0.009, +0.080]	0.045	diff.

Table 6 Permute-and-relearn control. Each species-using representation is compared against a twin in which the species labels were randomly permuted and the model re-trained from scratch; the permuted twin keeps the same per-species degrees of freedom, so a negative difference-in-differences (dd, RMSE on the $\log_{10}$ scale, real minus permuted) isolates the value of the *correct* species assignment from the value of the added capacity. Ninety-percent block-bootstrap confidence intervals and Benjamini–Hochberg q are in the same superiority family as the contrast from no species information to one-hot identity. The zeroed and dummy label controls and the capacity-matched frozen random projection (Table 2) are further controls of the same design.

Representation vs. permuted twin	Backbone	ADORE			ECOTOX-external		
		dd	90% CI	q	dd	90% CI	q
species mean offset	D-MPNN	-0.077	[-0.113, -0.042]	<0.001	-0.056	[-0.110, +0.001]	0.127
	GraphConv	-0.048	[-0.075, -0.018]	0.008	-0.080	[-0.125, -0.033]	0.020
out-of-fold residual calibration	D-MPNN	-0.273	[-0.350, -0.210]	<0.001	-0.402	[-0.471, -0.333]	<0.001
	GraphConv	-0.288	[-0.364, -0.223]	<0.001	-0.403	[-0.472, -0.334]	<0.001
one-hot identity	D-MPNN	-0.128	[-0.177, -0.074]	<0.001	-0.134	[-0.205, -0.060]	0.007
	GraphConv	-0.106	[-0.156, -0.049]	0.002	-0.158	[-0.231, -0.084]	0.002
source-authority taxonomy	D-MPNN	-0.146	[-0.204, -0.083]	<0.001	-0.118	[-0.174, -0.065]	<0.001
	GraphConv	-0.118	[-0.175, -0.061]	<0.001	-0.116	[-0.176, -0.061]	0.002
NCBI taxonomy	D-MPNN	-0.134	[-0.198, -0.069]	0.002	-0.110	[-0.163, -0.057]	0.002
	GraphConv	-0.127	[-0.181, -0.071]	<0.001	-0.102	[-0.166, -0.045]	0.015
learned embedding	D-MPNN	-0.104	[-0.156, -0.049]	0.002	-0.104	[-0.153, -0.057]	<0.001
	GraphConv	-0.113	[-0.155, -0.066]	<0.001	-0.093	[-0.139, -0.049]	0.003

4.3 The result is not driven by species with little data

Decomposing the primary equivalence pairs by the pre-registered per-species training-record bins showed that the cells whose interval excluded zero were not concentrated in the sparse bins: six of twenty-four such cells on the ADORE data and three of twenty-four on the ECOTOX-external set (Table 7, exploratory support-bin family). The closest prior comparison reports a substantially larger one-hot-to-taxonomy gain on a

Table 7 Gatekeeping census, reported as equivalent / significant / indeterminate counts per family; for the equivalence (TOST) families the middle column is the pre-registered “different” verdict (interval excludes zero), and for the superiority families it is a significant improvement. The confirmatory family is absent by construction from the ECOTOX-external set, which has no second partition. No comparison reached the species-cold gate stage on either dataset.

Family	ADORE (eq/sig/indet)	ECOTOX-external (eq/sig/indet)
Primary: superiority	0/24/0	0/23/1
Primary: equivalence (TOST)	0/1/5	0/1/5
Confirmatory: superiority	0/20/4	–
Confirmatory: equivalence	0/2/4	–
Deterministic: equivalence	3/0/0	2/1/0
Exploratory: rank truncation	0/1/11	2/4/6
Exploratory: support bin	0/6/18	0/3/21
Exploratory: taxon group	0/4/14	0/1/17
Exploratory: scaffold	0/42/18	1/51/8
Exploratory: leakage control	1/28/1	0/28/2
Exploratory: cross-backbone	0/1/5	0/0/6
Ensemble sensitivity	0/5/7	0/1/5

corpus several times larger [20]; ours ranged from 0.017 to 0.043 on the ADORE single-condition partition (779 of the 1,267 species in that dataset). Because the external set carries 1,975 species and the difference neither concentrated in the sparse bins nor enlarged, the discrepancy with that work is not explained by species count.

4.4 The auxiliary data are missing for most species

At the scale of the full ECOTOX aquatic species pool (3,268 species), phylogenetic distances exist for 17.7% of species but 55.7% of records, and energy-budget traits for 7.6% of species but 51.2% of records. The record scale exceeds the species scale by approximately three-fold for phylogenetic distance and seven-fold for energy-budget traits, because each resource is concentrated in species that already carry many measurements (Fig. 3, Table 8). Coverage collapses outside fish: phylogenetic distance reaches 63% of fish species but about 15% of both algae and crustaceans, and under 1% of the remaining taxa, which are largely insects, molluscs, and other invertebrates. Energy-budget traits are absent for algae entirely; an energy-budget model describes animal physiology and does not extend to algae [14], a boundary of the resource rather

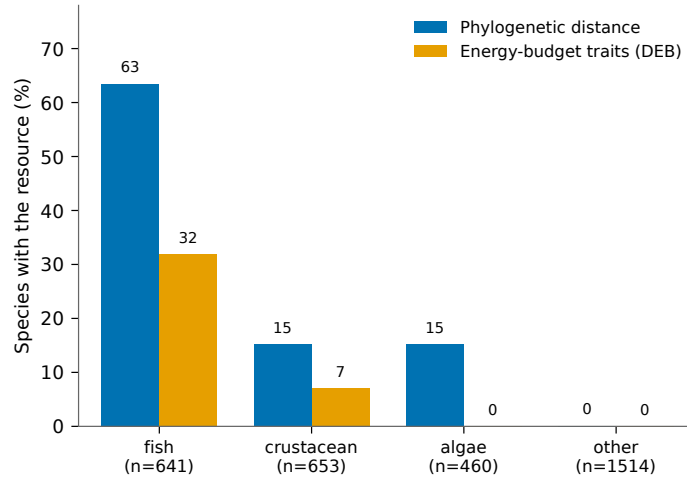


Fig. 3 Coverage of the two auxiliary resources by taxonomic group at field scale. The field scale is the full ECOTOX aquatic species pool (3,268 species). Percentages are rounded to the nearest integer; phylogenetic distance reaches 0.07% of the remaining taxa. Energy-budget traits are absent for algae entirely, and both resources collapse outside fish.

than a curation gap. The ADORE public subset is enriched on exactly this variable; the 203 species it retains have 100% phylogenetic-distance and 77.3% energy-budget trait coverage, against 30.4% and 7.6% for the 1,064 species it drops (32.2% and 16.9% by record weight). Taxonomy, by contrast, is available for every species.

4.5 Cost falls into two layers

The representations are therefore separated by the reason they are set aside. Taxonomy carries no availability penalty: the source authority covers all species, and an independent NCBI resolution recovers four ranks for roughly 86% of species on both datasets, validated against the original resolution (rank-name agreement 99.7%, 98.0%, 99.7%, and 100% at class, order, family, and genus respectively). Phylogenetic distance and energy-budget traits, by contrast, are set aside not for performance but because they cannot be computed for most species, and where they can be, they concentrate on the species already richest in data. Taxonomy is thus usable but without demonstrated benefit; the two continuous resources are unusable at scale.

Table 8 Availability of the two auxiliary resources, on two scales: raw species count and record-weighted. Source-authority taxonomy is available for all species, and the NCBI pipeline returns a representation for every species although it resolves all four reported ranks for approximately 86%; neither is shown. The public modelling subset is enriched for covered species relative to those it drops. The algae row is a field-scale subset, not a subset of the row above it.

Species set	<i>n</i>	Phylogenetic distance		Energy-budget traits	
		species %	record %	species %	record %
ECOTOX aquatic species pool (field scale)	3,268	17.7	55.7	7.6	51.2
ADORE processed table	1,267	41.5	75.3	18.8	68.3
ADORE public subset (retained)	203	100.0	100.0	77.3	97.7
dropped from public subset	1,064	30.4	32.2	7.6	16.9
algae (field scale)	460	15.2	20.7	0.0	0.0

4.6 The ordering of representations is not stable across datasets

A post-hoc estimate of how much species-axis variance each representation recovers (Fig. 4) is a restatement of the effect sizes rather than independent evidence, and its denominator includes measurement noise, so it is not bounded above by 100%. It makes one observation legible: the ordering of representations is not stable across datasets. On ADORE, taxonomy recovered slightly more than one-hot identity (23–27% versus 17–19%); on the external set, one-hot identity recovered as much or more (36–47%). A species-by-chemical interaction is present in the data: species-sensitivity rankings rearrange across chemical classes (median $\rho = 0.23$ over the twenty-eight pairs of chemical classes that could be compared, and median $\rho = 0.31$ over the fifteen pairs among the six classes for which a within-class split-half reliability could be computed). The latter is about half the median reliability ceiling for those classes ($\rho = 0.63$ after Spearman–Brown correction), so the rearrangement is not attributable to measurement noise alone. However, this interaction is suggested rather than resolved, and the reading that these representations are not separated by how much species-axis variance they recover rests on the instability of the ordering, not on a low recovery rate.

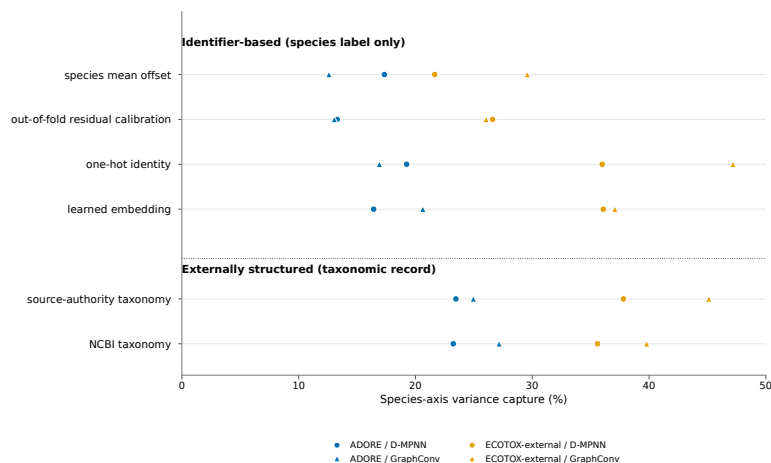


Fig. 4 Species-axis variance capture for each species-using representation, per dataset and backbone. Colour distinguishes the two datasets and marker shape the backbone, as in Figure 2. Capture is zero by construction for the no-species condition, which is therefore not shown. The representations are grouped by family and are listed in the order used in the pre-registration (Table 2); the vertical order is not a ranking. This is a calculation whose denominator includes measurement noise, so it is not bounded above by 100%, and it restates the effect sizes rather than providing independent evidence: the two datasets differ in level, and the ordering among representations does not agree between them.

536 4.7 The apparent taxonomy advantage is backbone-dependent

537 The one comparison in which taxonomy was better than one-hot identity arose under
 538 GraphConv and not under D-MPNN. We first checked the implementation: the read-
 539 out feed-forward network is shared between the two backbones, and the species-path
 540 dimensionality, parameter count (16,272 for NCBI taxonomy on the single-condition
 541 partition), and fusion point are identical. The asymmetry is therefore not an imple-
 542 mentation artefact. It is consistent with an architecture-by-representation interaction,
 543 but we did not test that interaction directly, and the ADORE effect itself does not
 544 survive multiplicity correction ($q = 0.060$). Because both networks compute the molec-
 545 ular vector without species information, a species representation can only fill what
 546 the molecular summary leaves, and the two backbones leave different residuals. This
 547 is a single post-hoc observation, and we do not read it as a recommendation to pair
 548 taxonomy with a particular backbone.

5 Discussion

Our results place the choice of species representation on a different footing from predictive accuracy. Species information is a definite benefit, lowering the error in every cell examined, but among the representations that carry it, performance does not single one out. Most differences fall within the models’ own re-run noise, and the handful that reach significance disagree in direction across datasets and evaluation subsets. A performance-first reading would report these as null results and treat the representations as interchangeable. The equivalence tests are more careful: they returned “indeterminate” rather than “equivalent” because certifying interchangeability requires showing positively that the difference lies within the margin, and our precision was not sufficient to do so at the margin we derived. Either way, accuracy does not tell a practitioner which representation to adopt. This is not a finding that the models fail to use species information: under the same fixed configuration, every representation beat its permuted twin, and each recovered a substantial share of the species-axis variance. What the evidence does not support is a stable ranking among them, and whether a differently configured species path – wider, tuned per representation, or fused earlier – would produce one is a question this design leaves open.

What separates the representations is the data each requires. Taxonomy is available for every species: the source authority is complete, and an independent resolution against NCBI recovers four ranks for roughly 86% of species on both datasets. It showed no reproducible predictive advantage over one-hot identity, and equivalence at the margin we derived was not established. Phylogenetic distance and energy-budget traits are different: they are undefined for most species, and energy-budget traits are undefined for algae entirely. They therefore cannot be applied across most species in the field, regardless of any advantage they might show on the covered subset. The two are thus set aside for different reasons: taxonomy for the absence of a performance

576 gain, and the continuous resources for the absence of coverage. A prescription that
577 ignores the second reason will recommend representations that most species cannot
578 use. This shape – performance failing to separate the options, so the choice falls to
579 cost – has begun to appear elsewhere: in molecular-representation benchmarking, two
580 pretrained embeddings judged practically equivalent in accuracy are then separated
581 by their computational and memory cost [9]. Ours differs in the axis, not the form,
582 with data availability rather than compute as the operative cost.

583 We controlled the false discovery rate within families rather than the family-wise
584 error rate over a single pool. The object of the study is a set of largely independent
585 equivalence claims – one per representation pair, backbone, and evaluation regime
586 – not the selection of a single best model from a ranked list, so the concern that
587 motivates family-wise control in classifier comparison does not apply in the same way
588 [24]. A family-wise correction over a pool that grows with every added regime would
589 instead erode the power of the primary comparison, the null we most need to resolve.

590 A larger advantage of taxonomy over one-hot identity has been reported elsewhere
591 on a corpus several times larger than ours [20]. We do not reproduce it, and our
592 design indicates why it is unlikely to be a matter of scale: the ECOTOX-external
593 set carries 1,975 species, against ADORE’s 1,267, yet the difference neither grew nor
594 concentrated in the species with the fewest records, where a scale effect would appear.
595 The discrepancy is more plausibly attributable to the model family, or to the fact that
596 such comparisons are run on the subset of species for which all representations exist
597 – a subset that, by construction, is data-rich and well-covered.

598 Finally, the equivalence margin is best treated as a procedure rather than a con-
599 stant. Re-deriving it on the ECOTOX-external set gave a value nearly twice that on
600 ADORE, so a margin frozen on one dataset and reused on another would misjudge
601 equivalence. Margins of this kind should be derived per dataset, and reporting the
602 two side by side is itself informative. Our central claim is a statement about what

603 performance cannot settle, supported by differences within noise and by inconsistent
604 significance rather than by a positive demonstration of equivalence; the continuous
605 representations, in turn, were assessed for availability, not accuracy.

606 5.1 Limitations

607 **Scope.** The study concerns acute lethal and immobilisation endpoints in aquatic
608 organisms and one class of molecular encoders; results may not transfer to architectures
609 that admit species information earlier in the network. Within that architecture the
610 species path is also fixed: the taxonomy and learned-embedding representations enter
611 the readout as 16-dimensional vectors, and the regularisation was held constant rather
612 than tuned per representation, so a representation whose useful structure exceeds that
613 width or whose optimum lies elsewhere in the regularisation space would be under-
614 served here. The two partitions differ in endpoint composition: the single-condition
615 partition is LC_{50} at 96 h and is almost entirely mortality, whereas the seven-condition
616 partition spans LC_{50} and EC_{50} at 24–96 h and includes intoxication, population,
617 growth, and physiological endpoints. Differences observed only in the seven-condition
618 partition therefore confound species with test conditions and endpoint definitions; we
619 interpret them accordingly. Our statements about what previous comparisons do not
620 report are likewise restricted to the studies we surveyed.

621 **Equivalence.** Our results do not establish that the representations are interchange-
622 able. Phylogenetic distance and energy-budget traits were not evaluated for predictive
623 performance: they are unavailable for most species, so a comparison would have been
624 restricted to the covered subset – the same restriction this study is about. Our claim
625 about these representations concerns availability, not accuracy. Not one comparison
626 between species representations returned an equivalence verdict on the trained net-
627 works in the primary or confirmatory family on either dataset. Our precision was
628 insufficient to establish equivalence at the network margin we derived, because that

margin measures seed re-run noise whereas the intervals are dominated by between-compound heterogeneity; we examine this in a post-hoc precision analysis (Section 4). The deterministic baselines did return equivalence against a margin several times wider, and we recorded this expectation before unblinding. Our margin measures re-run noise, not decision relevance; no established quantitative acceptance criterion exists in this area [28], and we did not construct one after seeing the results.

Uniformity of the null result. The null result is not uniform across families: no comparison between representations that differ in information requirement survives multiplicity correction within a primary family on either dataset, but one – NCBI taxonomy against one-hot identity under GraphConv – does survive within the confirmatory family on ADORE. The one comparison in a primary family that survived correction on the ECOTOX-external set (learned embedding versus one-hot identity, $q = 0.045$) became indeterminate when the 341 test rows (7.5%) whose species are absent from training were excluded. We report the pre-registered verdict and this sensitivity together. No verdict was reached for unseen species, because the two-stage gate was empty (Section 4). We therefore make no claim about generalisation to species absent from training and do not contest studies that evaluate that axis directly [3].

Data. Availability must be read on two scales, and the gap between them is itself a result: phylogenetic distances exist for 17.7% of species but cover 55.7% of records, and energy-budget traits for 7.6% of species but 51.2% of records. The auxiliary information is concentrated in species that already carry many measurements, and any single-scale reading overstates or understates the constraint. We did not apply the water-solubility filter used in adjacent studies [2, 20], a decision that cuts both ways (Section 3.1). The source database is not static; 2,541 of the 70,670 records used here (3.6%) are absent from the June 2026 release, having been removed or revised upstream. A benchmark built on a snapshot is therefore reproducible against that snapshot rather than against the current database.

Generalisation across datasets. The two partitions of the ADORE data are not independent repeats: they are subsets of one curated table separated by experimental condition, sharing compounds, species, and curation lineage. We therefore describe the seven-condition partition as a test of generalisation across conditions, not across datasets. The ECOTOX-external set is likewise not an independent database: it is drawn from the same source, disjoint at the record level from the ADORE data but overlapping in species (517 of the 1,267 ADORE species recur in it) and in compounds, and its reproduction of the design is partial: the confirmatory family and the corresponding portion of the gate could not be reproduced, and exploratory families were not reproduced. It is also enriched for species with few measurements, which works against the result we observe: because it consists of records the original curation excluded, its species carry fewer records each, and sparse per-species data disadvantages representations that must estimate a parameter for every species (one-hot identity and the learned embedding) while taxonomic representations share parameters across ranks and degrade more gradually. The composition therefore favours taxonomy, and yet the advantage of taxonomy seen on ADORE did not reappear.

Analysis. A small fraction of warm test rows involves species absent from training, and these rows do not behave as expected. Because splitting is by compound, 61 of 396 test species (69 of 2,029 rows) on ADORE and 187 of 934 test species (341 of 4,561 rows) on the ECOTOX-external set do not appear in training and retain an untrained embedding. Removing them moves the comparison in favour of taxonomic representations, which implies that taxonomy performed worse than an untrained identifier on those rows. For an unseen species, the one-hot column contributes essentially nothing, whereas taxonomic features support an extrapolation that is confident but not reliable. This is a post-hoc observation and not a substitute for a species-cold evaluation. The mechanistic analyses – rank correlation, attenuation control, variance-capture estimates, concentration-fold interpretation, the comparison of one-hot identity with

683 the species mean offset, the benchmark coverage analysis, the solubility check, and
684 the decomposition of per-seed variance by processor – were specified after unblinding.
685 They do not carry the status of the pre-registered verdicts.

686 6 Conclusion

687 The choice of species representation in multi-species aquatic acute-toxicity prediction
688 is not resolved by predictive performance. Species information helps in every setting
689 we tested, but the representations that carry it differ by less than the models' own
690 re-run noise, and the few significant differences do not reproduce across datasets or
691 evaluation subsets. What constrains the choice is the availability and acquisition cost
692 of the auxiliary data each representation requires: taxonomy is universal but showed no
693 reproducible predictive advantage, whereas phylogenetic distances and energy-budget
694 traits are absent for most species, and energy-budget traits are absent for algae entirely.
695 They therefore cannot be used for the majority of species in the field. Practitioners
696 should select a species representation by availability and cost rather than by reported
697 accuracy. Those who build or curate benchmarks should be aware that reduced mod-
698 elling subsets are enriched for information-rich species and understate this constraint.
699 Whether phylogenetic distance or energy-budget traits improve prediction where they
700 can be computed is a question our design deliberately leaves open; answering it, on
701 the covered subset and with the same controls, is the natural next step.

702 **Supplementary information. Additional file 1.** Format: .pdf. Title: Supporting
703 analyses, full enumerations, and terminology. Description: terminology disambigua-
704 tion; the water-solubility check; the complete run inventory; the decomposition
705 of per-seed variance by processor; naive-predictor results for the two taxonomy
706 representations; the full gatekeeping enumeration; the rank-truncation, support-
707 bin, taxon-group, scaffold, leakage-control, and cross-backbone exploratory families;
708 the ensemble sensitivity analysis; the species-axis variance-capture estimates; the

species-by-chemical interaction analysis; the out-of-vocabulary re-aggregation; the concentration-fold interpretation; and the comparison of one-hot identity with the species mean offset.

List of abbreviations. BH-FDR: Benjamini–Hochberg false discovery rate; CASRN: Chemical Abstracts Service registry number; d.o.f.: degrees of freedom; dd: difference-in-differences; DEB: dynamic energy budget; D-MPNN: directed message-passing neural network; EC₅₀/LC₅₀: median effect/lethal concentration; NCBI: National Center for Biotechnology Information; OOF: out-of-fold; OOV: out-of-vocabulary; RMSE: root-mean-square error; TOST: two one-sided tests.

Declarations.

- **Funding.** The authors did not receive any funding for this study.
- **Competing interests.** The authors declare that they have no competing interests.
- **Ethics approval and consent to participate.** Not applicable.
- **Consent for publication.** Not applicable.
- **Use of large language models.** A large language model was used as described in Section 3.11. It did not design the study, generate or analyse the data, or determine any scientific claim, and all content was verified by the authors.
- **Availability of data and materials.** The datasets supporting the conclusions of this article are available in the Zenodo repository [29], <https://doi.org/10.5281/zenodo.22245340>. The ADORE benchmark, including the intermediate processed table analysed here, is available under CC-BY 4.0 [5], and ECOTOX is distributed by the US EPA as a public ASCII export that can be downloaded without registration. Because that export is replaced quarterly and superseded releases are not served from the ECOTOX site, the aggregated ECOTOX-external table and the species-level coverage table underlying Figure 3 and Table 8 are the tables deposited there under CC0 1.0. We do not redistribute the raw ECOTOX records themselves: the full export is large, its authoritative copy is the one distributed

736 by the US EPA, and the aggregated tables deposited here are what the analyses
737 reported in this paper require. The code repository at [https://github.com/Pluswick/
738 species-representation-toxicity](https://github.com/Pluswick/species-representation-toxicity)[30], released under the Apache License 2.0, provides
739 the analysis library and pipeline, the pre-registration, the frozen equivalence mar-
740 gins, the full gatekeeping enumerations, the figure-generation scripts, and the audit
741 artefacts, so that every reported statistic can be recomputed from these three tables.
742 Because NCBI taxonomy is versioned, the NCBI taxonomy condition is fixed by
743 the taxonomy snapshot archived with the code, not by re-resolving against a live
744 database.

- 745 • **Authors’ contributions.** SK: conceptualisation, methodology, software, formal
746 analysis, writing – original draft. HC: conceptualisation, methodology, software,
747 validation, data curation, visualisation. KJ: formal analysis, writing – review and
748 editing. SP: investigation, writing – review and editing. DJ: supervision, project
749 administration. All authors read and approved the final manuscript.
- 750 • **Acknowledgements.** Not applicable.

751 References

- 752 [1] Sheffield TY, Judson RS (2019) Ensemble QSAR modeling to predict multispecies
753 fish toxicity lethal concentrations and points of departure. *Environmental Science
754 & Technology* 53(21):12793
- 755 [2] Zubrod JP, Galic N, Vaugeois M, et al (2023) Physiological variables in machine
756 learning QSARs allow for both cross-chemical and cross-species predictions.
757 *Ecotoxicology and Environmental Safety* 263:115250
- 758 [3] Forastiere M, Vijver MG, Posthuma L, et al (2026) Bridging chemical structure
759 and genetic similarity: a Bio-QSAR model for cross-species toxicity predictions.
760 *Ecotoxicology and Environmental Safety* 314:120057

- 761 [4] Viljanen M, Minnema J, Wassenaar PNH, et al (2023) What is the ecotoxicity
762 of a given chemical for a given aquatic species? predicting interactions between
763 species and chemicals using recommender system techniques. SAR and QSAR in
764 Environmental Research 34(10):765
- 765 [5] Schür C, Gasser L, Perez-Cruz F, et al (2023) A benchmark dataset for machine
766 learning in ecotoxicology. Scientific Data 10:718
- 767 [6] Hooker G, Mentch L, Zhou S (2021) Unrestricted permutation forces extrapola-
768 tion: variable importance requires at least one more model, or there is no free
769 variable importance. Statistics and Computing 31:82
- 770 [7] Schuirmann DJ (1987) A comparison of the two one-sided tests procedure and the
771 power approach for assessing the equivalence of average bioavailability. Journal
772 of Pharmacokinetics and Biopharmaceutics 15:657–680
- 773 [8] Lakens D (2017) Equivalence tests: a practical primer for t tests, correlations,
774 and meta-analyses. Social Psychological and Personality Science 8:355–362
- 775 [9] Praski M, Adamczyk J, Czech W (2025) Benchmarking pretrained molecu-
776 lar embedding models for molecular representation learning. arXiv preprint
777 arXiv:2508.06199
- 778 [10] Ahlmann-Eltze C, Huber W, Anders S (2025) Deep-learning-based gene pertur-
779 bation effect prediction does not yet outperform simple linear baselines. Nature
780 Methods 22:1657
- 781 [11] Posthuma L, Price OR, Viljanen M (2025) Improving the ecotoxicological haz-
782 ard assessment of chemicals by pairwise learning. Environmental Science &
783 Technology 59(31):16250

- 784 [12] Wu J, D'Ambrosi S, Ammann L, et al (2022) Predicting chemical hazard across
785 taxa through machine learning. *Environment International* 163:107184
- 786 [13] Schlender T, Viljanen M, van Rijn JN, et al (2023) The bigger fish: a comparison
787 of meta-learning QSAR models on low-resourced aquatic toxicity regression tasks.
788 *Environmental Science & Technology* 57(46):17818
- 789 [14] Kooijman SALM (2010) *Dynamic Energy Budget Theory for Metabolic Organi-*
790 *sation*, 3rd edn. Cambridge University Press, Cambridge
- 791 [15] Zubrod JP, Galic N, Vaugeois M, et al (2024) Bio-QSARs 2.0: unlocking a
792 new level of predictive power for machine learning-based ecotoxicity predictions
793 by exploiting chemical and biological information. *Environment International*
794 186:108607
- 795 [16] Guénard G, von der Ohe PC, de Zwart D, et al (2011) Using phylogenetic infor-
796 mation to predict species tolerances to toxic chemicals. *Ecological Applications*
797 21:3178
- 798 [17] Guénard G, von der Ohe PC, Walker SC, et al (2014) Using phylogenetic
799 information and chemical properties to predict species tolerances to pesticides.
800 *Proceedings of the Royal Society B* 281:20133239
- 801 [18] Yang Y, Yang Y, Pan X, et al (2026) Multimodal integration of chemical and bio-
802 logical descriptors for cross-species prediction of fish acute toxicity. *Environmental*
803 *Science & Technology* 60(3):2556
- 804 [19] Gasser L, Schür C, Perez-Cruz F, et al (2024) Machine learning-based prediction
805 of fish acute mortality: implementation, interpretation, and regulatory relevance.
806 *Environmental Science: Advances* 3(8):1124

- 807 [20] Yuan Y, Shaw S, Tang X, et al (2025) Cross-chemical and cross-species toxicity
808 prediction: benchmarking and a novel 3D-structure-based deep learning model.
809 bioRxiv 2025.11.24.690199
- 810 [21] Rücker C, Rücker G, Meringer M (2007) y-randomization and its variants in
811 QSPR/QSAR. *Journal of Chemical Information and Modeling* 47:2345
- 812 [22] Breiman L (2001) Random forests. *Machine Learning* 45:5–32
- 813 [23] Robinson AP, Froese RE (2004) Model validation using equivalence tests.
814 *Ecological Modelling* 176:349–358
- 815 [24] Demšar J (2006) Statistical comparisons of classifiers over multiple data sets.
816 *Journal of Machine Learning Research* 7:1–30
- 817 [25] Gilmer J, Schoenholz SS, Riley PF, et al (2017) Neural message passing for
818 quantum chemistry. In: *Proceedings of the 34th International Conference on*
819 *Machine Learning, Proceedings of Machine Learning Research*, vol 70. PMLR, pp
820 1263–1272
- 821 [26] Yang K, Swanson K, Jin W, et al (2019) Analyzing learned molecular represen-
822 tations for property prediction. *Journal of Chemical Information and Modeling*
823 59(8):3370
- 824 [27] Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undis-
825 closed flexibility in data collection and analysis allows presenting anything as
826 significant. *Psychological Science* 22:1359–1366
- 827 [28] OECD (2007) Guidance document on the validation of (Q)SAR models (GD69).
828 OECD Series on Testing and Assessment, no 69

- 829 [29] Kim S (2026) ECOTOX-external set and species coverage table for a study of
830 species representation in aquatic acute-toxicity prediction. Zenodo, <https://doi.org/10.5281/zenodo.22245340>, data set
- 832 [30] Kim S (2026) species-representation-toxicity: analysis pipeline, pre-
833 registration, and audit artefacts. [https://github.com/Pluswick/](https://github.com/Pluswick/species-representation-toxicity)
834 [species-representation-toxicity](https://github.com/Pluswick/species-representation-toxicity), accessed 3 Sept 2026

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.pdf](#)