

Advances in IT and Digital Security

ISSN 3143-4266

DAC Insight Publishers

<https://journals.dacinsightpublishers.com/AIDY>



MACHINE LEARNING FOR BIOMOLECULAR DATA WRANGLING AND CURATION: A REVIEW OF LEARNED EXTRACTION, ERROR DETECTION, INTEGRATION, AND CURATION WORKFLOWS

Kingsley Chinazaekpere Ndupu^{1*}, Chukwudera Obumneke Anunagba², Stanley Nwakamma³, Serif Oyindamola Oyesiji⁴

¹Kennesaw State University, Georgia, USA

²Barclays, USA

³Independent Researcher, USA

⁴Harrisburg University of Science and Technology, Harrisburg, PA

Corresponding Author*: kingsleynd34@gmail.com

ABSTRACT

The biomolecular sciences generate data faster than they can be made usable. Between instrument output and analyzable knowledge lies the unglamorous majority of bioinformatics labor: extracting assertions from literature, standardizing identifiers and vocabularies, detecting and repairing errors, resolving duplicate and conflicting records, imputing what is missing, correcting what batches distorted, and integrating what heterogeneous sources report differently. This wrangling and curation layer has historically been manual, expert, and chronically under-resourced relative to data growth, and machine learning has been moving into it from several directions at once. This review synthesizes the literature on machine learning for biomolecular data wrangling and curation. It organizes the field by curation function rather than by model family: learned extraction and annotation, where biomedical text mining and pre-trained language models convert literature into structured assertions and where computational function prediction annotates sequences at scale; learned error detection, from the foundational quality scores of sequencing through statistical validation of proteomics identifications to general anomaly detection over database entries; learned entity resolution and integration, where deduplication and record matching meet deep matching models; learned imputation and batch correction, where the missing-data and batch-effect literatures supply principled repair; and learning-directed curation workflows, where active learning, weak supervision, and human-in-the-loop systems allocate scarce expert attention. Cross-cutting infrastructure, ontologies and minimum-information standards, workflow managers, and the FAIR agenda, is reviewed as the substrate that makes learned curation deployable, and the evidence on pitfalls, leakage, label imperfection, and drift, is weighed. The review closes with open challenges, including curation of predicted rather than measured knowledge, and directions for the field.

Keywords: *Event-Driven Architecture, REST and WebSocket Integration, Channel Semantics, Offline-First Mobile Systems, Read-Then-Subscribe Splice, Backpressure Policy, State Convergence, Reconnection Resumption Contracts*

Received 1 May 2026; Accepted 22 May 2026, Available online 30 May 2026

1. INTRODUCTION

The scale argument for computational curation was settled by the data itself. High-throughput sequencing pushed biology into the exabyte conversation (Reuter et al., 2015; Stephens et al., 2015), mass spectrometry did the same for proteins and their modifications (Aebersold & Mann, 2003; Aebersold & Mann, 2016), and the repositories that receive the output, nucleotide archives, protein knowledge bases, structure and expression databases, proteomics exchanges, became critical scientific infrastructure with submission volumes no curator workforce could exhaustively review (Benson et al., 2013; UniProt Consortium, 2021; Berman et al., 2000; Barrett et al., 2013; Perez-Riverol et al., 2019; Deutsch et al., 2017). The biocuration community named the resulting gap early and has documented it since: expert curation is the standard that makes data trustworthy, and it is precisely the resource that does not scale with the data (Howe et al., 2008; Burge et al., 2012; International Society for Biocuration, 2018). The same scaling pressure is documented outside the biomolecular repositories, in regulated life-sciences operations where data integrity failures, their controls, and their remediation have been catalogued systematically (Aneke et al., 2023), and in national health-data architectures where interoperability and data quality are treated as explicit design targets rather than emergent properties (Ilodigwe & Adesemoye, 2021).

Machine learning entered this gap along several fronts rather than as one program. Biomedical text mining grew a community evaluation tradition around extracting genes, proteins, and their relations from literature (Hirschman et al., 2005), matured into deployed annotation aids (Wei et al., 2013; Leaman & Lu, 2016), and was transformed by domain-pre-trained language models (Lee et al., 2020; Beltagy et al., 2019; Gu et al., 2021; Luo et al., 2022). Statistical learning became load-bearing inside instrument pipelines, most consequentially in shotgun proteomics, where empirical models and semi-supervised rescoring made identification error a governed quantity (Keller et al., 2002; Käll et al., 2007; Elias & Gygi, 2007), a lineage with roots in the learned quality scores that disciplined sequencing itself (Ewing & Green, 1998). Function prediction organized itself into community assessment at knowledge-base scale (Radivojac et al., 2013; Zhou et al., 2019), and structure prediction crossed a threshold that turned computational annotation into a first-class database content type (Jumper et al., 2021; Varadi et al., 2022). Meanwhile the data-management community built a general learned-cleaning literature, error detection, probabilistic repair, and cleaning oriented to downstream models (Rahm & Do, 2000; Chu et al., 2016; Ilyas & Chu, 2019; Rekatsinas et al., 2017; Heidari et al., 2019; Krishnan et al., 2016), and the machine learning community supplied the workflow-level instruments, active learning, weak supervision, anomaly detection, that convert curation from exhaustive review into directed attention (Settles, 2009; Ratner et al., 2017; Chandola et al., 2009). Adjacent applied literatures developed the same instruments on different substrates: anomaly detection matured against transaction and time-series records where rare legitimate events and true errors are hard to separate (Atakpa & Fobellah, 2023; Atakpa & Abetoh, 2022), human-in-the-loop composition was consolidated as a design pattern in its own right (Ladapo et al., 2022), and curation of training corpora became an explicit engineering stage in model-development pipelines rather than an unexamined precondition (Oyesiji et al., 2024).

What the field lacks is synthesis across these fronts, and that is this review's purpose: to organize machine learning for biomolecular wrangling and curation by the curation functions it serves, to weigh the evidence for what works and where it fails, and to identify the challenges the next phase of the field must answer. Section 2 states scope and approach. Section 3 characterizes the wrangling problem. Sections 4 through 8 review the five function areas, extraction and

annotation, error detection, entity resolution and integration, imputation and batch correction, and learning-directed workflows. Section 9 reviews the enabling infrastructure, Section 10 the challenges, Section 11 future directions, and Section 12 concludes.

2. SCOPE AND APPROACH OF THE REVIEW

The review covers machine learning applied to the preparation, quality assurance, and integration of biomolecular data, sequences, proteins and their modifications, structures, expression, metabolites, and the literature-derived assertions about them, as distinct from machine learning for downstream biological discovery, which is reviewed elsewhere (Libbrecht & Noble, 2015; Min et al., 2017). The boundary is functional: a method belongs here when its output is better data, cleaner, more complete, better annotated, better integrated, rather than a biological conclusion. Included are the community evaluation traditions where they assess curation-relevant capabilities, text mining assessments (Hirschman et al., 2005) and function prediction assessments (Radivojac et al., 2013; Zhou et al., 2019), the instrument-pipeline validation literatures whose statistical machinery is curation in situ (Keller et al., 2002; Ewing & Green, 1998), and the general data-cleaning literature where biomolecular deployments draw on it directly (Ilyas & Chu, 2019). Excluded are clinical data curation, imaging, and electronic health records, whose constraints differ enough to warrant their own syntheses. Also excluded, though drawn on for contrast where the mechanisms transfer, is the wider governance literature on data quality in regulated enterprise settings (Aneke et al., 2023; Aliliele et al., 2025b; Orise & Niniola, 2025).

The literature base spans bioinformatics and genome-resource venues, where repositories and their pipelines are described; the biocuration community's own literature, where the human workflows are documented; proteomics and glycomics methods venues, where validation machinery lives; natural language processing venues, where biomedical text mining is published; and data management venues, where learned cleaning originates. Two synthesis choices shape the presentation. First, organization is by curation function, because functions persist while model families turn over, and because deployments compose functions regardless of the models inside them. Second, evaluation evidence is weighed with the pitfalls literature in view, since curation models are trained on curated labels and evaluated on the same repositories they serve, conditions under which leakage and circularity are structural risks rather than exotic ones (Whalen et al., 2022; Kapoor & Narayanan, 2023). Where evidence from adjacent domains is cited in what follows, it is cited for mechanism rather than for biological content: enterprise data-governance, audit-analytics, and model-governance work is invoked because it addresses the same structural problems of provenance, agreement, and drift under different data (Ominyi & Anichukwueze, 2023; Ozowara et al., 2025).

3. THE BIOMOLECULAR WRANGLING PROBLEM

The problem's shape follows from the data's origins. Biomolecular records are produced by heterogeneous instruments and laboratories under evolving protocols, described in free text as much as in structured fields, identified through vocabularies that vary by community and era, and deposited under standards whose adoption is partial. The consequences recur across data types: identifier ambiguity and synonymy, gene symbols reused across species, protein accessions superseded, metabolite names colliding (Rahm & Do, 2000; Elmagarmid et al., 2007); incompleteness against the applicable minimum-information checklists, which the standards communities publish precisely because submissions vary (Wilkinson et al., 2016; Rojas-Macias et

al., 2019); value-level error from instrument artifact through transcription slips to unit confusion; duplication and near-duplication as the same result enters through multiple routes; source-level distortion, the batch effects that make laboratory and processing signatures masquerade as biology (Leek et al., 2010; Johnson et al., 2007); and inter-source conflict, where integrated resources report incompatible assertions about the same entity, the condition integration platforms confront daily (UniProt Consortium, 2021; Wishart et al., 2022; York et al., 2020). These failure modes are not unique to biomolecular resources. Identifier and schema heterogeneity across sources is the recurring obstacle in enterprise integration architectures (Badmus et al., 2019a; Badmus et al., 2022b; Ogundairo, 2025), cross-jurisdiction record exchange raises the same normalization burden under stricter auditability requirements (Adesuyi et al., 2023; Akomolafe et al., 2025), and data sensitivity classification with regulatory traceability supplies a working vocabulary for exactly the provenance metadata that curation requires (Aliliele et al., 2024a). Upstream of all of it, error entering at the point of measurement remains the dominant contributor, which is why pre-analytical and analytical process control in accredited laboratories bears directly on the quality of records that later reach repositories (Ayinde et al., 2026), and why chain-of-custody integrity has been formalized in comparably regulated distribution settings (Aneke & Adesemoye, 2019).

Two structural features distinguish biomolecular curation from generic data cleaning. First, its ground truth is expert judgment against community standards, ontologies with curated term hierarchies (Ashburner et al., 2000; Smith et al., 2007) and reporting guidelines per data type (Rojas-Macias et al., 2019), so labels are expensive, drift with standards versions, and encode the very expertise automation is meant to conserve (Burge et al., 2012). Second, its errors are consequential downstream in ways that resist local assessment: a mis-normalized gene propagates through every analysis that joins on it, which is why the field's most successful learned curation, proteomics validation, is precisely the case where error was made a measured, controlled quantity at the point of production (Elias & Gygi, 2007; Käll et al., 2007). These features explain the field's trajectory reviewed below: learning succeeds where it either inherits calibrated error control or directs, rather than replaces, expert judgment. The dependence on expert judgment as ground truth also makes inter-rater agreement a measurable property of the labeling process rather than an assumption about it, a point developed formally in work on the disagreement problem in human-in-the-loop learning, where systematic reviewer divergence is shown to carry information that averaging destroys (Ladapo et al., 2026). The downstream consequences are equally concrete: analyses that join on curated annotations, from gene-expression and circulating-biomarker studies (Okwah et al., 2026) through nutrition and DNA-marker association work (Ike et al., 2025) to phenotypic characterization (Ekeocha et al., 2021), inherit whatever normalization error the underlying resources carry, usually without any means of detecting it locally.

4. LEARNED EXTRACTION AND ANNOTATION

Literature extraction is the oldest front. The community assessment tradition established shared tasks for gene and protein mention recognition, normalization to database identifiers, and relation extraction, with the inaugural critical assessment fixing the evaluation culture (Hirschman et al., 2005). Deployed systems converted the capability into curation aids: large-scale pre-annotation services supplying entity and concept annotations across the literature (Wei et al., 2013) and joint recognition-normalization models addressing the identifier ambiguity that makes normalization the hard half of the task (Leaman & Lu, 2016). Pre-trained language models re-founded the area: biomedical-corpus pre-training lifted recognition, normalization, and relation

extraction benchmarks substantially (Lee et al., 2020), scientific-text pre-training generalized the recipe (Beltagy et al., 2019), domain-from-scratch pre-training showed that in-domain vocabularies and corpora matter beyond continued training (Gu et al., 2021), and generative biomedical models extended the interface toward extraction and summarization framed as text generation (Luo et al., 2022). For curation practice, the settled deployment pattern is assistive: extraction proposes, with provenance to the sentence, and curators dispose, a division the biocuration community has endorsed as attention-multiplication rather than replacement (International Society for Biocuration, 2018; Burge et al., 2012). Comparable extraction pipelines have been built over non-biomedical corpora, with systems for public health surveillance illustrating the same recognition-and-normalization architecture operating at scale over noisy text (Atakpa et al., 2023), and multi-agent quality control applied to the review of regulated documents showing how automated proposals are gated before acceptance (Ominyi et al., 2025). Parameter-efficient adaptation makes such pipelines deployable where full domain pre-training is not affordable (Oyesiji et al., 2023), and the explainability that curators require of any proposal has been treated as an architectural constraint rather than a reporting afterthought in adjacent applied systems (Sanni et al., 2026).

Sequence and structure annotation is the second extraction front, computational rather than textual. Protein function prediction matured under community assessment at repository scale, with successive challenges documenting real but bounded progress and persistent difficulty on specific ontology branches (Radivojac et al., 2013; Zhou et al., 2019), against the ontology infrastructure that makes function a computable target at all (Ashburner et al., 2000). Structure prediction's breakthrough changed the annotation economics outright: near-experimental accuracy at proteome scale (Jumper et al., 2021) and the ensuing database of predicted structures made computational annotation a primary content type alongside experimental record (Varadi et al., 2022). The curation consequence is a new object of governance, predicted knowledge with model-attached confidence, whose handling, provenance labeling, confidence thresholds, and update-on-retrain policy, is now a live curation question the challenges section returns to. Governing model outputs as content, rather than as analysis, is the general form of this problem, and it has been posed in the same terms in the model-governance literature, where release-gating and automated evaluation determine what a model is permitted to assert and under what recorded conditions (Nwakamma et al., 2024a).

5. LEARNED ERROR DETECTION AND VALIDATION

The exemplar is instrument-pipeline validation, where statistical learning made error a controlled quantity. Sequencing established the pattern first: learned base-call quality scores with calibrated error probabilities became the substrate on which all downstream filtering rests (Ewing & Green, 1998). Shotgun proteomics generalized it: empirical mixture models converted database-search scores into identification probabilities (Keller et al., 2002), target-decoy construction supplied distribution-free false discovery estimation (Elias & Gygi, 2007), and semi-supervised rescoring learned discriminative functions per dataset while preserving decoy-based error control (Käll et al., 2007), a composition whose governing idea, learn aggressively, control error externally, is this review's recurring lesson. Multiple-testing machinery underwrites the same discipline wherever screening is high-dimensional (Benjamini & Hochberg, 1995; Storey & Tibshirani, 2003).

Above the instrument layer, error detection generalizes as anomaly and constraint learning over records. The anomaly detection literature supplies the unsupervised repertoire, density,

isolation, and distance-based detectors with a consolidated taxonomy (Chandola et al., 2009; Breunig et al., 2000; Liu et al., 2008), applied to database entries as deviation from accepted-entry distributions and violation of compositional constraints. The data-management strand adds learned error detection proper: systematic frameworks for cleaning (Chu et al., 2016; Ilyas & Chu, 2019), probabilistic repair engines that combine integrity constraints, statistics, and external evidence into holistic corrections (Rekatsinas et al., 2017), few-shot error detectors that learn a resource's error distribution from small samples via augmentation (Heidari et al., 2019), and cleaning oriented explicitly to the downstream model, prioritizing repairs by their effect on analysis (Krishnan et al., 2016). Deployed biomolecular use composes the two strands: constraint and standards checks convert community rules into features, statistical screens flag value-level implausibility, and detectors route candidates to review, with the calibration and evaluation cautions of the prediction literature attached wherever flags gate decisions (Guo et al., 2017; Whalen et al., 2022). The applied anomaly-detection literature outside biology supplies useful evidence on how such detectors behave at operational scale: on financial time series, where seasonality and legitimate rare events confound naive thresholds (Atakpa & Fobellah, 2023); on industrial control telemetry, where physical constraints supply the analogue of compositional checks (Adebayo et al., 2023; Adegbite et al., 2023); on log and event streams, where triage volume rather than detection sensitivity is the binding constraint (Mbonu et al., 2019); and on transaction monitoring, where scores must be calibrated against a fixed review budget (Fadayomi et al., 2024; Ominyi et al., 2024a). Root-cause and corrective-action methodology, developed in regulated quality systems, addresses the step after detection that curation also requires, namely deciding whether a flagged record indicates a local error or a systematic process fault worth correcting at source (Aneke et al., 2022). Detection at the point of measurement remains the cheapest intervention of all, whether through field-deployable analytical methods (Aneke & Adesemoye, 2024) or through learned interpretation of biosensor output (Olorunkosebi et al., 2023).

6. LEARNED ENTITY RESOLUTION AND INTEGRATION

Entity resolution, deciding when records denote the same molecule, gene, sample, or study, is the integration bottleneck and a classical learning problem. The record-linkage tradition established the machinery, blocking to control candidate space, similarity features, and match classification, surveyed comprehensively for duplicate detection (Elmagarmid et al., 2007) and consolidated as an engineering discipline with evaluation practice (Christen, 2012). Deep matching brought representation learning to the comparison step, with systematic architectures showing where learned representations beat engineered similarity, textual and dirty attributes foremost, and where classical methods hold, clean structured records (Mudgal et al., 2018). Biomolecular resolution adds domain structure to the generic problem: identifier systems with explicit cross-references make much resolution rule-based, while the hard residue, literature-derived mentions, legacy records, cross-database assertions, is exactly the ambiguous-name regime where learned normalization earns its keep (Leaman & Lu, 2016; Wei et al., 2013). Integration engineering outside the biomolecular domain has converged on similar decompositions, with extract-transform-load design treating identifier mapping as a first-class concern rather than a preprocessing detail (Badmus et al., 2019a) and enterprise service-bus patterns formalizing how heterogeneous sources are reconciled without collapsing their provenance (Badmus et al., 2022b).

Integration platforms are where resolution meets conflict. Knowledge bases that aggregate across sources, protein-centric (UniProt Consortium, 2021), metabolome-centric (Wishart et al.,

2022), and glycoscience-centric (York et al., 2020), confront the full pipeline: map identifiers, merge duplicates, and adjudicate conflicting assertions with provenance retained, under ontology-governed vocabularies that make assertions comparable at all (Ashburner et al., 2000; Smith et al., 2007). The learned contribution across this pipeline is presently strongest at its edges, mention normalization into identifier space and duplicate candidate ranking, and weakest at adjudication, where conflict resolution remains policy and expertise; the probabilistic-repair line points at learned adjudication, evidence-weighted repair under constraints (Rekatsinas et al., 2017), but biomolecular deployments at that depth remain early, a gap Section 10 records. Where adjudication cannot yet be automated, the auditable alternative is to make the adjudication trail itself durable and inspectable, an approach developed in regulatory-transparency and audit-trail architectures for cross-organizational data exchange (Kumuyi et al., 2023; Oshoba et al., 2020).

7. LEARNED IMPUTATION AND BATCH CORRECTION

Missingness in biomolecular data is structured and informative, and its principled treatment rests on the missing-data framework, mechanisms distinguished, analysis matched to mechanism (Rubin, 1976; Little & Rubin, 2002), with chained-equations multiple imputation as the general-purpose workhorse and its pooling rules preserving uncertainty through downstream analysis (van Buuren & Groothuis-Oudshoorn, 2011; Sterne et al., 2009). The curation-relevant discipline is less the imputation model than its placement: imputation is fitted within, not before, whatever validation surrounds downstream use, and absence indicators are retained where missingness itself carries signal, the practice the pitfalls literature demands (Whalen et al., 2022; Kapoor & Narayanan, 2023). The same placement discipline recurs wherever imputed values feed forecasting: long-range statistical models prove sensitive to how gaps were filled well before they are sensitive to model class (Atakpa et al., 2024), and integration of heterogeneous covariate sources compounds the problem when missingness patterns differ systematically by source (Oparah et al., 2024).

Batch correction is the repair function the high-throughput era made unavoidable. The foundational demonstration that batch effects pervade high-throughput studies and can dominate biological signal reframed source structure as a first-class data-quality dimension (Leek et al., 2010), and empirical-Bayes adjustment became the standard repair for location-scale batch differences, robust at the modest per-batch sample sizes real studies have (Johnson et al., 2007). For curation practice, the two functions converge on one rule: repaired values are annotated values. Imputed entries, corrected intensities, and adjusted profiles enter resources with method, version, and provenance attached, because repair without provenance converts data quality problems into data provenance problems, and the FAIR agenda's reusability principle makes the annotation obligation explicit (Wilkinson et al., 2016). Source-structured distortion is not confined to laboratory batches. Systematic differences between data-generating units can produce models that are accurate in aggregate while being unreliable for particular subgroups, a failure mode documented directly in district-level prediction work where overall accuracy masked uneven performance across units (Komi, 2023). That result is the same argument, reached from a different direction, for treating source structure as a data-quality dimension rather than a nuisance parameter.

8. LEARNING-DIRECTED CURATION WORKFLOWS

The workflow layer is where learning meets the scarce resource directly. Active learning formalizes attention allocation: query strategies select the records whose labels most improve the

model, with the survey literature organizing uncertainty, disagreement, and expected-change strategies and their failure modes (Settles, 2009), and curation deployments using the same machinery to order review queues by expected yield. Weak supervision attacks the label bottleneck from the other side: labeling functions, heuristics, standards checks, distant supervision from existing resources, are combined by a generative model into probabilistic training labels, trading label quality for label volume under explicit noise modeling (Ratner et al., 2017), a natural fit for curation where community rules and legacy annotations abound and gold labels are dear. Human-in-the-loop composition is the deployed synthesis: models propose with provenance and calibrated confidence, curators adjudicate, adjudications return as labels, and the loop's own health, agreement rates, drift in proposal quality, reviewer burden, is monitored, the assistive pattern the biocuration community's assessments endorse and the interactive-cleaning line demonstrates for statistical workloads (International Society for Biocuration, 2018; Krishnan et al., 2016). State-of-the-art treatments of human-in-the-loop machine learning consolidate the same composition beyond curation, cataloguing where human judgment enters learning loops and what it costs to keep it there (Ladapo et al., 2022). Two further findings from that literature bear on curation loops specifically. The first is that reviewer disagreement is systematic rather than noise, so loops that average it away discard the signal that would otherwise flag ambiguous cases or drifting standards (Ladapo et al., 2026). The second is that federated and cross-organizational arrangements, in which review labor is distributed across institutions that cannot pool raw records, are a realistic deployment shape for community resources rather than an exotic one (Sanni et al., 2023).

The workflow layer also carries the automation-governance question. Curation labels are expert judgment, so inter-curator agreement bounds every model trained on them; automation bias, deference to model proposals, threatens the independence that gives adjudication its value; and feedback loops, submitters learning the checker, model retraining on its own accepted proposals, can entrench errors efficiently. The documented responses are procedural: blinded review streams that continuously measure model-curator disagreement, provenance separating human-asserted from machine-proposed content, and retraining gates on labeled, adjudicated data only, disciplines the reproducibility literature's leakage taxonomy motivates directly (Kapoor & Narayanan, 2023; Whalen et al., 2022). Comparable governance controls have been specified for automated decision systems in other regulated settings, where continuous auditing of model behavior in production (Aliliele et al., 2025a; Fobellah, 2025), documented model-risk management with named ownership (Ominyi & Anichukwueze, 2023), and structured acceptance testing before deployment (Eyetsemitan et al., 2023) are treated as prerequisites rather than refinements.

9. INFRASTRUCTURE AND REPRODUCIBILITY SUBSTRATE

Learned curation deploys on infrastructure the community built for other reasons, and inherits its strengths. Ontologies make annotation computable and interoperable, the gene ontology as the exemplar (Ashburner et al., 2000) and the foundry coordinating orthogonal, principled vocabularies across domains (Smith et al., 2007). Minimum-information standards convert completeness into checkable predicates per data type, with community roadmaps pressing standardization where heterogeneity is worst (Rojas-Macias et al., 2019), and the FAIR principles frame the target properties, findable, accessible, interoperable, reusable, that curation serves (Wilkinson et al., 2016). Repositories supply the deployment surface and the training substrate at once (UniProt Consortium, 2021; Perez-Riverol et al., 2019; Deutsch et al., 2017; Barrett et al., 2013), and workflow managers make curation pipelines themselves reproducible artifacts,

versioned, portable, and provenance-capturing (Di Tommaso et al., 2017; Köster & Rahmann, 2012; Wratten et al., 2021), on standard scientific tooling (Pedregosa et al., 2011). The substrate's reproducibility stakes are the field's: the deficits documented across preclinical research price what silent data problems cost downstream (Freedman et al., 2015; Baker, 2016; Ioannidis, 2005), which is the economic argument for curation infrastructure stated in the currency decision-makers use. The operational half of this substrate is less often written up but is equally load-bearing. Curation pipelines are software, and inherit the practices that make software auditable: version control with enforced review (Badmus et al., 2022a), continuous integration and deployment discipline (Badmus et al., 2018), declarative infrastructure definition so that environments are reproducible rather than remembered (Mbonu et al., 2020), and governed platform administration where the deployment target is itself regulated (Badmus et al., 2019b). Storage and access architecture carries comparable weight, with lakehouse governance addressing lineage and retention at repository scale (Aliliele et al., 2025b) and privacy-preserving architecture determining what may be integrated at all when sources fall under differing access regimes (Atakpa & Abolaji, 2022; Aliliele et al., 2024b). The legal and ethical framing of these obligations has itself been modeled explicitly, which matters for curation because access constraints shape what integration is permissible before they shape what is technically feasible (Mbonu et al., 2018; Orise & Niniola, 2025).

10. CHALLENGES AND OPEN PROBLEMS

Five challenges structure the field's near term. Label validity and circularity: curation models learn from curated labels and are evaluated on the repositories they serve, so agreement statistics, standards-version drift, and train-test contamination through shared provenance are structural threats, and the field under-reports all three; the remedy, agreement measurement before modeling, provenance-aware splitting, and temporal evaluation, is known and unevenly practiced (Whalen et al., 2022; Kapoor & Narayanan, 2023). Adjudication remains unlearned: extraction, flagging, matching, and ranking are demonstrably learnable, while conflict resolution across sources, the step that makes integrated resources authoritative, remains expert policy, with evidence-weighted probabilistic repair a promising but under-deployed direction (Rekatsinas et al., 2017; UniProt Consortium, 2021). Analogous accountability problems have been analyzed in settings where automated decisions must be defended to a regulator, and the resulting emphasis on documented model risk, contestability, and evidentiary trails maps closely onto what curation circularity audits would need to record (Ominyi & Anichukwueze, 2023; Ominyi & Anichukwueze, 2024; Annan, 2024).

Predicted knowledge governance is the newest challenge: repositories now hold model outputs at scale, structures foremost (Jumper et al., 2021; Varadi et al., 2022), with function prediction pressing the same way (Zhou et al., 2019), and curation must decide provenance labeling, confidence thresholds for downstream visibility, versioning under model updates, and the containment of circularity when predictions train successors. Language-model extraction raises the adjacent version: generative extraction is fluent, scalable, and capable of confident error, so the assistive pattern's provenance-to-sentence discipline becomes load-bearing exactly as the proposal volume grows (Luo et al., 2022; Gu et al., 2021), and human-centered treatments of generative-AI-assisted annotation reach the same conclusion from the annotation side, with loop design deciding whether oversight remains meaningful as automation proposes at scale (Ladapo et al., 2024). Finally, evaluation infrastructure lags the functions: text mining and function prediction have community assessments (Hirschman et al., 2005; Radivojac et al., 2013), while

error detection, resolution, and workflow-level curation lack shared benchmarks with realistic error distributions, leaving deployments to per-resource evidence, a gap the general cleaning literature's benchmark practice could seed (Ilyas & Chu, 2019; Heidari et al., 2019). Institutional readiness for the governance these challenges imply is uneven, as assessments of organizational preparedness for emerging AI regulation report (Ominyi et al., 2024b; Afrihyia et al., 2024), and the technical attack surface of automated proposal pipelines has itself become a governance object, with generative components exposed to manipulation of their inputs and tooling in ways that would corrupt proposals silently (Nwakamma et al., 2024b). Auditability of automated decision-making under transparency and anti-discrimination obligations supplies a further template for what a defensible provenance policy must record and retain (Annan, 2025).

11. FUTURE DIRECTIONS

Four directions follow. First, calibrated curation as a norm: every learned proposal, extraction, flag, match, repair, carries a calibrated confidence and provenance, with resource-level dashboards of proposal quality over time, extending the proteomics lesson, error as a governed quantity, across all curation functions (Elias & Gygi, 2007; Guo et al., 2017). Second, benchmark construction for the unbenchmarked functions: community datasets with realistic, documented error and duplication distributions, and evaluation protocols that respect provenance structure in their splits, giving error detection and resolution the assessment culture extraction has enjoyed (Hirschman et al., 2005; Christen, 2012). Third, weak-supervision-first curation stacks: community rules, standards checks, and legacy annotations compiled systematically into labeling functions per resource, making the label bottleneck an engineering problem rather than a staffing one (Ratner et al., 2017), with active learning allocating the expert labels that remain (Settles, 2009). Fourth, governance frameworks for predicted content: provenance ontologies distinguishing measured, curated, and predicted assertions, threshold and versioning policies for model-derived records, and circularity audits as standard repository practice, the curation community's next standards project as prediction becomes a primary content source (Varadi et al., 2022; Smith et al., 2007; Wilkinson et al., 2016). Two enabling conditions cut across all four. The first is that governance be specified as an institutional function with defined ownership and audit obligations, rather than as a set of per-project conventions that lapse with staff turnover (Ominyi et al., 2026; Ekechi et al., 2026). The second is that curation of training data be treated as a first-class stage of any model that will subsequently propose curation, closing the loop between corpus construction and downstream assertion quality (Oyesiji et al., 2024), including where inference must run under constrained resources close to the point of data generation (Nwakamma et al., 2025). Risk-based approaches to evidentiary sufficiency, developed where regulators must decide how much evidence is enough to act, offer a workable model for setting confidence thresholds on predicted content (Eze et al., 2024).

12. CONCLUSION

This review has synthesized machine learning for biomolecular data wrangling and curation across five functions: extraction and annotation, where text mining and computational prediction now propose at scales curation could never reach manually; error detection and validation, where the instrument-pipeline lineage shows learned scoring under external error control at its most successful; entity resolution and integration, where learning owns the edges and expertise still owns adjudication; imputation and batch correction, where principled repair is

available and provenance-annotated deployment is the discipline; and learning-directed workflows, where active learning, weak supervision, and human-in-the-loop composition convert scarce expertise into directed attention.

Three conclusions carry the synthesis. First, the field's durable successes share one architecture: models propose aggressively, error is controlled by mechanisms outside the model, and humans adjudicate where judgment is the ground truth. Second, the binding constraints are now evaluative and governance-shaped rather than algorithmic: label circularity, missing benchmarks for whole function areas, and the influx of predicted content are where progress is gated. Third, the economics remain decisive: data growth guarantees the curation gap widens on manual methods, and the reviewed literature shows learned curation is the only instrument that scales expert judgment rather than merely substituting for it. The agenda that follows, calibration as norm, benchmarks for the unbenchmarked, weak-supervision-first labeling, and governance for predicted knowledge, is the path by which the field's data infrastructure keeps pace with its instruments. That instrument is not exclusive to biology. The same architecture, aggressive proposal under external error control with human adjudication reserved for genuine judgment, has emerged independently wherever expert review became the bottleneck on data quality (Ladapo et al., 2022; Ominyi et al., 2025; Borah et al., 2022), which suggests the curation problem is structural rather than domain-specific, and that the biomolecular sciences should expect both to borrow from and to contribute to that wider literature.

REFERENCES

1. Adebayo, A., Adegbite, M. P., & Ahmed, M. O. (2023). AI augmented threat detection in industrial control systems: A systematic review of machine learning approaches for ICS anomaly detection. *International Journal of Engineering and Modern Technology*, 9(2), 1–32.
2. Adegbite, M. P., Adebayo, A., & Ahmed, M. O. (2023). ICS and SCADA threat detection architectures in energy sector networks: A systematic review of SIEM, NDR, and anomaly detection approaches. *World Journal of Innovation and Modern Technology*, 7(2), 121–150.
3. Adesuyi, M. O., Akomolafe, O., Olaogun, B. O., Ndukwe, V. U., & Sakyi, J. K. (2023). Global interoperability model for distributed ledger-based cross-border payment systems. *International Journal of Multidisciplinary Research and Growth Evaluation*, 4(6), 1–18.
4. Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928), 198–207.
5. Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620), 347–355.
6. Afrihyia, E., Akinse, S. G., & Ojukwu, P. U. (2024). Comparative governance of AI-driven healthcare management: Executive oversight, regulatory structures, and accountability in the United States and developing countries. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3087–3102.
7. Akomolafe, O., Olaogun, B. O., Adesuyi, M. O., Ndukwe, V. U., & Sakyi, J. K. (2025). Global standardization model for next-generation blockchain interoperability in payment systems. *International Journal of Multidisciplinary Research and Growth Evaluation*, 6(1), 1–19.
8. Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2024a). A conceptual framework for enterprise data sensitivity classification and regulatory traceability mechanisms. *International Journal of Advanced Multidisciplinary Research and Studies*, 4(6), 3103–3124.
9. Aliliele, C., Mbonu, I. S., & Iwuanyanwu, U. (2024b). Advances in HIPAA compliant data architecture and secure analytics frameworks for community healthcare organizations. *Shodhshauryam, International Scientific Refereed Research Journal*, 7(2), 277–324.
10. Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025a). A review of AI assisted continuous auditing systems in technology risk and cybersecurity oversight. *Gyanshauryam, International Scientific Refereed Research Journal*, 8(4), 210–250.

11. Aliliele, C., Mbonu, I. S., Uzoka, E., & Iwuanyanwu, U. (2025b). Advances in data lakehouse governance architectures for enterprise data loss prevention and compliance assurance. *Shodhshauryam, International Scientific Refereed Research Journal*, 8(4), 171–213.
12. Aneke, O. B., & Adesemoye, A. C. (2019). Toward an integrated conceptual framework for traceability and chain-of-custody integrity in pharmaceutical distribution. *Iconic Research and Engineering Journals*, 3(4), 655–672.
13. Aneke, O. B., & Adesemoye, A. C. (2024). Recent advances in analytical and field-deployable methods for detecting substandard and falsified pharmaceuticals. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 10(5), 1285–1323.
14. Aneke, O. B., Adesemoye, A. C., & Sydney, E. (2022). Corrective and preventive action effectiveness in pharmaceutical quality management systems: A systematic review of root-cause methodologies. *International Journal of Health and Pharmaceutical Research*, 7(2), 78–111.
15. Aneke, O. B., Sydney, E., & Adesemoye, A. C. (2023). Data integrity and governance in regulated life sciences operations: A review of failures, controls, and remediation strategies. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(10), 538–568.
16. Annan, A. O. (2024). Algorithmic accountability and trade secret protection in artificial intelligence. *Shodhshauryam, International Scientific Refereed Research Journal*, 7(5), 315–347.
17. Annan, A. O. (2025). Automated decision-making and anti-discrimination compliance under U.S. law. *International Journal of Advanced Multidisciplinary Research and Studies*, 5(2), 2522–2540.
18. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1), 25–29.
19. Atakpa, M. I., & Abetoh, N. F. (2022). A systematic review of machine learning advances in financial fraud detection for banking systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 8(2), 771–801.
20. Atakpa, M. I., & Abolaji, T. O. (2022). A privacy-preserving data architecture model for regulated industry analytics under GDPR and HIPAA compliance. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 8(3), 781–808.
21. Atakpa, M. I., & Fobellah, A. N. (2023). Anomaly detection in financial time-series data: A conceptual model for healthcare and banking applications. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(2), 967–997.
22. Atakpa, M. I., Abolaji, T. O., & Abetoh, N. F. (2024). Statistical time-series models for long-range public health trend forecasting: A review and conceptual framework. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 10(6), 2748–2780.
23. Atakpa, M. I., Abolaji, T. O., & Akeju, B. (2023). A review of natural language processing for social media-based public health surveillance and analytics. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(6), 1030–1060.
24. Ayinde, M., Ohunyon, G., & Ojukwu, P. U. (2026). Optimization of pre-analytical and analytical processes in CLIA-certified clinical laboratories: Recent innovations and future directions. *International Journal of Advanced Multidisciplinary Research and Studies*, 6(2), 585–600.
25. Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2018). A systematic review of CI/CD pipeline strategies in Salesforce DevOps: Tools, practices, and deployment outcomes. *Iconic Research and Engineering Journals*, 2(6), 1–19.
26. Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2019a). A conceptual model for ETL design and data integration in Salesforce-centric enterprise architectures. *Iconic Research and Engineering Journals*, 3(5), 1–20.
27. Badmus, O., Dosunmu, A. A., & Ozowara, D. E. (2019b). A governance framework for Salesforce platform management in regulated healthcare environments. *Iconic Research and Engineering Journals*, 3(6), 1–18.
28. Badmus, O., Jooda, D., Ozowara, D. E., & Anunagba, C. O. (2022a). A framework for Git-based version control and code review governance in Salesforce development teams. *Shodhshauryam, International Scientific Refereed Research Journal*, 5(2), 473–493.
29. Badmus, O., Jooda, D., Ozowara, D. E., & Anunagba, C. O. (2022b). A systematic review of MuleSoft ESB integration patterns in multi-cloud Salesforce architectures. *Gyanshauryam, International Scientific Refereed Research Journal*, 5(3), 462–481.
30. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454.
31. Barrett, T., Wilhite, S. E., Ledoux, P., et al. (2013). NCBI GEO: Archive for functional genomics data sets, update. *Nucleic Acids Research*, 41(D1), D991–D995.

32. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3615–3620.
33. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.
34. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(D1), D36–D42.
35. Berman, H. M., Westbrook, J., Feng, Z., et al. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.
36. Borah, S., Aliliele, K. C., Rakshit, S., & Vajjhala, N. R. (2022). Applications of artificial intelligence in software testing. In P. K. Mallick et al. (Eds.), *Cognitive informatics and soft computing (Lecture Notes in Networks and Systems, Vol. 375, pp. 727–736)*. Springer.
37. Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 93–104.
38. Burge, S., Attwood, T. K., Bateman, A., et al. (2012). Biocurators and biocuration: Surveying the 21st century challenges. *Database*, 2012, bar059.
39. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
40. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 15.
41. Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.
42. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2201–2206.
43. Deutsch, E. W., Csordas, A., Sun, Z., et al. (2017). The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, 45(D1), D1100–D1106.
44. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), 316–319.
45. Ekechi, N. V., Ozowara, D. E., & Anunagba, C. O. (2026). Conceptual framework for AI governance, data privacy compliance, and financial sustainability in digital health. *Computer Science and IT Research Journal*, 7(4), 275–299.
46. Ekeocha, A. H., Aganga, A. A., Adejoro, F. A., Oyeboji, A., Oluwadele, J. F., & Tawose, O. M. (2021). Phenotypic characteristics of indigenous chickens in selected regions of Nigeria. *Journal of World's Poultry Research*, 11(3), 352–358.
47. Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3), 207–214.
48. Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 1–16.
49. Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3), 186–194.
50. Eyetsemitan, R. A., Ambali, K. B., Oyeleye, A. O., & Fadayomi, O. (2023). User acceptance testing in small business technology deployment: A structured validation framework for lean operational environments. *International Journal of Multidisciplinary Research and Growth Evaluation*, 4(6), 1–16.
51. Eze, F. I., Akinleye, O. K., & Anene, U. N. (2024). Development of a risk-based accelerated approval decision model balancing speed, safety, and evidence quality in pharmaceutical regulatory science. *Research Journal of Pure Science and Technology*, 7(6), 120–161.
52. Fadayomi, O., Bello, A. D., Elebe, O., Hammed, N. I., & Omoegun, G. O. (2024). An adaptive fraud risk scoring model for real-time transaction monitoring at scale. *IIARD International Journal of Banking and Finance Research*, 10(10), 212–230.
53. Fobellah, A. N. (2025). Navigating digital transformation: Auditing artificial intelligence-powered financial systems: A conceptual review. *International Journal of Science and Research Archive*, 16(2), 23–28.
54. Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biology*, 13(6), e1002165.
55. Gu, Y., Tinn, R., Cheng, H., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 2.
56. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the International Conference on Machine Learning*, 1321–1330.

57. Heidari, A., McGrath, J., Ilyas, I. F., & Rekatsinas, T. (2019). HoloDetect: Few-shot learning for error detection. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 829–846.
58. Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1), S1.
59. Howe, D., Costanzo, M., Fey, P., et al. (2008). Big data: The future of biocuration. *Nature*, 455(7209), 47–50.
60. Ike, P. C., Enyinnaya, A. C., Ujam, C. G., Ayinde, M. K., Golden, F. I., Olatokun, T. P., & Nasir, K. (2025). Investigating the relationship between nutrition and DNA markers associated with heart health in middle-aged adults with type 2 diabetes. *International Journal of Convergent and Informatics Science Research*, 8(9), 1–18.
61. Ilodigwe, L., & Adesemoye, A. C. (2021). The data backbone of health system transformation: A governance and architecture framework for interoperability, data quality, and advanced analytics at national scale. *International Journal of Health and Pharmaceutical Research*, 6(3), 1–38.
62. Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. ACM Books.
63. International Society for Biocuration. (2018). Biocuration: Distilling data into knowledge. *PLoS Biology*, 16(2), e2002846.
64. Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
65. Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), 118–127.
66. Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.
67. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.
68. Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74(20), 5383–5392.
69. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, 4(11), 923–925.
70. Komi, N. M. (2023). When accurate forecasts are not fair forecasts: Deep learning load prediction and equity across districts. *Shodhshauryam, International Scientific Refereed Research Journal*, 6(6), 476–546.
71. Krishnan, S., Wang, J., Wu, E., Franklin, M. J., & Goldberg, K. (2016). ActiveClean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, 9(12), 948–959.
72. Köster, J., & Rahmann, S. (2012). Snakemake: A scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520–2522.
73. Kumuyi, O., Akeju, B., Uzoka, E., & Ozowara, D. E. (2023). Framework for blockchain-based cross-border data exchange and regulatory transparency. *International Journal of Multidisciplinary Futuristic Development*, 4(1), 99–113.
74. Ladapo, O. O., Dosunmu, A. A., Jooda, D., & Abolaji, T. O. (2022). Human-in-the-loop machine learning: A state of the art. *Journal of Frontiers in Multidisciplinary Research*, 3(1), 656–669.
75. Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2024). Keeping humans in the loop: Human-centered automated annotation with generative AI. *International Journal of Multidisciplinary Futuristic Development*, 5(1), 81–95.
76. Ladapo, O. O., Jooda, D., Dosunmu, A. A., & Abolaji, T. O. (2026). On the disagreement problem in human-in-the-loop federated machine learning. *International Journal of Multidisciplinary Research and Growth Evaluation*, 7(3), 178–192.
77. Leaman, R., & Lu, Z. (2016). TaggerOne: Joint named entity recognition and normalization with semi-Markov models. *Bioinformatics*, 32(18), 2839–2846.
78. Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240.
79. Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733–739.
80. Libbrecht, M. W., & Noble, W. S. (2015). *Machine learning applications in genetics and genomics*. *Nature Reviews Genetics*, 16(6), 321–332.
81. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

82. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. *Proceedings of the IEEE International Conference on Data Mining*, 413–422.
83. Luo, R., Sun, L., Xia, Y., et al. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), bbac409.
84. Mbonu, I. S., Aliliele, C., Iwuanyanwu, U., & Oluoha, O. M. (2018). A conceptual framework for legal and ethical risk modeling in enterprise data protection governance systems. *Iconic Research and Engineering Journals*, 2(2), 207–226.
85. Mbonu, I. S., Iwuanyanwu, U., Aliliele, C., & Uzoka, E. (2020). Advances in infrastructure as code governance for secure terraform based enterprise cloud deployments. *International Journal of Multidisciplinary Research and Growth Evaluation*, 1(5), 811–828.
86. Mbonu, I. S., Iwuanyanwu, U., Uzoka, E., & Oluoha, O. M. (2019). Advances in enterprise log analytics and automated incident response architectures using Python and SIEM platforms. *Iconic Research and Engineering Journals*, 3(2), 1000–1019.
87. Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5), 851–869.
88. Mudgal, S., Li, H., Rekatsinas, T., et al. (2018). Deep learning for entity matching: A design space exploration. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 19–34.
89. Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024a). LLM-as-a-judge for automated model governance: A review of methods, biases, and release-gating practices. *World Journal of Innovation and Modern Technology*, 8(6), 185–216.
90. Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2024b). Securing agentic AI enterprise workflows against prompt injection, tool poisoning, memory manipulation, and excessive agency threats. *World Journal of Innovation and Modern Technology*, 8(6), 217–248.
91. Nwakamma, S., Ojukwu, J., & Oyesiji, S. O. (2025). On-device multimodal small language models for privacy-preserving edge intelligence through quantization, pruning, distillation, and energy-efficient inference. *World Journal of Innovation and Modern Technology*, 9(12), 271–302.
92. Ogundairo, K. M. (2025). Hardening the autonomous value chain: Lean Six Sigma guardrails and multi-enterprise ETL architecture for agentic supply chain orchestration. *IIARD International Journal of Economics and Business Management*, 11(12), 441–478.
93. Okwah, M. N., Abah, M. A., Oladosu, M. A., Blessing, T. B., Anku, A. I., & Agida, O. D. (2026). Anti-inflammatory effects of omega-3 fatty acids: Evidence from circulating biomarkers and inflammatory gene-expression studies. *African Research Reports*, 2(4), 365–382.
94. Olorunkosebi, M. T., Ayinde, M. K., & Mensah, G. K. (2023). Integrating artificial intelligence and advanced biosensors for rapid and sensitive detection of antimicrobial resistant foodborne pathogens. *Journal of Advances in Microbiology Research*, 4(2), 175–182.
95. Ominyi, M., & Anichukwueze, C. C. (2023). A review of algorithmic accountability and model risk management approaches in financial services. *Journal of Accounting and Financial Management*, 9(12), 218–238.
96. Ominyi, M., & Anichukwueze, C. C. (2024). Conceptualizing responsible AI governance structures for automated decision systems in regulated industries. *International Journal of Social Sciences and Management Research*, 10(11), 403–430.
97. Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024a). Developing a conceptual framework for AI-assisted compliance monitoring and anomaly detection in high-volume transaction environments. *World Journal of Innovation and Modern Technology*, 8(6), 204–228.
98. Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2024b). A systematic review of organizational readiness for the EU Artificial Intelligence Act in multinational enterprises. *World Journal of Innovation and Modern Technology*, 8(6), 185–203.
99. Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2025). A conceptual framework for multi-agent AI quality control in the review of regulated documents. *International Journal of Social Sciences and Management Research*, 11(8), 534–557.
100. Ominyi, M., Anichukwueze, C. C., & Uzougbo, N. S. (2026). Conceptualizing enterprise AI governance functions as institutional standards for auditable and defensible risk management. *Journal of Accounting and Financial Management*, 12(7), 326–356.
101. Oparah, O. S., Ezech, F. E., Olatunji, G. I., & Ajayi, O. O. (2024). Framework for integrating climate data and health outcomes to improve mortality risk prediction systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 10(2), 1128–1150.

- 102.Orise, C., & Niniola, S. (2025). Data governance in digital health learning systems: A systematic review of implementation frameworks and best practices. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(5), 520–538.
- 103.Oshoba, T. O., Hammed, N. I., & Odejebi, O. D. (2020). Blockchain-enabled compliance and audit trail model for cloud configuration management. *Journal of Frontiers in Multidisciplinary Research*, 1(1), 193–201.
- 104.Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2023). Parameter-efficient multilingual adaptation of on-device language models: A survey. *International Journal of Engineering and Modern Technology*, 9(3), 287–320.
- 105.Oyesiji, S. O., Nwakamma, S., & Ojukwu, J. (2024). A conceptual framework for context-aware data curation in fine-tuning pipelines. *International Journal of Engineering and Modern Technology*, 10(11), 197–228.
- 106.Ozowara, D. E., Adebayo, A., & Anunagba, C. O. (2025). Advanced conceptual model for strengthening audit quality using data analytics across financial institutions. *International Journal of Advanced Multidisciplinary Research and Studies*, 5(6), 2246–2268.
- 107.Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- 108.Perez-Riverol, Y., Csordas, A., Bai, J., et al. (2019). The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Research*, 47(D1), D442–D450.
- 109.Radivojac, P., Clark, W. T., Oron, T. R., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3), 221–227.
- 110.Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.
- 111.Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3), 269–282.
- 112.Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11), 1190–1201.
- 113.Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular Cell*, 58(4), 586–597.
- 114.Rojas-Macias, M. A., Mariethoz, J., Andersson, P., et al. (2019). Towards a standardized bioinformatics infrastructure for N- and O-glycomics. *Nature Communications*, 10, 3275.
- 115.Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.
- 116.Sanni, J. O., Iwuanyanwu, U. A., & Essien, M. A. (2026). Designing explainable AI based marketing automation architectures for healthcare and financial applications. *World Scientific News*, 213, 119–148.
- 117.Sanni, J. O., Iwuanyanwu, U. A., Essien, M. A., & Attah, A. (2023). Lifecycle-aware marketing automation using federated learning for secure cross-organizational data management. *Gyanshauryam, International Scientific Refereed Research Journal*, 6(6), 337–364.
- 118.Settles, B. (2009). Active learning literature survey (Computer Sciences Technical Report 1648). University of Wisconsin-Madison.
- 119.Smith, B., Ashburner, M., Rosse, C., et al. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11), 1251–1255.
- 120.Stephens, Z. D., Lee, S. Y., Faghri, F., et al. (2015). Big data: Astronomical or genomics? *PLoS Biology*, 13(7), e1002195.
- 121.Sterne, J. A. C., White, I. R., Carlin, J. B., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*, 338, b2393.
- 122.Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.
- 123.UniProt Consortium. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489.
- 124.Varadi, M., Anyango, S., Deshpande, M., et al. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1), D439–D444.
- 125.Weï, C.-H., Kao, H.-Y., & Lu, Z. (2013). PubTator: A web-based text mining tool for assisting biocuration. *Nucleic Acids Research*, 41(W1), W518–W522.
- 126.Whalen, S., Schreiber, J., Noble, W. S., & Pollard, K. S. (2022). Navigating the pitfalls of applying machine learning in genomics. *Nature Reviews Genetics*, 23(3), 169–181.

127. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
128. Wishart, D. S., Guo, A., Oler, E., et al. (2022). HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research*, 50(D1), D622–D631.
129. Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, 18(10), 1161–1168.
130. York, W. S., Mazumder, R., Ranzinger, R., et al. (2020). GlyGen: Computational and informatics resources for glycoscience. *Glycobiology*, 30(2), 72–73.
131. Zhou, N., Jiang, Y., Bergquist, T. R., et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20, 244.