Entomology and Applied Science Letters
Volume 13, Issue 1, Page No: 61-74

Available Online at: www.easletters.com

ISSN No: 2349-2864

Foundation Models for Insect Molecular Biology Must Learn across Genomes, Transcriptomes, Proteomes, Phenotypes, Behaviour, and Chemical Space

Sarah Jensen[1]*, Pieter Vos[2], Lars Olsen[3], Klaus Weber[4]

[1]Department of Computational Biology and AI-Driven Genomics, Faculty of Biosciences, University of Copenhagen, Copenhagen, Denmark.
[2]Department of Multimodal Data Integration and Machine Learning, Faculty of Engineering, Wageningen University, Wageningen, Netherlands.
[3]Department of Bioinformatics and Evolutionary Systems, Faculty of Science, University of Queensland, Brisbane, Australia.
[4]Department of Molecular Systems Biology, Faculty of Life Sciences, University of Bonn, Bonn, Germany.

ABSTRACT

Insect biology increasingly depends on heterogeneous molecular, organismal, and environmental observations that cannot be interpreted reliably through isolated analytical models. Genome assemblies, single-cell transcriptomes, proteomic and metabolomic measurements, chemical structures, images, behavioural trajectories, and ecological records describe different biological objects at different spatial and temporal scales. Yet current artificial-intelligence applications in entomology are commonly developed around one modality, one species, one laboratory setting, or one narrowly defined prediction task. This architecture article addresses the resulting representational gap by proposing a non-validated multimodal foundation-model structure for insect molecular biology. The approach organizes evidence-supported computational capabilities into modality-specific encoders, biologically conditioned alignment layers, shared and modality-private representations, provenance and uncertainty controls, and task-specific outputs. The synthesis integrates genomic and transcriptomic representation, proteomic and metabolomic learning, chemical-space modelling, computer vision, behavioural time series, ecological context, pretraining, transfer, and biological alignment. The strongest defensible conclusion is that broadly pretrained models may support reusable insect representations only when modality-specific meaning, taxonomy, life stage, tissue, environment, measurement process, and uncertainty remain explicit. Multimodal scale is therefore not equivalent to biological understanding; pretraining performance is not downstream scientific validity; representational alignment is not causal mechanism; and benchmark performance is not evidence of transfer across species, modalities, or environments. Major limitations include taxonomic imbalance, incomplete molecular annotation, sparse cross-modal observations, laboratory–field divergence, leakage, reference-atlas bias, and insufficient biologically independent validation. The central implication is that insect foundation models should be developed as auditable, context-conditioned scientific infrastructures whose outputs can be rejected, qualified, or experimentally tested rather than treated as universal biological explanations or deployment-ready digital twins.

Keywords: Predictive entomology, Multimodal foundation models, Artificial intelligence, Machine learning, Computer vision, Biological representation learning.

HOW TO CITE THIS ARTICLE: Jensen S, Vos P, Olsen L, Weber K. Foundation Models for Insect Molecular Biology Must Learn across Genomes, Transcriptomes, Proteomes, Phenotypes, Behaviour, and Chemical Space. Entomol Appl Sci Lett. 2026;13(1):61-74. https://doi.org/10.51847/1GGg63ErEt

Corresponding author: Sarah Jensen
E-mail ✉ sarah.jensen@bio.ku.dk
Received: 10/10/2025
Accepted: 27/02/2026

INTRODUCTION

Insect molecular biology is entering a period in which the volume and diversity of available data increasingly exceed the organizing capacity of conventional task-specific analysis. Genome assemblies now span a growing range of insect lineages, while transcriptomic atlases, mass-

spectrometric profiles, chemical assays, imaging systems, behavioural platforms, and environmental monitoring generate additional views of biological state. These sources are not interchangeable. A genomic sequence describes inherited molecular structure; a transcriptome records condition-dependent expression; a proteome reflects translation, degradation, localization, and trafficking; and an image or movement trajectory records observable organismal outcomes. Their joint value therefore depends on preserving their distinct biological meanings. Insect genome resources have expanded rapidly, but taxonomic imbalance and assembly-quality heterogeneity mean that sequence scale alone does not create a representative biological substrate [1].

Foundation models offer a possible computational response because they are pretrained on broad data collections and subsequently adapted to multiple tasks. In principle, such a model could reduce dependence on independently engineered pipelines for species identification, molecular annotation, functional prediction, phenotype recognition, behavioural classification, or ecological forecasting. In practice, however, the foundation-model label can conceal important differences among corpus size, self-supervised objectives, modality coverage, transfer procedures, and scientific validation. A system trained on many records may still learn dominant species, laboratory protocols, database conventions, or technical artefacts rather than transferable insect biology. Foundation-model paradigms suggest that self-supervised representations can be reused across modalities and tasks, although that proposition remains unvalidated for insect biology [2].

The central problem is consequently not the absence of algorithms capable of combining data. It is the absence of an insect-specific representational logic that determines what should be combined, what must remain separate, which contextual variables condition interpretation, and what evidence is required before an aligned representation can support a biological claim. Multimodal integration methods show that complementary measurements can improve state resolution when modality-specific structure is retained. Nevertheless, a joint latent space can also erase biologically meaningful differences, absorb batch effects, or create numerical proximity between observations that are not mechanistically connected. Multimodal integration can improve cellular-state resolution by retaining modality-specific neighbourhood structure rather than collapsing all measurements into one undifferentiated representation [3].

This article therefore develops an original, explicitly non-validated foundation-model architecture for insect molecular biology. Its scope extends from genomic and transcriptomic representations to proteins, metabolites, chemical space, phenotypes, behaviour, ecological context, pretraining, transfer, and biological alignment. The proposed contribution is not a new experiment, benchmark, field evaluation, regulatory framework, or deployment programme. It is an evidence-grounded scholarly structure that specifies candidate components, relations, boundary conditions, failure modes, validation requirements, and decision points. Transfer-learning evidence indicates why such reuse may be scientifically valuable under limited labelled data, while simultaneously demonstrating why successful adaptation in one biological domain cannot establish cross-species entomological validity. Large-scale transcriptomic pretraining can improve limited-data adaptation, but successful transfer within human network biology cannot be assumed to extend across insect species and environments [4].

Why insect biology requires multimodal foundation models

A foundation model for insect biology is justified only if it addresses a biological problem that separate specialist models cannot resolve adequately. The proposed synthesis begins with the fragmentation of insect evidence across species, repositories, molecular layers, measurement technologies, and observational scales. Genomic data may be available for one lineage, behavioural video for another, and ecological time series for a third, with little direct pairing among them. The architecture must therefore learn reusable structure without implying that missing relationships have been empirically observed. It also requires provenance-aware ingestion because public availability does not ensure consistent

annotation, licensing, assembly quality, or taxonomic coverage. Open arthropod genomics has expanded the available sequence substrate, yet uneven coverage and inconsistent accessibility still constrain comparative learning across insect lineages [5].

Biological context must enter the representation as an active conditioning layer rather than descriptive metadata appended after prediction. Insect genomes reflect lineage history, symbiosis, population structure, and environmental selection, while transcriptomic states vary among cells, tissues, ages, sexes, developmental stages, and challenges. A chromosome-level leafhopper genome demonstrates that insect genomic representation is inseparable from ecological history and microbial symbiosis, even though one species cannot define a general architecture [6]. At finer resolution, embryonic cell states illustrate why one aggregate expression vector cannot represent a developing organism. Single-cell resolution in the Drosophila embryo shows that bulk transcriptomic summaries can obscure spatially organized developmental states that a biologically aligned model must preserve [7]. The proposed model must consequently distinguish inherited sequence, dynamic molecular state, developmental identity, and ecological conditioning rather than allowing one modality to stand in for the others.

Phenotypic and behavioural modalities are equally necessary because molecular measurements do not directly specify what an insect looks like, how it moves, which resources it uses, or how it interacts with conspecifics, hosts, vectors, predators, or environments. Imaging and video can scale detection, identification, morphology, pose estimation, and behavioural observation, but visual models are particularly vulnerable to background cues, imaging hardware, annotation conventions, occlusion, and taxonomic imbalance. Their outputs must therefore retain confidence estimates and acquisition context and must be validated beyond the cameras, sites, seasons, and taxa represented during training. Entomological computer vision can scale observations of identity, abundance, traits, and behaviour, but its value depends on representative labels and validation beyond the acquisition environment [8].

Ecological information completes the proposed multimodal structure by defining the conditions under which molecular and organismal states are expressed. Temperature, habitat, land use, resource availability, microbial exposure, chemical exposure, geography, season, and sampling effort can alter both the observed biological state and the probability that it is recorded. These variables cannot be reduced to a universal environmental embedding without risking the conversion of spatial or temporal confounding into apparent biological structure. Insect population change reflects interacting pressures whose effects vary across taxa, habitats, and time, making ecology a conditioning modality rather than decorative metadata [9]. Accordingly, the proposed architecture requires modality-specific encoders, explicit contextual modifiers, shared and private representations, uncertainty and rejection mechanisms, and validation across biologically independent units. Its outputs should be treated as conditional representations or hypotheses whose scientific meaning depends on the data, scale, and validation supporting them.

The proposed components, evidence bases, boundary conditions, failure modes, and validation requirements are organized in **Table 1**.

Table 1. Why Insect Biology Requires Multimodal Foundation Models: Components, Evidence Basis, Relations, Boundary Conditions, Failure Modes, and Validation Requirements

Proposed component	Purpose	Evidence basis	Relation or mechanism	Input or precondition	Expected output	Boundary condition or failure mode	Validation requirement
Insect foundation-model backbone	Learn reusable representations across related insect tasks	Broad pretraining may support adaptation, but insect corpora remain	Self-supervised learning within curated biological modalities,	Versioned, licensed, provenance-rich data from diverse insect lineages	Reusable base representations with calibrated uncertainty	Corpus size may amplify dominant-species, repository, or protocol bias	Contamination-controlled tests on independently held-out taxa and tasks

		taxonomically uneven	followed by task-specific adaptation				
Genomic representation	Encode inherited sequence, genome structure, annotation, and lineage context	Insect assemblies differ in contiguity, completeness, and taxonomic coverage	Long-context sequence encoding conditioned on assembly quality and orthology	Genome sequence, annotations, chromosome context, and quality metadata	Genome embeddings linked to explicit lineage and provenance	Availability is not comparability; fragmented assemblies can create artefacts	Hold out species, genomic regions, and sequencing technologies
Transcriptomic representation	Preserve dynamic cell, tissue, developmental, and challenge states	Single-cell insect atlases resolve states obscured by bulk measurements	Hierarchical encoding of genes, cells, tissues, stages, and conditions	Expression matrices, cell ontology, tissue, stage, sex, and treatment context	Contextual cell- and tissue-state representations	Batch effects or reference labels may be mistaken for stable biology	Cross-study mapping, novelty detection, and orthogonal marker confirmation
Proteomic representation	Separate protein sequence, structure, abundance, localization, and trafficking	Protein measurements can differ from local transcription and may reflect inter-organ movement	Typed relations connect protein sequence, predicted structure, tissue origin, abundance, and function	Proteomic data, sequence, structure confidence, tissue provenance, and matched context	Protein representations with explicit origin and confidence	Detection or structural similarity is not equivalent to physiological function	Cross-family and cross-species testing with functional confirmation
Metabolomic and chemical-space representation	Represent partially identified metabolites, xenobiotics, and molecular relations	Chemical evidence often supports probabilistic classes rather than exact identities	Spectral and molecular encoders connect compounds to chemical ontologies and assay context	Raw spectra or structures, platform metadata, biological context, and reference standards	Calibrated chemical-class or molecular-relation predictions	Chemical similarity or class assignment is not mechanism, efficacy, or safety	Instrument-, scaffold-, and assay-held-out evaluation with targeted confirmation
Phenotype vision encoder	Capture morphology, traits, body configurations, and observable condition	Computer vision can scale insect detection and trait measurement	Images or video are converted into confidence-aware morphology and pose features	Representative annotations, camera metadata, taxonomic labels, and acquisition conditions	Morphological and pose representations	Background, equipment, occlusion, or species morphology may drive predictions	Independent cameras, field sites, annotators, and taxa
Behavioural sequence encoder	Model longitudinal movement, interactions, and responses	Automated observation can produce time-resolved behavioural data	Temporal models and interaction graphs link identities, poses, events, and interventions	Identity-linked trajectories, timestamps, environmental context, and intervention provenance	Behavioural sequences and interaction representations	Motion or proximity is not equivalent to behavioural mechanism or intent	Longitudinal validation, blinded annotation, perturbation, and field comparison
Ecological-context adapter	Condition molecular and organismal predictions on place, time, habitat, and sampling	Insect outcomes vary across taxa, ecosystems, pressures, and observation designs	Environmental variables modify rather than replace modality-specific representations	Site, season, climate, habitat, land use, exposure, and sampling-effort metadata	Context-specific predictions and uncertainty estimates	Spatial or temporal correlation may be mistaken for causal environmental influence	Leave-site, leave-year, and leave-ecosystem-out validation
Provenance, uncertainty, and	Preserve traceability and prevent	Heterogeneous data quality and	Every representation retains	Versioned data, computationa	Auditable predictions, uncertainty,	Imputed or aligned representatio	Independent reproducibility, calibration,

abstention layer	unsupported interpretation	missing modalities limit defensible inference	source, transformation, missingness, confidence, and out-of-distribution indicators	l records, quality metrics, and predefined rejection rules	or refusal to predict	ns may be reported incorrectly as observations	leakage, and missing-modality audits

Figure 1 shows conceptual synthesis designed to show the multimodal learning architecture for insect foundation models. Arrows and grouping indicate proposed or evidence-supported relations, not measured effect sizes or universal causal pathways. Relations must be qualified in the text with the approved references.

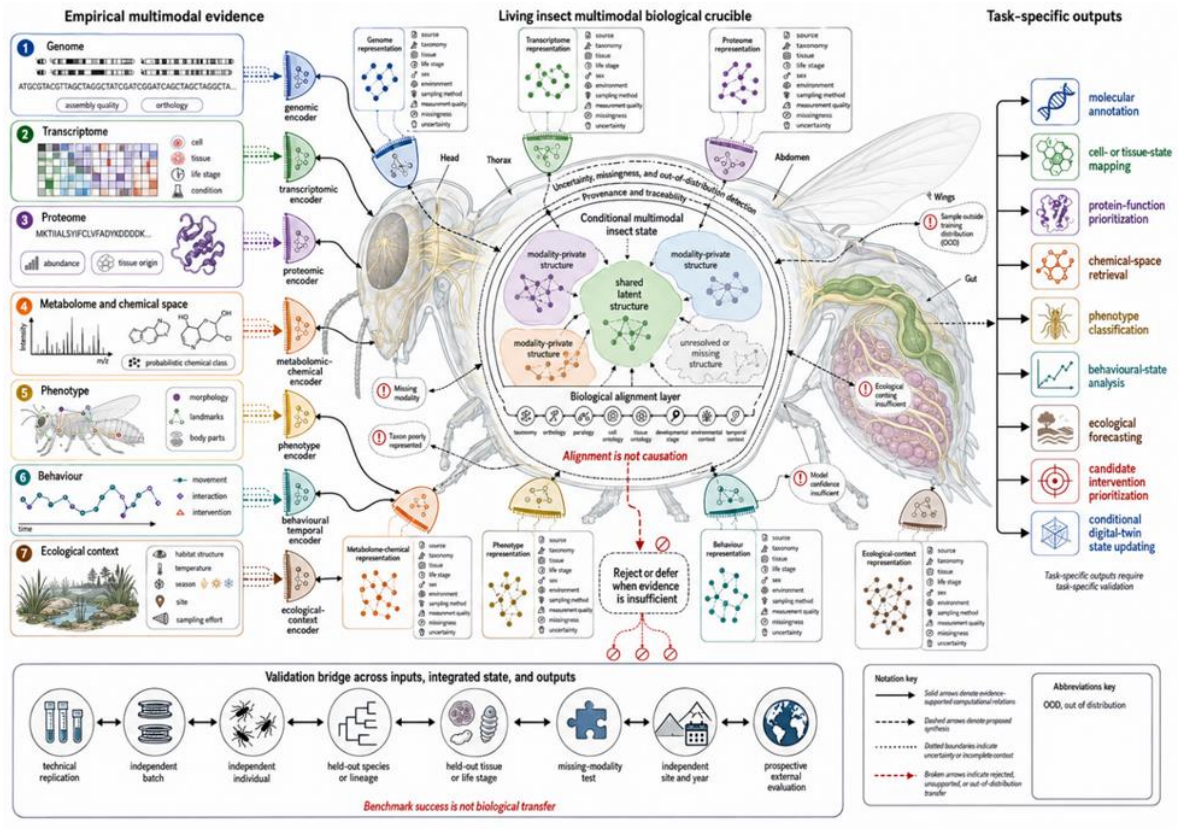

Figure 1. The multimodal learning architecture for insect foundation models

Abbreviations key: OOD, out of distribution.
Uncertainty notation: Solid arrows denote evidence-supported computational relations; dashed arrows denote proposed integrative relations; bordered warning symbols denote uncertainty, missingness, or possible failure.

Genomic and transcriptomic representation

Genomic representation must capture more than short sequence motifs. Regulatory relationships may depend on long-range sequence context, chromosome organization, gene-family structure, assembly quality, and lineage-specific evolutionary history. Long-context architectures developed in regulatory genomics demonstrate that predictive information can be distributed across extensive sequence windows, but their assumptions, training organisms, and regulatory distances cannot be transferred automatically to insects. Sequence encoders can benefit from long-range regulatory context, although architectures calibrated to mammalian genomes require explicit revalidation before insect use [10]. Nucleotide pretraining further suggests that a shared sequence model may be adaptable to several genomic tasks, yet insect genomes introduce major differences in genome size, repeat composition, chromosome organization, annotation completeness, and lineage-specific gene expansion. Nucleotide language models

show that broad pretraining can support varied genomic tasks, but insect deployment requires lineage-aware tokenization and held-out-species tests [11]. The appropriate genomic component is therefore a long-context encoder conditioned on taxonomy, assembly quality, orthology, paralogy, and annotation provenance rather than a universal sequence representation detached from its biological source.

Transcriptomic representation requires a different hierarchy because expression is dynamic, compositional, and context-dependent. A gene-expression profile is shaped by cell type, tissue composition, developmental stage, age, sex, physiological state, exposure, infection, and sampling time. A model that collapses these distinctions may generate a stable numerical embedding while misrepresenting the biological object that was measured. The aging Drosophila brain contains cell-type-specific transcriptional trajectories that would be lost if tissue, age, and cell identity were treated as interchangeable labels [12]. Accordingly, the proposed transcriptomic encoder should preserve nested biological identifiers and distinguish directly observed labels from transferred or inferred labels. It should also represent uncertainty where reference atlases are incomplete, where marker genes are ambiguous, or where technical batch is associated with biological condition.

Mosquito immune-cell studies illustrate both the promise and the boundary of cross-species transcriptomic alignment. Mosquito single-cell profiling reveals immune-cell states and differentiation structure that require species, cell-state, and challenge context to remain explicit in the representation [13]. Independent profiling provides convergent evidence that mosquito hemocytes are heterogeneous, but it also shows that subtype boundaries and differentiation signatures depend on assay design, sampled conditions, and validation strategy. Independent mosquito hemocyte profiling further supports immune-cell heterogeneity while showing that cell labels and differentiation signatures remain conditional on assay and biological context [14]. A foundation model should therefore align transcriptomic states through confidence-scored orthology and ontology relations while retaining a novelty-detection option. Similarity between expression profiles may support a hypothesis of correspondence, but it cannot by itself establish cell homology, shared function, differentiation mechanism, or causal immune activity. Genomic and transcriptomic alignment must remain connected to, but analytically distinct from, protein abundance, metabolism, phenotype, and organismal outcome.

Proteomic, metabolomic, and chemical-space learning

Proteomic learning must separate at least five constructs: amino-acid sequence, predicted structure, measured abundance, tissue origin, and biological function. These constructs are related, but none is a universally reliable substitute for another. Integrated insect evidence shows why provenance and spatial origin are essential. Multi-omic mapping of Drosophila secretomes shows that protein abundance must be linked to tissue origin and trafficking rather than treated as a direct proxy for local transcription [15]. The proposed proteomic encoder should consequently represent proteins as linked objects that retain sequence, structural confidence, post-translational or abundance measurements, tissue and cellular provenance, and evidence for movement between organs. This structure would permit cross-modal reasoning while preventing a circulating protein from being attributed automatically to the tissue in which it was detected.

Large protein-sequence models provide an enabling representation layer because they can encode statistical regularities associated with structure and function. Protein language modelling can recover structure- and function-related information from sequence, but those embeddings still require insect-specific functional and interaction validation [16]. For insect biology, this requirement is especially important where rapid gene-family expansion, paralogy, symbiotic interaction, venom or salivary specialization, detoxification, or lineage-specific proteins weaken simple transfer from broadly represented families. A sequence embedding may prioritize candidate relations, support retrieval, or guide downstream experiments, but it cannot establish expression, localization, molecular interaction, physiological effect, ecological fitness, or safety. Protein representations should therefore expose family novelty and out-of-distribution status and should

be linked to experimentally observed functions only through traceable evidence.

Metabolomic and chemical-space learning introduce additional uncertainty because many detected compounds lack complete structural identification, and measured profiles depend on extraction, instrument platform, ionization, diet, sex, stage, infection, microbiota, and environment. Mass-spectral learning can classify unknown compounds at the chemical-class level, but class assignment must not be reported as exact structural identification or biological mechanism [17]. This makes probabilistic chemical ontologies more defensible than forced exact labels, particularly when spectra are unmatched to authenticated standards. In mosquitoes, the interpretive problem is compounded by strong biological context dependence. Mosquito metabolomic profiles vary with life stage, diet, infection, sex, and analytical platform, so chemical representations require explicit biological and technical context [18]. The proposed architecture should therefore connect spectra, candidate structures, chemical classes, proteins, assays, and phenotypes through typed relations carrying uncertainty and provenance. Chemical similarity may guide hypothesis generation, but it does not establish shared target activity, efficacy, toxicity, ecological consequence, or causal metabolic mechanism.

Phenotypes, behaviour, ecology, and environmental context

Molecular representations become biologically useful only when they can be related cautiously to observable organismal states. Images and video can encode morphology, body configuration, locomotion, feeding, courtship, host seeking, social interaction, and responses to experimental or environmental change. Markerless pose-estimation systems demonstrate that fine-grained body-part trajectories can be extracted with relatively limited manual annotation, but their predictions remain dependent on camera geometry, occlusion, morphology, training labels, and recording conditions. Markerless pose estimation makes fine-grained behavioural phenotyping scalable, but the learned landmarks remain conditional on annotation quality and imaging conditions [19]. A foundation model should therefore preserve keypoint confidence, image provenance, acquisition conditions, and taxonomic context rather than converting visual outputs directly into biological interpretations.

Behavioural representation requires temporal continuity and reliable identity tracking. Multi-animal systems can represent individual movement and interaction structure, but crowding, identity switching, and unfamiliar environments can weaken apparent behavioural patterns. Multi-animal tracking extends phenotype representation to interaction structure, although identity continuity and social interpretation remain vulnerable to occlusion and domain shift [20]. Longitudinal platforms add another layer by recording repeated behavioural states and experimentally controlled interventions. High-throughput Drosophila ethomics demonstrates the value of longitudinal behavioural streams, while also exposing the boundary between controlled laboratory behaviour and field expression [21]. Accordingly, pose, movement, interaction, and behaviour should remain separate analytical constructs: a coordinate is not an action, proximity is not a social mechanism, and laboratory activity is not automatically an ecologically expressed phenotype.

Ecological context defines where, when, and under which sampling conditions molecular and behavioural states are observed. Habitat, season, climate, resource availability, land use, exposure history, and sampling effort can alter both insect biology and the probability of detection. Contrasting terrestrial and freshwater abundance trends show why ecosystem type, geography, and sampling design must condition any population-level representation [22]. Ecological variables should therefore modify predictions through explicit spatial and temporal context rather than being compressed into an unqualified environmental score. Generalization must be tested across independent sites, years, ecosystems, cameras, taxa, and behavioural settings. The evidence dimensions and interpretive boundaries for phenotypes behaviour ecology and environmental context are summarized in **Table 2**.

Table 2. Phenotypes, Behaviour, Ecology, and Environmental Context: Data Inputs, Analytical Tasks, Biological Representations, Validation, Explainability, Generalization, and Decision Boundaries

Analytical component or task	Input data	Model or representation	Expected output	Validation requirement	Explainability or interpretation need	Generalization risk	Decision boundary
Morphology and pose estimation	Images, video, body-part annotations, camera metadata	Confidence-aware visual and keypoint representation	Body-part locations and morphological features	Independent cameras, annotators, taxa, backgrounds, and field conditions	Show keypoint confidence and image regions influencing prediction	Occlusion, equipment cues, morphology, and background shift	Pose coordinates are not behaviour, condition, or fitness
Multi-individual tracking	Video, identity labels, trajectories, interaction events	Temporal multi-instance representation	Identity-linked movement and interaction sequences	Crowded scenes, independent densities, species, and recording systems	Expose identity confidence, track breaks, and uncertainty	Identity swaps and unfamiliar social configurations	Proximity or synchronized motion is not a social mechanism
Longitudinal ethomics	Continuous observations, timestamps, interventions, environmental conditions	Behavioural time-series representation	Repeated activity states and intervention-linked trajectories	Repeated experiments, blinded labels, perturbation, and field comparison	Retain observable events and intervention provenance	Laboratory apparatus and constrained environments	Laboratory behaviour is not automatically field behaviour
Ecological conditioning	Site, habitat, climate, season, land use, exposure, and sampling effort	Spatial-temporal context representation	Context-specific predictions and uncertainty	Leave-site, leave-year, and leave-ecosystem-out testing	Identify influential contextual variables and missing covariates	Spatial leakage, uneven sampling, and unmeasured pressures	Ecological association or trend is not causal attribution
Molecular–phenotype linkage	Molecular profiles paired with images or behavioural records	Typed cross-modal correspondence	Candidate relations between molecular state and observed phenotype	Matched individuals, tissues, times, and independent interventions	Distinguish observed pairing from inferred alignment	Different individuals or time points may be aligned spuriously	Representation similarity is not a molecular mechanism
Decision-support output	Phenotypic, behavioural, ecological, and uncertainty representations	Task-specific calibrated head with abstention	Conditional classification, forecast, prioritization, or rejection	Independent operational context and prospective evaluation	Report uncertainty, coverage, provenance, and failure conditions	Benchmark data may not represent target environments	Model output is not field effectiveness or management readiness

Pretraining, transfer, and biological alignment

Pretraining should be understood as learning reusable statistical structure within large data collections, not as demonstrating biological validity. Reference-atlas transfer provides one model for adapting representations to new cellular datasets, but it can also force unfamiliar states into existing categories. Reference-atlas transfer can organize new cellular data efficiently, but novelty detection is essential to prevent genuinely new insect states from being forced into inherited labels [23]. For insect applications, donor and target distributions should be characterized by taxonomy, phylogenetic distance, tissue, life stage, sex, environment, assay technology, and sampling design. Adaptation should permit rejection when target observations are insufficiently represented rather than assigning every record to the nearest known class.

Generative single-cell pretraining shows how one representation may support annotation, integration, and perturbation-related tasks. Generative single-cell pretraining supports multiple downstream tasks, yet insect use must test whether token semantics survive changes in genome annotation, cell ontology, and species [24]. Similar caution applies to protein representations. Accurate protein-structure prediction can enrich molecular representations, but structure must remain analytically separate from function, interaction, and insect fitness [25]. A biologically aligned architecture should therefore retain the distinctions among predicted structure, measured abundance, tissue localization, interaction evidence, and organismal consequence. Alignment may identify correspondence or shared statistical structure, but it does not establish causal mechanism or functional equivalence.

Generative protein models further illustrate why model capability and scientific justification must be evaluated separately. Protein language models can generate functional sequences within learned families, although family-specific activity does not establish ecological safety or general insect utility [26]. The proposed insect architecture consequently separates pretraining objectives, transfer procedures, alignment operations, and downstream validation. Pretraining may include masked reconstruction, contrastive learning, generative prediction, or graph completion within individual modalities. Transfer may use adapters or task-specific heads. Biological alignment should use taxonomy, orthology, ontology, tissue, life stage, and context information while preserving modality-private variation. Each stage requires its own success criterion, stress test, uncertainty estimate, and prohibited inference.

Proposed foundation-model architecture

The proposed architecture begins with independent encoders for genomic sequence, transcriptomic state, protein sequence and structure, proteomic abundance, metabolomic spectra, chemical structures, phenotype imagery, behavioural trajectories, and ecological context. These encoders should not be merged through unrestricted early fusion because some variation is shared across modalities while other variation remains biologically specific. A latent-factor framework demonstrates the value of separating common from modality-specific information. The proposed architecture therefore preserves both shared and modality-specific latent factors, following evidence that forcing all signals into one common space can erase informative variation [27]. Each encoded observation also retains taxonomy, tissue, life stage, sex, environment, sampling method, data quality, and provenance.

The fusion layer is probabilistic and uncertainty-aware. It should estimate a shared biological state only where supported while retaining modality-private residuals and explicit missingness. A probabilistic fusion layer should represent measurement noise and posterior uncertainty explicitly rather than returning a single deterministic aligned state [28]. The resulting structure is surrounded by an orthology and ontology graph, a provenance ledger, an out-of-distribution detector, and an abstention mechanism. Task-specific heads may support annotation, retrieval, classification, forecasting, candidate prioritization, or conditional digital-twin state updating, but each output remains limited by its validation evidence. **Figure 2** maps the relationship between pretraining objectives, biological representations, and downstream entomological tasks within the analytical logic developed in this section.

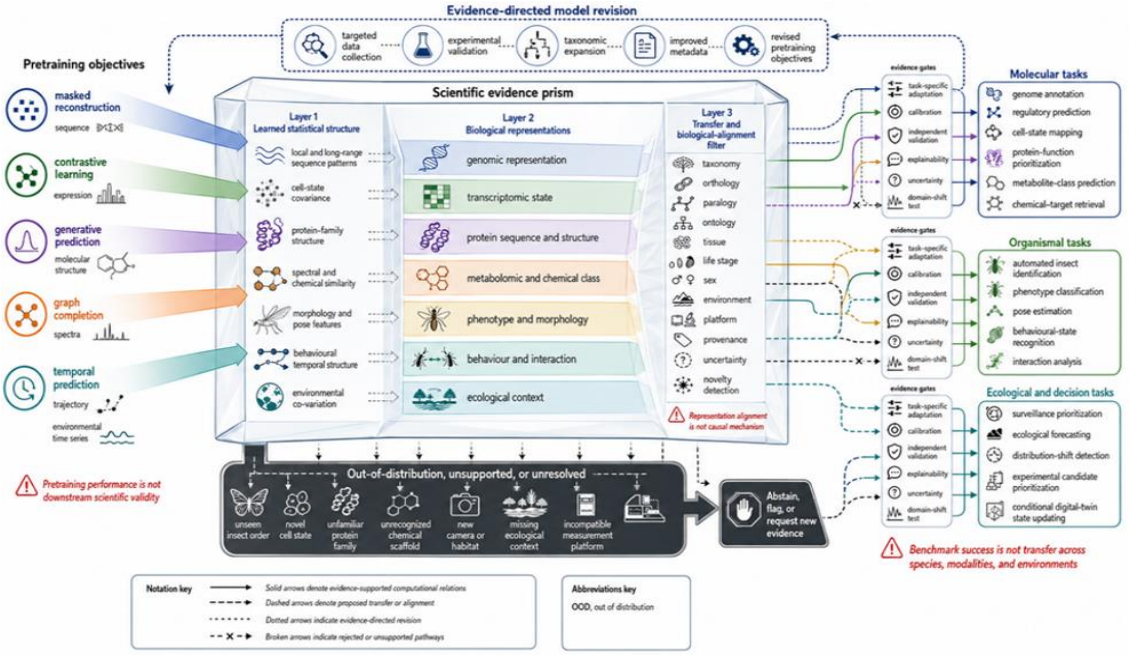

Figure 2. The relationship between pretraining objectives, biological representations, and downstream entomological tasks

Abbreviations key: OOD, out of distribution.

Uncertainty notation: Solid arrows denote evidence-supported computational relations; dashed arrows denote proposed transfers or alignments; warning boundaries mark tasks for which biological, ecological, or external validation is incomplete.

Relational information can be represented through graphs connecting genes, proteins, metabolites, organisms, observations, and environments. Graph-based fusion offers a way to preserve relations among samples and molecular features, but predictive graph structure must not be interpreted automatically as biological mechanism [29]. Missing modalities, sparsity, batch effects, and unpaired measurements must also be treated as central design problems. The architecture must treat missing modalities, batch effects, heterogeneity, scalability, and interpretability as design constraints rather than post hoc corrections [30]. Imputed representations should be marked as inferred, accompanied by posterior uncertainty, and excluded from claims requiring direct observation.

Cross-species alignment is organized through confidence-scored orthology, paralogy, taxonomy, and cell-ontology relations. Cross-species atlas mapping supports an orthology-aware alignment layer, provided that conserved expression is not equated automatically with homologous function [31]. The architecture therefore produces conditional representations rather than a universal insect state. It should support abstention when taxonomic distance, missing modalities, unfamiliar chemical space, novel cell states, or environmental shift exceeds the evidence represented during training. The proposed components, evidence bases, boundary conditions, failure modes, and validation requirements are organized in **Table 3**.

Table 3. Proposed Foundation-Model Architecture: Components, Evidence Basis, Relations, Boundary Conditions, Failure Modes, and Validation Requirements

Proposed component	Purpose	Evidence basis	Relation or mechanism	Input or precondition	Expected output	Boundary condition or failure mode	Validation requirement
Modality-specific encoders	Preserve distinct biological objects before integration	Shared and private variation can coexist across biological modalities	Separate sequence, expression, protein, chemical, visual, behavioural, and ecological encoders	Versioned modality data with quality and context metadata	Modality-specific embeddings	Early fusion may erase meaningful variation	Modality-ablation and independent task validation
Shared and private latent state	Represent convergent evidence without removing modality-specific information	Latent-factor integration separates shared and modality-specific structure	Gated fusion estimates common state while retaining private residuals	Paired or partially paired modalities with missingness indicators	Joint state plus modality-private information	Shared factors may capture batch or confounding	Negative controls, batch holdouts, and orthogonal assays
Probabilistic fusion and missingness layer	Represent noise, incomplete observations, and uncertainty	Joint probabilistic models can separate biological signal from technical variation	Posterior inference integrates observed modalities and marks inferred values	Measurement models, quality metrics, and explicit missingness	Posterior representations, uncertainty, or rejection	Imputation may be mistaken for observation	Masking experiments and genuinely missing-modality tests
Biological relation graph	Preserve relations among genes, proteins, chemicals, samples, and organisms	Graph learning can encode relational structure across heterogeneous data	Typed nodes and edges connect modalities through traceable evidence	Curated relations, provenance, confidence, and negative examples	Relational embeddings and candidate links	Predictive connectivity is not causal mechanism	Independent relation tests and experimental confirmation
Context adapter	Condition outputs on taxonomy, stage, tissue, environment, site, and time	Biological and technical heterogeneity changes representation meaning	Context variables modulate rather than replace core modality representations	Complete contextual metadata	Context-specific predictions and uncertainty	Missing context may produce confident but invalid transfer	Leave-context-out and prospective external validation
Orthology and	Support cross-species	Cross-species atlas mapping	Confidence-scored mappings connect	Orthology, paralogy,	Traceable candidate	Expression similarity may	Phylogenetically blocked tests and

ontology layer	correspondence without assuming equivalence	can identify conserved relationships	genes, cells, tissues, and taxa	taxonomy, cell ontology, and annotation quality	correspondences	reflect convergence or annotation bias	functional confirmation
Provenance and uncertainty ledger	Make predictions auditable and scientifically qualified	Every integration step can introduce transformation, leakage, or uncertainty	Record source, version, preprocessing, model, split, confidence, and inferred status	Reproducible computational environment and data lineage	Auditable output with uncertainty and abstention	Traceability alone does not establish validity	Independent rerun, calibration, and leakage audit
Task-specific heads	Adapt the shared infrastructure to distinct scientific questions	Transfer is task- and distribution-dependent	Separate heads perform annotation, retrieval, forecasting, or prioritization	Defined target labels, claim limits, and independent evaluation data	Task output with calibration or rejection	One benchmark cannot validate all downstream tasks	Task-specific external and prospective validation
Conditional digital-twin updater	Integrate repeated observations into a changing insect-state estimate	Longitudinal multimodal data can support state updating	Time-stamped observations update a conditional representation	Repeated molecular, phenotype, behavioural, and ecological observations	Uncertainty-aware state estimate or forecast	A state estimator is not a validated causal digital twin	Prospective temporal prediction and intervention testing

Validation, bias, and research implications

The first priority is to replace random record-level validation with biologically independent evaluation. Genomic and multimodal observations may share individuals, species, homologous sequences, sampling sites, processing batches, reference atlases, or temporal structure across training and test sets. Validation must split data at biologically independent levels because random records can leak species, lineage, site, batch, or individual information into both training and testing sets [32]. Progress would be demonstrated by preregistered evaluations that hold out entire species, lineages, gene families, tissues, life stages, sites, years, instruments, or modalities according to the intended claim. Performance should be accompanied by calibration, abstention, subgroup uncertainty, and explicit statements of the largest defensible inference.

The second priority is auditable reproducibility and leakage control. Reproducibility requires versioned data, executable code, model specifications, environments, and reporting that make each result traceable to its computational and biological inputs [33]. Such transparency must extend to preprocessing, feature construction, reference atlases, pretrained checkpoints, homology searches, data exclusions, and all transformations performed before splitting. Leakage auditing must extend beyond duplicate records to preprocessing, feature construction, reference atlases, phylogeny, and temporally overlapping sampling [34]. Progress would be shown by independent reruns, contamination reports, data-lineage records, and evaluation designs that remain valid after correlated biological units are removed.

The third priority is representation governance across taxa, geography, life stage, sex, tissue, environment, and measurement platform. Taxonomic bias must be measured explicitly because underrepresented insect lineages can be invisible in aggregate benchmarks while still defining the model's largest deployment failures [35]. Dataset documentation should report which lineages and contexts are represented, which are missing, how licensing and access shape the corpus, and where uncertainty prevents meaningful evaluation. Research investment should prioritize paired multimodal observations from underrepresented insects, independent field datasets, experimentally validated molecular relations, negative results, and prospective transfer studies. These measures would not validate a universal insect model, but they would make failure visible and establish whether a proposed representation transfers beyond the species, modalities, and environments from which it was learned.

CONCLUSION

A scientifically defensible foundation model for insect molecular biology cannot be built by pooling heterogeneous records and treating the

resulting scale as understanding. It must preserve the different biological objects represented by genomes, transcriptomes, proteins, metabolites, chemical structures, phenotypes, behaviours, and ecological observations while making their conditional relations explicit. The proposed architecture therefore combines modality-specific encoders, shared and private representations, biological context adapters, orthology and ontology relations, probabilistic uncertainty, provenance, abstention, and task-specific validation. Its strongest implication is methodological: insect foundation models should be judged by whether they retain biological meaning and withstand independent species, lineage, life-stage, modality, site, and environmental shifts—not by aggregate pretraining or benchmark performance alone. The architecture remains a non-validated scholarly proposal whose relations require prospective testing, experimental confirmation, and transparent governance before it can support consequential scientific or decision applications.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, et al. Long reads are revolutionizing 20 years of insect genome sequencing. Genome Biol Evol. 2021;13(8). doi:10.1093/gbe/evab138
2. Moor M, Banerjee O, Shakeri Hossein Abad Z, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. doi:10.1038/s41586-023-05881-4
3. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87. doi:10.1016/j.cell.2021.04.048
4. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616-24. doi:10.1038/s41586-023-06139-9
5. González VL, Devine AM, Trizna M, Mulcahy DG, Barker KB, Coddington JA. Open access genomic resources for terrestrial arthropods. Curr Opin Insect Sci. 2018;25:91-8. doi:10.1016/j.cois.2017.12.003
6. Vasquez YM, Li Z, Xue AZ, Bennett GM. Chromosome-level genome assembly of the aster leafhopper (Macrosteles quadrilineatus) reveals the role of environment and microbial symbiosis in shaping pest insect genome evolution. Mol Ecol Resour. 2024;24(3). doi:10.1111/1755-0998.13919
7. Karaiskos N, Wahle P, Alles J, Boltengagen A, Ayoub S, Kipar C, et al. The Drosophila embryo at single-cell transcriptome resolution. Science. 2017;358(6360):194-9. doi:10.1126/science.aan3235
8. Høye TT, Ärje J, Bjerge K, Hansen OLP, Iosifidis A, Leese F, et al. Deep learning and computer vision will transform entomology. Proc Natl Acad Sci U S A. 2021;118(2). doi:10.1073/pnas.2002545117
9. Wagner DL, Grames EM, Forister ML, Berenbaum MR, Stopak D. Insect decline in the Anthropocene: Death by a thousand cuts. Proc Natl Acad Sci U S A. 2021;118(2). doi:10.1073/pnas.2023989118
10. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196-203. doi:10.1038/s41592-021-01252-x
11. Dalla-Torre H, Gonzalez L, Mendoza-Revilla J, Lopez Carranza N, Grzywaczewski AH, Oteri F, et al. Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. Nat Methods. 2025;22(2):287-97. doi:10.1038/s41592-024-02523-z
12. Davie K, Janssens J, Koldere D, De Waegeneer M, Pech U, Kreft Ł, et al. A single-cell transcriptome atlas of the aging Drosophila brain. Cell. 2018;174(4):982-98. doi:10.1016/j.cell.2018.05.057

13. Raddi G, Barletta ABF, Efremova M, Ramirez JL, Cantera R, Teichmann SA, et al. Mosquito cellular immunity at single-cell resolution. Science. 2020;369(6507):1128-32. doi:10.1126/science.abc0322
14. Kwon H, Mohammed M, Franzén O, Ankarklev J, Smith RC. Single-cell analysis of mosquito hemocytes identifies signatures of immune cell subtypes and cell differentiation. Elife. 2021;10. doi:10.7554/eLife.66192
15. Bosch JA, Beltran PMJ, Cavers C, LaGraff JT, Melanson R, Singh A, et al. Multi-omic mapping of Drosophila protein secretomes reveals tissue-specific origins and inter-organ trafficking. Nat Commun. 2026;17(1):5425. doi:10.1038/s41467-026-71763-8
16. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15). doi:10.1073/pnas.2016239118
17. Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol. 2021;39(4):462-71. doi:10.1038/s41587-020-0740-8
18. Horvath TD, Dagan S, Scaraffia PY. Unraveling mosquito metabolism with mass spectrometry-based metabolomics. Trends Parasitol. 2021;37(8):747-61. doi:10.1016/j.pt.2021.03.010
19. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21(9):1281-9. doi:10.1038/s41593-018-0209-y
20. Pereira TD, Tabris N, Matsliah A, Turner DM, Li J, Ravindranath S, et al. SLEAP: A deep learning system for multi-animal pose tracking. Nat Methods. 2022;19(4):486-95. doi:10.1038/s41592-022-01426-1
21. Geissmann Q, Garcia Rodriguez L, Beckwith EJ, French AS, Jamasb AR, Gilestro GF. Ethoscopes: An open platform for high-throughput ethomics. PLoS Biol. 2017;15(10). doi:10.1371/journal.pbio.2003026
22. van Klink R, Bowler DE, Gongalsky KB, Swengel AB, Gentile A, Chase JM. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. Science. 2020;368(6489):417-20. doi:10.1126/science.aax9931
23. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol. 2022;40(1):121-30. doi:10.1038/s41587-021-01001-7
24. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods. 2024;21(8):1470-80. doi:10.1038/s41592-024-02201-0
25. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. doi:10.1038/s41586-021-03819-2
26. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol. 2023;41(8):1099-106. doi:10.1038/s41587-022-01618-2
27. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111. doi:10.1186/s13059-020-02015-1
28. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 2021;18(3):272-82. doi:10.1038/s41592-020-01050-x
29. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445. doi:10.1038/s41467-021-23774-w
30. Ma A, McDermaid A, Xu J, Chang Y, Ma Q. Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. 2020;38(9):1007-22. doi:10.1016/j.tibtech.2020.02.013

31. Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. Elife. 2021;10. doi:10.7554/eLife.66747
32. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet. 2022;23(3):169-81. doi:10.1038/s41576-021-00434-9
33. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5. doi:10.1038/s41592-021-01256-7
34. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
35. Troudet J, Grandcolas P, Blin A, Vignes-Lebbe R, Legendre F. Taxonomic bias in biodiversity data and societal preferences. Sci Rep. 2017;7(1):9132. doi:10.1038/s41598-017-09084-6