



Available online at
<https://jiicet.gnt.com.pk/index.php/jiicet/>

Journal of Innovative Intelligent Computing and Emerging Technologies (JIICET)

Vol. 1, No. 1, 2026



Explainable Artificial Intelligence in Bioinformatics: Applications, Challenges, and Future Opportunities

Sajjad Ahmed¹, Shaman Ali¹, Nimra¹, Shahid Ali¹

¹ Information Technology Centre, Sindh Agriculture University, Tandojam Hyderabad, Sindh

* Correspondence: ahmedsajjadgopang@gmail.com, shamanalimahar462@gmail.com

Article History

Received 05 August 2026
Revised 07 October 2026
Accepted 9 October 2026
Available Online 23 October 2026

Keywords:

Explainable AI (XAI); bioinformatics interpretable machine learning; multi-omics
biomarker discovery; SHAP
attention; foundation models
digital pathology

Abstract

Machine learning now sits at the centre of bioinformatics, delivering strong performance on tasks from gene-expression classification and protein-structure prediction to digital pathology and drug-target scoring. The best-performing models, however, are the hardest to interrogate, and in biology a prediction that cannot be explained has limited scientific or clinical value. Explainable artificial intelligence (XAI) addresses that gap by producing human-readable accounts of model behaviour. This review surveys the techniques applied to biological data, SHAP, LIME, Grad-CAM, attention, layer-wise relevance propagation, graph explainers and biologically constrained architectures, and their use across genomics, transcriptomics, proteomics, sequence analysis, structural bioinformatics, multi-omics integration and biomedical imaging. Six structured tables and seven figures compare methods, applications, metrics and open problems, each accompanied by a statement of what the comparison contributes. We then weigh these methods against the critical evidence published between 2024 and 2026, which is more sobering than the enthusiasm of the primary literature suggests: attention-derived scores in single-cell foundation models add no predictive value over simple gene-level statistics, and widely used saliency methods can survive randomisation of the very model they claim to explain. Grouping ten challenges by cause rather than by symptom produces the review's central argument. Instability under collinearity, computational cost and the limits of attention are properties of current methods and will yield to better algorithms; the absence of mechanistic ground truth, the scarcity of interventional data and the gap between statistical importance and biological causality will not yield to algorithms at all, and depend instead on experimental and curatorial effort. We close with research priorities: causal and counterfactual explanation, interpretability of biological foundation models, community benchmarks, human-in-the-loop systems, multi-scale explanation and prospective clinical evaluation, and argue that XAI should be judged by whether its explanations survive experimental test, not by how plausible they look.



Copyright: © 2025 by the authors. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License. (<https://creativecommons.org/licenses/by-nc/4.0/>)

1. Introduction

High-throughput technologies have produced biological datasets of a size and complexity that conventional statistics handles poorly: whole genomes, transcriptomes, proteomes, metabolomes and high-resolution images, in which the signal of interest is usually non-linear and distributed across many weakly informative variables. Machine learning absorbed that problem rather than

solving it outright, and deep models are now standard equipment in bioinformatics [1, 2], reaching competitive accuracy in protein-structure prediction, disease-subtype classification, drug-target scoring and image-based diagnosis.

Accuracy, however, is not the whole requirement. The best-performing models are generally the least transparent, and in biology and medicine a correct prediction whose basis cannot be examined is of limited use. Researchers need

explanations they can check against prior knowledge and turn into testable hypotheses; clinicians need to know why a particular patient received a particular recommendation; regulators increasingly require both [3, 4]. Rudin has argued that for high-stakes decisions the sensible response is to use interpretable models from the outset rather than to explain opaque ones after the fact [40], and that argument has shaped much of the debate since.

Explainable artificial intelligence aims to make model behaviour legible. Applied to biological data, it has been used to rank genes and variants, highlight discriminative tissue regions, recover sequence motifs and surface pathway-level patterns. Several surveys map parts of this literature [5-8], but they tend to treat the methods and the critiques separately, and the critiques have sharpened considerably in the last two years.

Three developments motivate a fresh synthesis. First, biological foundation models have emerged, with internal representations that are both more powerful and harder to interpret than the architectures XAI methods were designed for. The second is a growing body of negative results: careful evaluations showing that popular explanation methods are less faithful than their use implies. The third is regulatory, as transparency requirements for high-risk clinical AI move from guidance towards obligation. This review therefore does two things at once: it surveys what XAI can do in bioinformatics, and it takes seriously the published evidence on where it fails. Figure 1 shows where explanation sits in a typical pipeline and why the loop it belongs to is often left open.

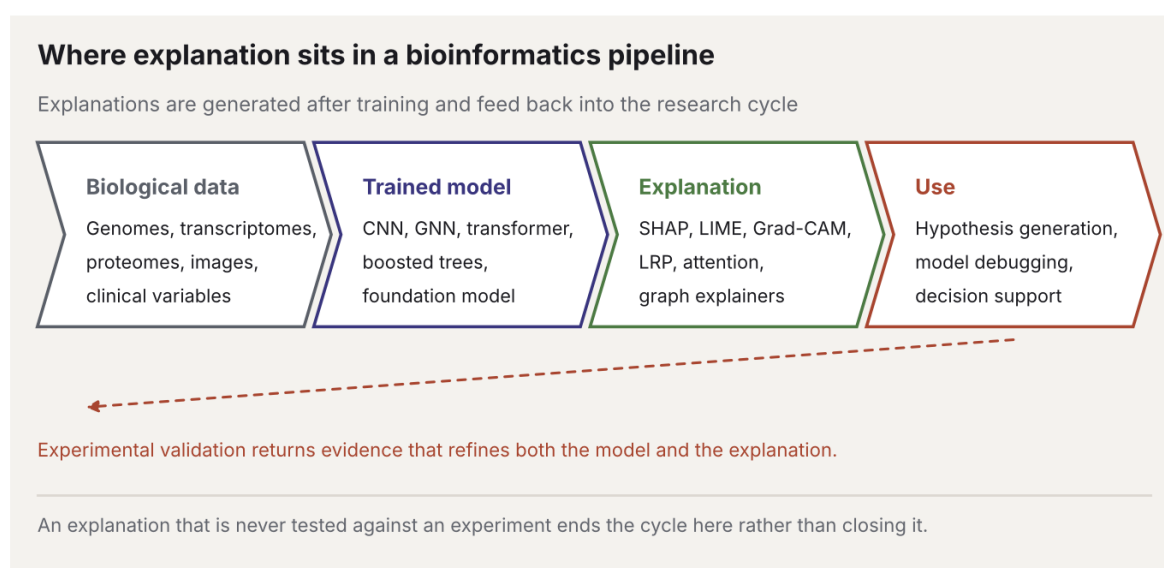


Figure 1. Explanations are produced after a model is trained, and feed back into hypothesis generation, debugging and clinical decisions rather than terminating the analysis.

2. Materials and Methods

This article is a narrative review. We searched PubMed, Scopus, Web of Science, arXiv and bioRxiv, combining explanation terms ("explainable AI", "interpretable machine learning", "SHAP", "Grad-CAM", "attention", "counterfactual") with domain terms ("bioinformatics", "genomics", "transcriptomics", "proteomics", "multi-omics", "digital pathology", "single-cell"). We prioritized peer-reviewed work and well-cited preprints published between 2020 and early 2026, along with foundational XAI papers from earlier years that remain in routine use.

A study was retained if it either introduced an explanation method applied to biological data or reported a substantial case study in which the explanation produced biological

insight. We also retained evaluation studies that tested explanation quality, whether or not they proposed a method, since these carry most of the evidence on reliability. Purely methodological XAI papers without a biological application were excluded unless they supplied essential background. We then organized studies by data modality and explanation technique.

No new experiments were performed. All quantitative statements are those reported by the original authors, and where the primary and evaluation literature disagree, we say so rather than reconciling them.

3. Results and Discussion

We begin with the methods and what distinguishes them, move through their application by data modality, then

compare both against evidence of how well the explanations hold up.

3.1. A working taxonomy of explanation methods

Explanation methods divide along three axes that matter in practice: whether interpretability is built into the model or added afterwards, whether the method needs access to the model's internals, and whether it explains a single prediction or the model as a whole. Figure 2 places the main families on the first and third of these; Figure 3 adds the two practical constraints that usually decide the matter, cost and stability; and Table 1 summarises the families themselves.

Model-agnostic methods treat the trained predictor as a black box and probe its input-output behaviour. SHAP assigns each feature an importance equal to its average marginal contribution across coalitions of features, with efficient approximations such as TreeSHAP and DeepSHAP making it tractable for omics-scale data [11, 41]. LIME fits a sparse linear surrogate in the neighbourhood of the instance being explained [12]. Both are flexible and widely used; however, they become unstable when features are strongly correlated, which omics data almost always are.

Where the architecture is known, gradient-based methods are cheaper and often sharper. Grad-CAM produces class-

discriminative localisation maps from the final convolutional layer [13], integrated gradients satisfy sensitivity and implementation-invariance axioms [32], and layer-wise relevance propagation redistributes the prediction score backwards under conservation rules [30, 31]. These dominate imaging and convolutional sequence models.

Attention is often described as intrinsic interpretability, since the model produces the weight itself and, in sequence models, they frequently coincide with known motifs. That reading has been contested since Jain and Wallace showed that attention weights can be altered substantially without changing predictions [43]; Wiegreffe and Pinter replied that this depends on what one asks of an explanation [44]. The disagreement is unresolved and, as Section 3.9 shows, recent evidence from single-cell models favours the sceptical side.

Two further families deserve mention. Biologically informed networks embed pathway or interaction graphs directly into the architecture, so that information flow follows known biology and explanations arrive in units biologists already use [14, 45]. Graph-specific explainers such as GNNExplainer and PGExplainer identify the subgraphs or nodes that most influence a graph network's prediction [22, 33].

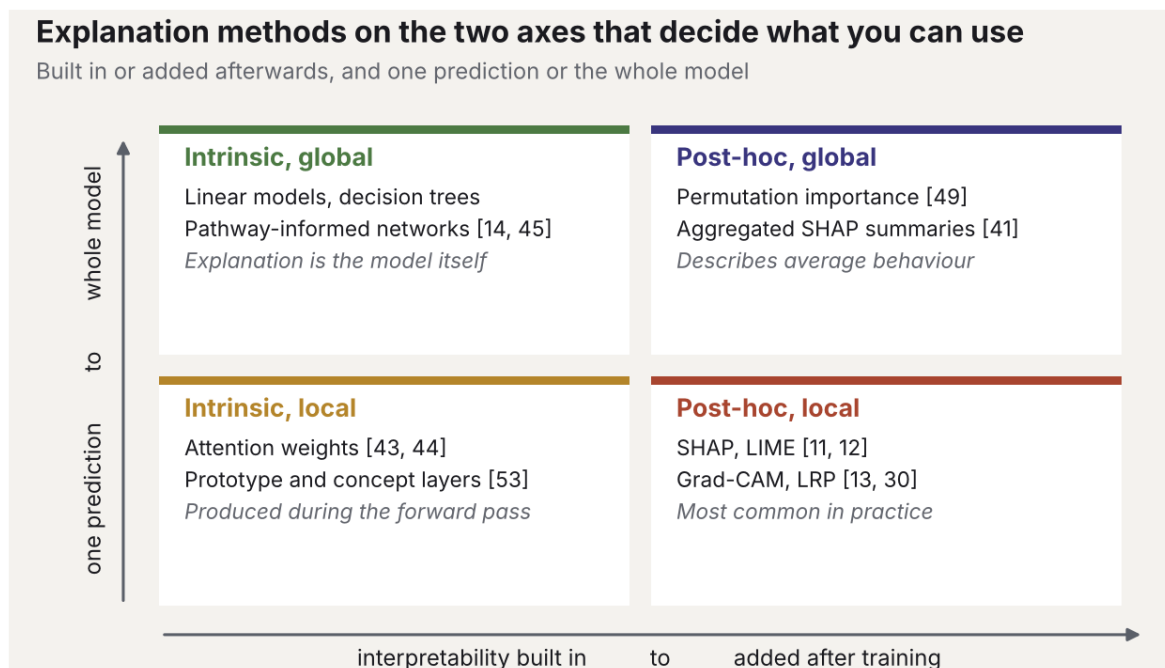


Figure 2. The two axes that matter operationally: whether interpretability is built in or added afterwards, and whether the explanation covers one prediction or the whole model.

Table 1. Families of explanation methods applied to biological data.

Category	Representative methods	Scope	Typical use in bioinformatics	Ref.
Intrinsic/interpretable by design	Linear models, decision trees, attention, pathway-informed networks	Global and local	Biomarker ranking, motif highlighting, pathway-level explanation	[14, 40, 45]
Post-hoc model-agnostic	SHAP, LIME, permutation importance	Mainly local	Gene ranking, patient-level explanation, multi-omics feature importance	[11, 12, 49]
Post-hoc gradient-based	Grad-CAM, integrated gradients, SmoothGrad	Local, visual	Pathology localisation, sequence contribution maps	[13, 32]
Propagation-based	LRP, DeepLIFT	Local	Relevance maps in deep genomic and imaging networks	[30, 31]
Graph-specific	GNNExplainer, PGExplainer, GraphSHAP	Local, subgraph	Drug-target interfaces, molecular graph attribution	[22, 33]
Example-based and counterfactual	Prototypes, counterfactuals, influence functions	Local	What-if clinical scenarios, nearest supporting examples	[51]

Contribution of Table 1. Grouping the methods this way makes the practical constraint visible, which a popularity-based list would not. Model access determines what is available: without it only perturbation-based methods apply, and those are the ones most damaged by correlated features. Scope determines what the explanation can support: local methods answer why

this patient, global methods answer what the model learned, and conflating the two is a common source of overstated claims. The table also shows that the intrinsic family is the only one whose explanations are guaranteed to describe the computation actually performed, which is the substance of the argument for interpretable models in high-stakes settings [40].

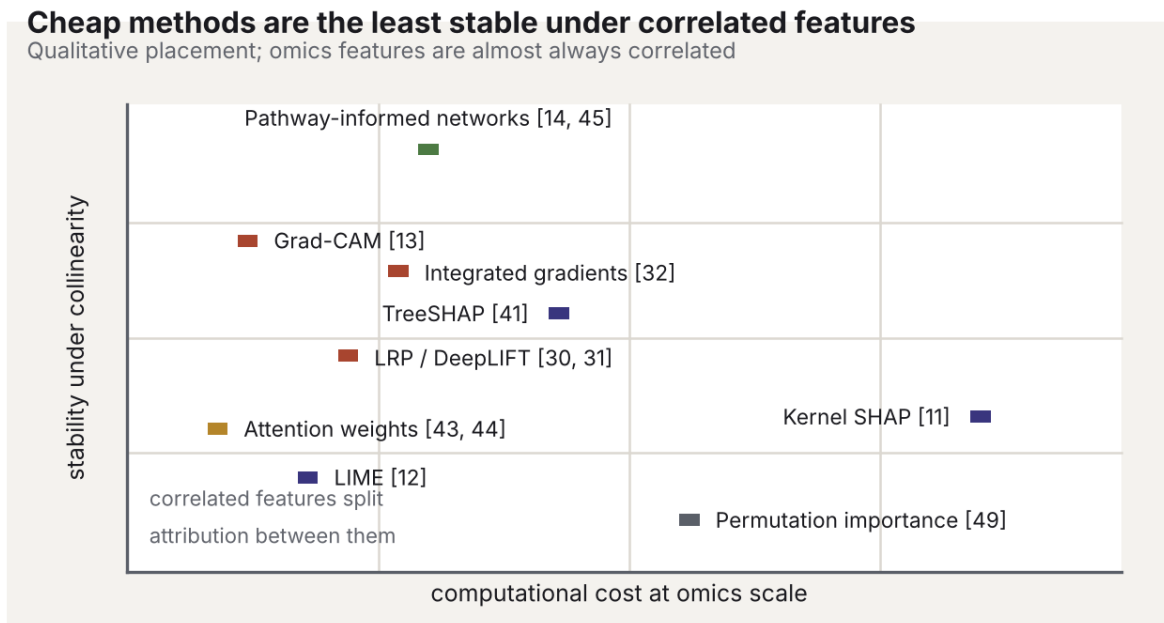


Figure 3. Qualitative placement of the main families: the methods that need least access to the model are also the ones most affected by correlated features and the most expensive to compute at omics scale.

3.2. Genomics and genome-wide association studies

Genomic data pose the attribution problem in its hardest form. Feature counts run from hundreds of thousands of variants to millions of genomic windows, and the correlation structure is not incidental but biological, arising from linkage disequilibrium and population stratification. Deep models trained for variant-effect prediction, polygenic risk scoring, chromatin-feature prediction and disease classification [46-48] have become standard, and

explanation methods are used to identify which variants or regions drive their outputs.

Layer-wise relevance propagation and SHAP have both been applied to highlight disease-associated variants and regulatory elements, and attention-based sequence models trained on DNA recover known transcription-factor binding motifs, which amounts to a form of in-silico mutagenesis [15, 16]. Explanations of this kind do more than build confidence: they prioritise candidates for experimental follow-up, which is where their value is

ultimately tested. A systematic mapping of XAI for omics confirms that DNA-based datasets account for a substantial share of the published literature, with feature-relevance methods the dominant paradigm [7].

Three difficulties are specific to this setting. Population structure must be corrected for, or the model will learn ancestry rather than biology. Rare variants contribute little individually and are easily lost in aggregate rankings. And when features are collinear, attribution among them is close to arbitrary, a problem Section 3.9.1 treats in full, and one that applies with particular force to linkage-disequilibrium blocks.

3.3. Transcriptomics and gene-expression analysis

Expression matrices, bulk RNA-seq, microarray or single-cell, remain the most common input for supervised learning in bioinformatics, and TreeSHAP and DeepSHAP are used routinely to rank genes by their contribution to phenotype prediction, subtype classification or survival modelling [17, 18, 41]. Systematic evaluation suggests DeepSHAP strikes a reasonable balance between the number of significant prognostic factors recovered and the stability of those rankings across repeated runs [19].

Single-gene rankings are rarely the end point. Attributions are increasingly aggregated to pathways, gene sets, or co-expression modules, and this step is not cosmetic: gene-level attributions are noisy and unstable under collinearity, whereas pathway-level summaries are more reproducible and easier to interpret biologically. Attention mechanisms in multi-omics architectures add another layer by highlighting cross-modal interactions. Single-cell data complicate matters again, since explanations may be required at the level of individual cells and simultaneously at the level of cell types or developmental trajectories.

3.4. Proteomics and pathway analysis

Proteomic data carry their own technical burdens, missing values, batch effects, post-translational modifications, but they sit closer to cellular phenotype than transcript abundance does, which makes their explanations correspondingly more actionable. Networks that encode protein-protein interactions or pathway hierarchies directly in their architecture improve both accuracy and interpretability, and when combined with SHAP or integrated gradients, they produce pathway-level accounts that align with known disease mechanisms [14].

P-NET best demonstrates what this approach can deliver. Built as a sparse network whose layers correspond to genes, pathways and biological processes, it predicted treatment resistance in prostate cancer [15] and, because its units were biologically meaningful by construction, implicated specific molecular processes that were

subsequently validated experimentally [45]. The distinction matters: the explanation did not have to be reconstructed from a black box because the model was structured in the vocabulary of the question from the start.

3.5. Biological sequence analysis

Sequence models have moved from convolutional and recurrent architectures to transformer-based foundation models trained on DNA, RNA or protein corpora, DNABERT, the Nucleotide Transformer, ESM and ProtTrans among them [35, 36, 50]. All of them benefit from methods that identify which positions or motifs drive a prediction. Grad-CAM, attention rollout and gradient attribution have been used to recover motifs in promoters, enhancers, splice sites and protein domains [5, 20], and SHAP computed on sequence embeddings or one-hot encodings can isolate residues whose mutation the model expects to alter function.

These explanations serve two distinct purposes that are worth separating. One is hypothesis generation about molecular recognition and catalysis. The other, less often discussed but arguably more reliable, is model debugging: attribution maps reveal when a model has latched onto an artefact, such as a GC-content gradient or a batch-specific adapter sequence, rather than the biology it was meant to learn.

Peptide activity prediction illustrates how these two purposes come apart in practice. StackAHTPs, a stacked ensemble that identifies antihypertensive peptides from heterogeneous sequence features, couples its prediction to feature-level explanation so that the physicochemical descriptors driving each call can be inspected rather than assumed [54]. Explanation here is not a presentation layer: it is what allows the contribution of a given descriptor family to be compared against what is known about the peptide chemistry.

3.6. Structural bioinformatics and drug discovery

Graph neural networks operating on molecular graphs, residue graphs or protein-ligand complexes are now common for affinity prediction, binding-site identification and interaction modelling. Graph-specific explainers, together with SHAP on graph embeddings, identify the atoms, bonds, or residues that most influence a prediction [21, 22, 33], and these attributions are directly usable: a medicinal chemist optimising a lead compound can act on an atom-level attribution in a way that no global feature ranking permits.

Evaluation in this domain has advanced faster than elsewhere, for a structural reason. Because molecular graphs are discrete and synthesizable, we can construct white-box oracle models whose decision rules are known

by design and then measure how accurately an attribution method recovers them. That gives cheminformatics something most of bioinformatics lacks, a ground truth against which explanation quality can be scored rather than merely inspected.

Target triage benefits from the same treatment one step earlier in the pipeline. Cascade deep forest models for druggable protein identification reach competitive accuracy while retaining the feature-importance accounting that tree ensembles provide by construction [55], which matters because an early druggability call determines which targets receive expensive structural and chemical effort later.

3.7. Biomedical imaging and digital pathology

Convolutional networks and vision transformers classify histopathology slides, radiology scans and microscopy images at high accuracy, and Grad-CAM and its variants remain the standard tools for producing class-activation maps over diagnostically relevant tissue [6, 23]. Embedded in multiple-instance learning frameworks, they supply a slide-level diagnosis together with region-level evidence a pathologist can inspect directly.

Visual explanation is valued in clinical settings because it maps onto the spatial reasoning pathologists and radiologists already perform, which lowers the cost of adoption. That same intuitiveness is a hazard. A heatmap over plausible tissue is persuasive whether or not it reflects the model's computation, and the sanity checks discussed in Section 3.9.3 were motivated precisely by saliency maps that looked convincing while being largely independent of the trained model [42].

The scale of adoption is now substantial: a systematic review of XAI in medical imaging covering 133 studies maps saliency, attention, model-agnostic and rule-based methods onto radiology and pathology workflows, and reports that explanations raise clinician confidence while

standardisation, interpretability and data bias remain unresolved [57]. Adoption and validation, in other words, have advanced at different rates.

3.8. Multi-omics integration

Integrating genomics, transcriptomics, proteomics, metabolomics and clinical variables is widely regarded as necessary for precision medicine, and explanation methods applied to such models can indicate which modalities, and which features within them, drive stratification, prognosis or drug response [24]. Attention-based fusion architectures yield modality-level importance directly; SHAP can be computed on concatenated feature vectors or on learned latent representations.

The difficulties compound rather than merely accumulate. Batch effects differ across platforms, modalities are missing for some patients, and the interactions worth explaining are precisely the higher-order ones that span biological layers and are hardest to attribute. Even so, published case studies have surfaced cross-omics signatures that univariate analysis would not have found, which is the strongest practical argument for the approach.

A recent study of host response shows what this looks like when it works. By integrating host genetics, multi-omics, and challenge-study transcriptomes into an explainable model, the authors separated a shared type I interferon backbone from infection-specific signatures, with coagulation, endothelial, and IL-6 inflammatory programmes distinguishing COVID-19 from the stronger interferon and T-cell signatures of influenza [58]. The explanation carried the finding: a comparable black-box classifier would have reported accuracy and nothing about mechanism.

Table 2 summarises where each technique is used, and Figure 4 shows the same pattern at a glance.

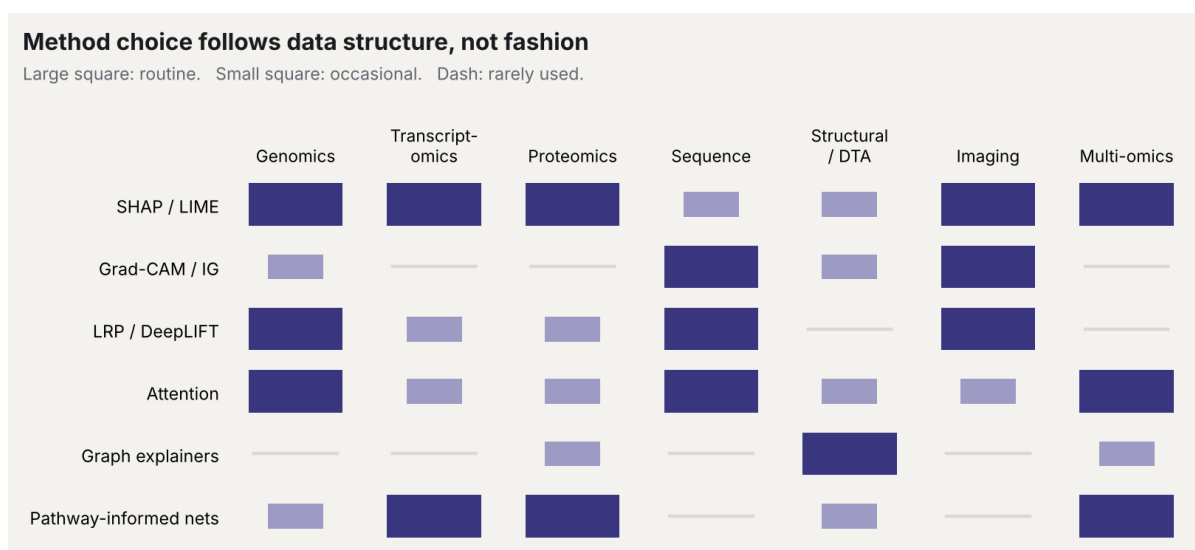


Figure 4. Method choice follows data structure: spatial inputs favour gradient methods, relational inputs favour graph explainers, and tabular inputs favour perturbation methods.

Table 2. Representative applications of explanation methods across bioinformatics domains.

Domain	Primary tasks	Common techniques	Typical insight obtained	Ref.
Genomics and GWAS	Variant prioritisation, disease risk	SHAP, LRP, attention	Disease-associated variants, recovered transcription-factor motifs	[7, 15, 16]
Transcriptomics	Subtype classification, survival prediction	TreeSHAP, DeepSHAP, pathway aggregation	Prognostic gene signatures, pathway narratives	[17-19, 41]
Proteomics	Biomarker and pathway analysis	Biologically informed networks with SHAP	Pathway-level mechanistic explanation	[14, 45]
Sequence analysis	Motif discovery, function prediction	Grad-CAM, attention, SHAP	Critical residues and sequence motifs	[5, 20, 50]
Structural and drug-target	Binding affinity, interaction sites	GNNExplainer, GraphSHAP	Key atoms and binding interfaces	[21, 22, 33]
Biomedical imaging	Diagnosis, region localisation	Grad-CAM, LIME	Pathologically relevant tissue regions	[6, 23]
Multi-omics	Patient stratification, drug response	Attention fusion, SHAP	Cross-omics feature importance	[24]

Contribution of Table 2. Read down the technique column, the table shows how strongly method choice is determined by data structure rather than by preference: gradient methods where inputs are spatial, graph explainers where they are relational, perturbation methods where they are tabular. Read down the insight column, it shows something less comfortable. The insights are mostly rankings and localizations: which features, which region, which atoms, and almost never mechanisms. That gap between what these methods return and what biologists want from an explanation is the theme of the rest of this section.

3.9. Challenges and limitations

The obstacles below are not of one kind, and the difference matters for deciding where effort should go. Some are properties of current algorithms; others follow from what biological data can supply.

3.9.1. High dimensionality, sparsity and collinearity

Omics matrices routinely contain tens of thousands of features measured on tens to hundreds of samples, and the features are rarely independent: co-expressed genes, linkage-disequilibrium blocks and co-abundant proteins all form tightly correlated groups. Most attribution algorithms distribute importance within such a group more or less arbitrarily, so rankings shift under small changes to the training set or the random seed. Permutation-based measures are especially vulnerable, since permuting one member of a correlated group leaves the information available through its neighbours [49].

The usual responses- aggressive feature selection, explicit modelling of correlation structure, aggregation from features to pathways or modules- all help, and all introduce modelling choices of their own. Aggregation in particular trades resolution for stability, which is often the right trade

in practice but should be reported as a choice rather than presented as a result.

The size of the effect is easy to underestimate, so we ran a small illustrative simulation. Five correlated variables carrying one shared signal were combined with five independent noise variables, and a gradient-boosted classifier was refitted on forty bootstrap resamples with permutation importance recorded each time. Across those runs, the share of total importance assigned to any single

member of the correlated block swung from essentially zero to 0.48, with an interquartile range 1.35 times its own median, while the block's combined share stayed near 0.81 with a relative interquartile range of 0.13, roughly ten times more stable. Figure 5 shows the result. Nothing about the underlying biology changed between runs; only the resampling did. Aggregating to the level of the block, which in real data means the pathway or module, recovers almost all of the lost stability, and this is the practical reason that pathway-level reporting has become standard.

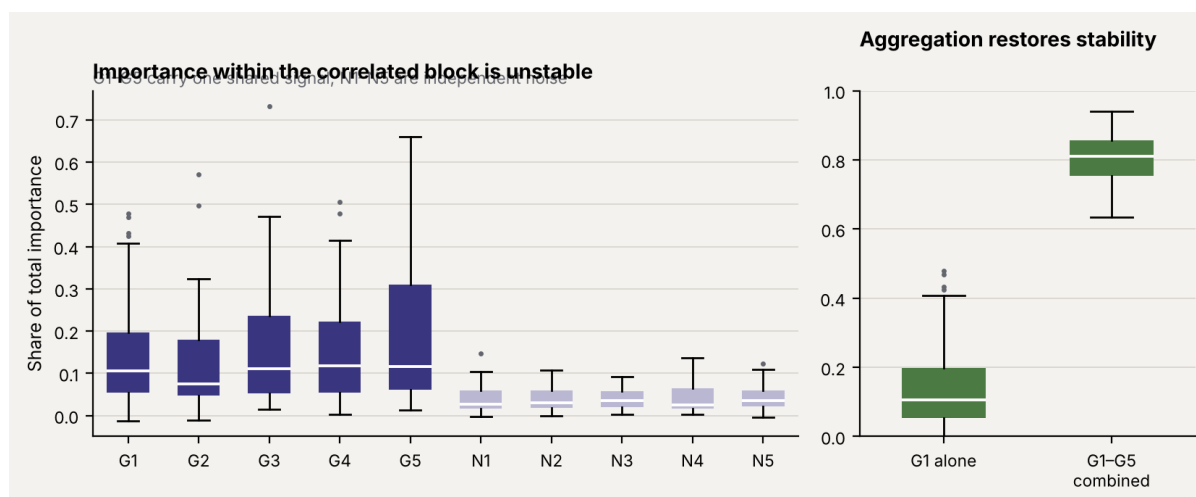


Figure 5. Illustrative simulation performed for this review: attribution moves freely within a correlated block across bootstrap resamples, while the block total does not.

3.9.2. The absence of mechanistic ground truth

In computer vision, approximate localisation ground truth can be obtained from human annotators. In molecular biology, a definitively correct explanation rarely exists. Evaluation therefore falls back on proxy metrics, faithfulness under perturbation, stability across runs, sparsity, or downstream biological validation such as overlap with curated pathways, enrichment of known disease genes, or experimental follow-up.

As a result, the field has no shared yardstick. Benchmark datasets with partially known mechanistic ground truth remain scarce; without them, method comparison stays largely qualitative, and progress is hard to accumulate [8, 25]. The white-box oracle approach used in cheminformatics (Section 3.6) is one of the few constructions that escapes this, and it does so by sacrificing biological realism for verifiability.

3.9.3. Evaluation, reproducibility and the evidence on faithfulness

A range of quantitative metrics exists: deletion and insertion curves, stability under input or model perturbation, agreement with knowledge bases, human usefulness ratings, but no consensus on which best reflects biological utility. Many published studies still rest primarily on qualitative visualisation, which makes systematic

comparison impossible, and reproducibility suffers further from incomplete reporting of preprocessing, hyperparameters, seeds and software versions.

The accumulated evidence is not reassuring. Adebayo and colleagues showed that several widely used saliency methods produce visually similar maps whether the model is trained or randomly initialised, meaning the maps were substantially independent of what the model had learned [42]. In bioinformatics specifically, a systematic evaluation of single-cell foundation models found that attention-derived edge scores from scGPT and Geneformer add no predictive value over simple gene-level statistics such as variance, mean expression and dropout rate when predicting CRISPR perturbation responses; ablating the attention heads ranked as most regulatory did not degrade performance [52]. A complementary study of attention in these models reached a similar conclusion, finding that attention largely captures co-expression rather than unique regulatory signal. These results do not show that explanation is hopeless, but they do show that plausibility is not evidence, and that a method's popularity says nothing about its faithfulness.

A comparative atlas of sparse autoencoder features in Geneformer and scGPT sharpens the point. The features recovered are biologically organised, corresponding to

pathway membership, co-activation modules and a layer-wise hierarchy, yet when they were tested against CRISPRi perturbation data only 3 of 48 transcription factors, that is 6.2 per cent, produced responses specific to their known regulatory targets [59]. Organised structure and causal regulatory wiring are evidently not the same thing, and an explanation can recover the first while saying nothing about the second.

Figure 6 maps the available checks against what each cannot establish, and Table 5 collects the studies that applied them. A recent review of multi-omics practice reaches a complementary conclusion from the other direction: interpretability, stability and reproducibility are typically studied in isolation and reported inconsistently, and the authors propose a structured framework and checklist covering stability testing, multi-cohort benchmarking and model-card reporting as the minimum that should accompany any published explanation [56].

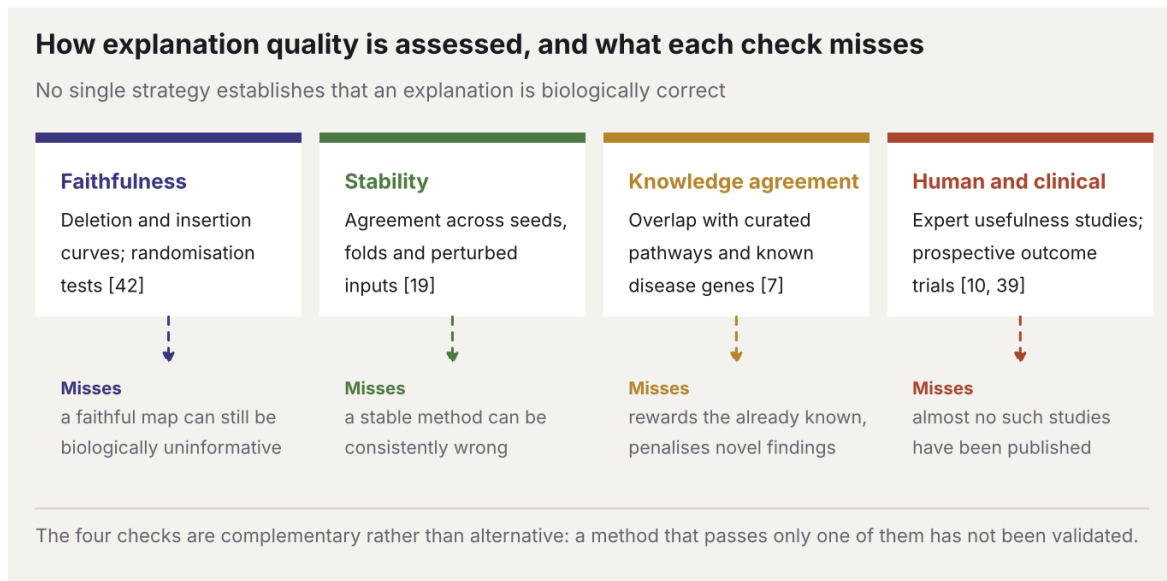


Figure 6. Four complementary checks on explanation quality; each has a characteristic blind spot, and passing only one does not validate the explanation.

Table 3. Metrics used to assess explanation quality.

Metric or test	What it measures	What it cannot establish	Ref.
Deletion/insertion curves	Change in prediction as attributed features are removed or added	Whether the attributed features are biologically meaningful	[9]
Model randomisation test	Whether the explanation changes when model weights are randomised	Faithfulness in any positive sense; it is a necessary minimum	[42]
Stability across seeds and folds	Reproducibility of the ranking under resampling	Correctness; a method can be stably wrong	[19]
Sparsity	How concentrated the attribution is	Whether the concentration is on the right features	[25]
Knowledge-base enrichment	Overlap with curated pathways or known disease genes	Novel mechanism, which by definition is not yet curated	[7]
Attention ablation	Performance loss when highly attended components are removed	Whether attention encodes signal used elsewhere in the network	[52]
Expert usefulness rating	Whether domain experts find the explanation helpful	Whether decisions actually improve	[10]
Prospective outcome study	Effect of explanations on diagnosis, confidence or outcomes	Generalisation beyond the studied setting; almost none exist	[39]

Contribution of Table 3. The third column is the one worth reading closely. Every commonly used metric establishes something weaker than the claim it is usually cited to support,

and two of them, the randomisation test and attention ablation, are purely negative checks: passing them means only that an explanation has not been shown to be spurious.

Together, the table argues for reporting several metrics from different rows rather than optimising one, and for treating enrichment against curated knowledge with caution, since it systematically rewards explanations that recover what was already known.

3.9.4. Statistical importance is not biological causality

Attribution methods quantify how much an input contributes to a model's output. They say nothing about whether that input causes the phenotype. Confounding, batch effects, collider bias and selection bias can all drive high attribution for non-causal features, and nothing in the attribution itself distinguishes these cases from genuine mechanisms.

Closing this gap requires integration with formal causal inference, with interventional data such as CRISPR screens or drug-perturbation experiments, and with prospective validation [19, 51]. Until that integration is routine, a claim that an explanation has uncovered a biological mechanism should be read as a hypothesis with a ranking attached, not as a finding.

3.9.5. Human factors and stakeholder needs

A faithful, stable explanation is worth little if its intended audience cannot act on it. Biologists, clinicians, data

scientists and regulators bring different background knowledge, ask different questions and tolerate different levels of complexity, yet user studies comparing explanation formats across these groups remain rare. Designing interfaces that let domain experts query, challenge and refine explanations is a human-computer interaction problem as much as a machine learning one, and it has received correspondingly little attention in this literature [10].

3.9.6. Cost and the barriers to clinical translation

Exact Shapley computation is exponential in the number of features, and even good approximations become expensive at genome-wide or multi-omics scale, which rules out real-time explanation in interactive decision support. Regulatory pathways for AI-based medical devices emphasise transparency but offer little concrete guidance on which methods are acceptable or how they should be validated. Most tellingly, prospective clinical studies measuring whether explanations actually improve diagnostic accuracy, clinician confidence or patient outcomes are almost absent, the one form of evidence that would settle whether XAI delivers clinical value at all [39].

Table 4. Challenges, their consequences and available mitigations.

Challenge	Consequence	Mitigation	Ref.
High dimensionality and collinearity	Unstable or arbitrary feature rankings	Feature selection, pathway aggregation, correlation-aware attribution	[49]
No mechanistic ground truth	Method comparison stays qualitative	Proxy metrics, biological enrichment, synthetic and oracle benchmarks	[8, 25]
Unfaithful explanations	Plausible maps unrelated to model computation	Sanity checks, randomisation tests, ablation of attributed components	[42, 52]
Inconsistent evaluation and reporting	Poor reproducibility, no cumulative progress	Community benchmarks, reporting checklists, released code and seeds	[8, 29]
Correlation mistaken for causality	Spurious mechanistic claims	Causal frameworks, interventional data, experimental follow-up	[51]
Stakeholder mismatch	Low adoption by biologists and clinicians	User-centred design, interactive explanation interfaces	[10]
Computational cost	No real-time explanation at omics scale	Approximate algorithms, distillation, surrogate models	[11, 41]
Absent clinical evidence	Unknown effect on outcomes	Prospective studies measuring diagnostic accuracy and confidence	[39]

Contribution of Table 4. The mitigation column separates the tractable problems from the structural ones. Instability, cost and unfaithfulness all have algorithmic remedies, and progress on them depends on method development alone. The remaining entries do not: ground truth, causal evidence and clinical outcome data must be generated experimentally, and

no improvement in attribution algorithms will substitute for them. Section 3.10 is organised around that division.

3.9.7. What the recent evaluation literature establishes

Table 5 collects the studies on which the preceding assessment rests.

Table 5. Critical evaluation evidence published between 2018 and 2026.

Study	What was tested	Principal finding	Ref.
Sanity checks for saliency maps (2018)	Saliency methods under model and label randomisation	Several widely used methods produce similar maps for trained and randomly initialised models	[42]
Attention is not explanation (2019)	Whether attention weights determine predictions	Attention can be altered substantially without changing model output	[43]
Attention is not not explanation (2019)	The above claim under alternative definitions	Conclusion depends on what an explanation is required to do	[44]
XAI for omics mapping study (2024)	Published XAI applications across omics modalities	DNA-based data dominate; feature-relevance methods are the prevailing paradigm	[7]
Single-cell foundation model interpretability (2026)	Attention-derived edge scores in scGPT and Geneformer	No incremental value over gene-level statistics; ablating top attention heads does not degrade performance	[52]
Concept discovery in scRNA-seq models (2025)	Sparse autoencoder concepts with counterfactual attribution	Concepts were more interpretable to a domain expert than individual neurons	[53]
Sparse autoencoder atlas (2026)	Geneformer and scGPT features against CRISPRi perturbation data	Features are biologically organised, but only 6.2 per cent of transcription factors show target-specific responses	[59]
Trustworthy multi-omics review (2026)	Reproducibility, stability and interpretability across multi-omics studies	The three are studied separately and reported inconsistently; a structured checklist is proposed	[56]

Contribution of Table 5. This table is why the review reaches a more cautious conclusion than the application literature alone would support. The negative results are not isolated complaints about one method; they span saliency, attention and foundation-model interpretability, and in each case the failure is the same in form: an explanation that looks informative while being weakly connected to the computation it describes. The table also shows the productive response,

visible in the final row: not abandoning explanation, but testing it against an intervention rather than against intuition.

3.10. Future directions

Six directions follow from the challenges above. Figure 7 groups them by what each needs to succeed, and Table 6 explicitly maps them to the challenges.

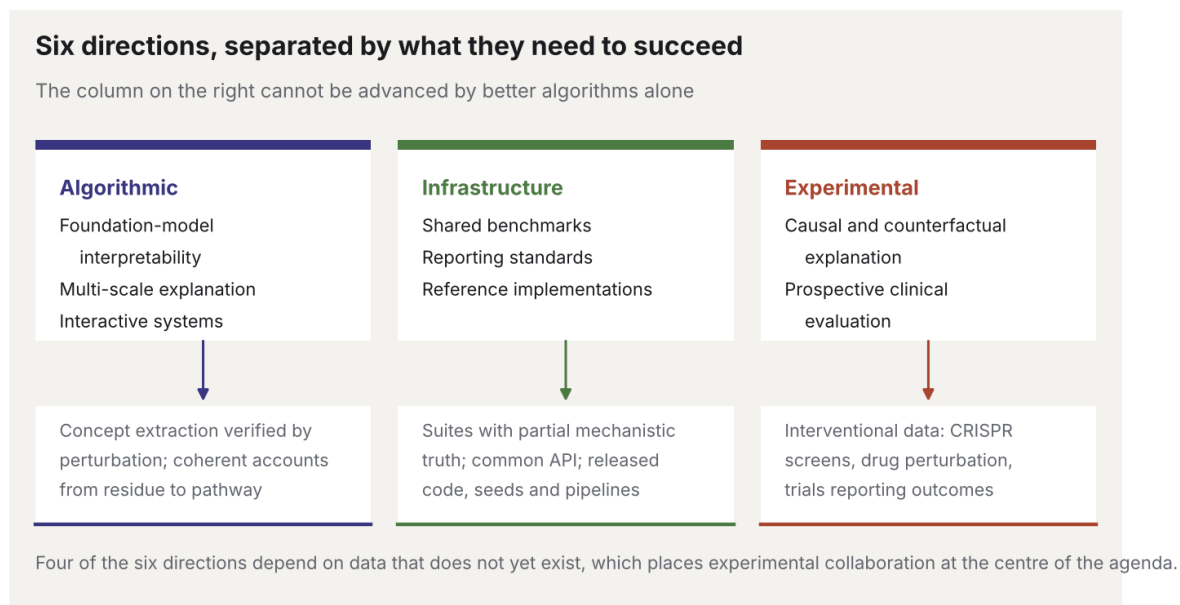


Figure 7. Six research directions, each tied to the challenges it addresses; the directions on the right require experimental rather than algorithmic work.

3.10.1. Causal and counterfactual explanation

The clearest priority is to move from correlative attribution towards explanations that support causal claims. Combining structural causal models, interventional

reasoning or counterfactual generation with deep architectures offers a route to distinguishing drivers from confounders. CausCell illustrates what this looks like in practice: a generative framework that combines a structural causal model with a diffusion process to disentangle single-

cell data into causally related concepts and then generate counterfactual cells, with the counterfactuals serving as the explanation [51]. Adapting such approaches to the dimensionality and sparsity of bulk omics remains open.

3.10.2. Interpretability of biological foundation models

Foundation models trained on DNA, RNA, protein and single-cell data are becoming the default substrate for biological prediction [35, 36, 38, 50], which makes their interpretability a question about the field's whole direction rather than about one model class. The evidence in Section 3.9.3 indicates that attention weights are not the answer [52]. More promising are methods that extract interpretable concepts from internal representations: sparse autoencoders applied to single-cell models recover concepts that a domain expert found more interpretable than individual neurons, with counterfactual perturbation used to establish which genes drive each concept [53]. The limits of that approach are already visible, since sparse-autoencoder features recover pathway structure far more reliably than regulatory causation [59]. Mechanistic interpretability and concept-based explanation defined over biological vocabulary, motifs, domains, pathways, and cell states are the natural next step.

3.10.3. Benchmarks and evaluation protocols

The field needs benchmark suites containing datasets with partially known mechanistic ground truth, whether synthetic, from well-characterised model organisms, or from curated interventional experiments; a defined set of metrics spanning faithfulness, stability, sparsity and biological plausibility; and reference implementations under a common interface. The oracle-model approach from cheminformatics (Section 3.6) and the randomisation tests of Adebayo and colleagues [42] both indicate what rigorous evaluation looks like; what is missing is the shared infrastructure that would make it routine rather than exceptional. A reporting checklist covering stability testing,

multi-cohort benchmarking and model cards would raise the floor immediately, without waiting for the benchmark suites themselves [56].

3.10.4. Human-in-the-loop and interactive systems

Platforms that let biologists and clinicians query a model, request alternative explanations, inject domain knowledge and iteratively refine both model and explanation address a failure mode that better algorithms cannot: an explanation that is faithful but unusable. Treating the domain expert as an active participant rather than a passive recipient of static heatmaps is a socio-technical agenda that has received little attention in bioinformatics relative to its likely impact on adoption [10].

3.10.5. Multi-modal and multi-scale explanation

Biological systems operate across scales, from nucleotides and residues to proteins, pathways, cells, tissues and organisms, and explanations should be inspectable at more than one of them. Hierarchical attention, graph models spanning multiple scales, and concept-bottleneck architectures that compute over intermediate biological concepts all offer substrates for this. P-NET [45] demonstrates the principle at a single scale; extending it across modalities while keeping the units biologically meaningful is the harder problem.

3.10.6. Clinical translation and prospective evaluation

For clinical use, explanation methods must be robust across institutions and patient populations, efficient enough for interactive use, and documented with their failure modes. Regulators must be engaged to define what counts as adequate explainability for a medical device. Above all, prospective studies are needed to establish whether explanations improve diagnostic accuracy, clinician confidence, treatment decisions or outcomes. This evidence is almost entirely absent, and until it exists the clinical case for XAI rests on plausibility rather than demonstration.

Table 6. Research directions mapped to the challenges they address.

Direction	Challenge addressed	What would constitute progress	Ref.
Causal and counterfactual explanation	Correlation mistaken for causality	Attributions validated against interventional data rather than enrichment alone	[51]
Foundation-model interpretability	Opaque representations; unfaithful attention	Concepts in biological vocabulary, verified by perturbation	[52, 53]
Benchmarks and protocols	No ground truth; inconsistent evaluation	Shared suites with partial mechanistic truth and reference implementations	[8, 42]
Human-in-the-loop systems	Stakeholder mismatch	User studies showing improved decisions, not improved ratings	[10]
Multi-scale explanation	Explanations at the wrong level of abstraction	Coherent accounts from residue to pathway within one model	[14, 45]
Clinical translation	Absent outcome evidence	Prospective trials measuring diagnostic accuracy and clinician confidence	[39]

Contribution of Table 6. Pairing each direction with a concrete criterion for progress is intended to make the agenda testable. The third column deliberately excludes the kinds of evidence this field has accepted too readily, a plausible heatmap, an enrichment p-value, a user rating of satisfaction, in favour of outcomes that could falsify a claim: a perturbation that confirms or refutes an attribution, a decision that improves, a trial that reports. As Table 4 shows, four of the six directions depend on data that does not yet exist, placing experimental collaboration at the centre of the agenda rather than at its periphery.

4. Conclusions

Explainable AI has become a standard component of bioinformatics practice. The methods reviewed here, SHAP, LIME, Grad-CAM, attention, biologically constrained networks and graph explainers, have supported biomarker discovery, hypothesis generation, model debugging and bias detection across genomics, transcriptomics, proteomics, sequence analysis, structural biology, imaging and multi-omics integration, and they have done so at a scale that manual analysis could not reach.

Weighing those applications against the evaluation literature gives a more guarded picture. Widely used saliency methods can produce similar maps for trained and randomly initialised models [42]; attention weights can be changed substantially without changing predictions [43]; and in single-cell foundation models, attention-derived scores add nothing over simple gene-level statistics for predicting perturbation responses [52]. None of this makes explanation futile, but it does mean that an explanation's plausibility is not evidence of its faithfulness, and that the field has often accepted the former in place of the latter.

This review classifies the methods, compares them across six tables and seven figures, and separates the open problems by cause. That separation is our central argument. Instability under collinearity, computational cost and unfaithful attribution are properties of current algorithms and will yield to better methods. The absence of mechanistic ground truth, the scarcity of interventional data and the gap between statistical importance and biological causality will not; they require experimental and curatorial work that no algorithm substitutes for. Treating these as one undifferentiated list of open problems obscures where effort is actually needed.

The directions most likely to produce genuine progress therefore combine methodological innovation with experimental coupling: causal and counterfactual explanation tested against interventions, concept-based interpretability for foundation models, shared benchmarks with partial mechanistic truth, interactive systems designed around how experts actually work, and prospective clinical

evaluation. Pursued that way, XAI can become what its proponents claim for it, an instrument of discovery rather than a reassurance applied after the fact. The standard it should be held to is simple enough: an explanation is worth having when it survives being tested, not when it looks convincing.

Author Contributions: Conceptualization, methodology, literature screening, data extraction, formal analysis, visualization, writing (original draft preparation and writing), and review and editing were carried out by the author(s). All authors have read and agreed to the published version of the manuscript.

Funding: This study received no external funding.

Ethical Clearance: Not applicable. This is a literature review and does not involve human or animal subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were generated. All sources cited in this review are publicly available through the respective publishers or preprint servers.

Acknowledgments: The authors thank colleagues who provided informal feedback on earlier drafts of this manuscript.

Conflicts of Interest: All authors declare no conflicts of interest.

References

1. Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583-589.
2. Eraslan, G., et al. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20, 389-403.
3. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793-4813.
4. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160.
5. Zhou, Z., et al. (2023). XAI meets biology: a comprehensive review of explainable AI in bioinformatics applications. *arXiv:2312.06082*.
6. Karim, M. R., et al. (2023). Explainable AI for bioinformatics: methods, tools and applications. *Briefings in Bioinformatics*, 24(5), bbad236.
7. Toussaint, P. A., et al. (2024). Explainable artificial intelligence for omics data: a systematic mapping study. *Briefings in Bioinformatics*, 25(1), bbad453.
8. Budhkar, A., et al. (2025). Demystifying the black box: a survey on explainable artificial intelligence (XAI) in bioinformatics. *Computational and Structural Biotechnology Journal*, 27, 346-359.
9. Samek, W., et al. (2021). Explaining deep neural networks and beyond: a review of methods and applications. *Proceedings of the IEEE*, 109(3), 247-278.
10. Mohseni, S., et al. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4), 24.
11. Ghulam, A., Ali, T., Sikander, R., Ali, F., Masmoudi, A., & Alsini, R. (2026). An Explainable Deep Learning Model for Clathrin Protein Prediction Using a DCT-Enhanced Position-Specific Scoring Matrix. *Current Drug Targets*, 27(5), 345-361.

12. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD*, 1135-1144.
13. Selvaraju, R. R., et al. (2017). Grad-CAM: visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE ICCV*, 618-626.
14. Hartman, E., et al. (2023). Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nature Communications*, 14, 5359.
15. Ghulam, A., Ali, F., Sikander, R., Ahmad, A., Ahmed, A., & Patil, S. (2022). ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemometrics and Intelligent Laboratory Systems*, 226, 104589.
16. Talukder, A., et al. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, 22(3), bbab177.
17. Withnell, E., et al. (2021). XOMiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in Bioinformatics*, 22(6), bbab315.
18. Yap, M., et al. (2021). Verifying explainability of a deep learning tissue classifier trained on RNA-seq data. *Scientific Reports*, 11, 2641.
19. Chen, H., et al. (2024). Applying interpretable machine learning in computational biology: pitfalls, recommendations and opportunities for new developments. *Nature Methods*, 21, 1454-1461.
20. Chen, J., et al. (2021). xDeep-AcPEP: deep learning method for anticancer peptide activity prediction based on convolutional neural network and multitask learning. *Journal of Chemical Information and Modeling*, 61(8), 3789-3803.
21. Gu, Y., et al. (2022). SEGSA-DTA: structure-enhanced graph self-attention network for drug-target affinity prediction. *Proceedings of IEEE BIBM*.
22. Ghulam, A., Sikander, R., Ali, F., Swati, Z. N. K., Unar, A., & Talpur, D. B. (2022). Accurate prediction of immunoglobulin proteins using machine learning model. *Informatics in Medicine Unlocked*, 29, 100885..
23. van der Velden, B. H. M., et al. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 79, 102470.
24. Hussein, A., et al. (2024). Explainable AI methods for multi-omics analysis: a survey. *arXiv:2410.11910*.
25. Arrieta, A. B., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
26. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.).
27. Guidotti, R., et al. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 93.
28. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36-43.
29. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*.
30. Bach, S., et al. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), e0130140.
31. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *Proceedings of the 34th ICML*, 3145-3153.
32. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th ICML*, 3319-3328.
33. Luo, D., et al. (2020). Parameterized explainer for graph neural network. *Advances in Neural Information Processing Systems*, 33.
34. Avsec, Ž., et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18, 1196-1203.
35. Ji, Y., et al. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics*, 37(15), 2112-2120.
36. Rives, A., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), e2016239118.
37. Theodoris, C. V., et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618, 616-624.
38. Cui, H., et al. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21, 1470-1480.
39. Gerlings, J., Shollo, A., & Constantiou, I. (2021). Reviewing the need for explainable artificial intelligence (XAI). *Proceedings of the 54th Hawaii International Conference on System Sciences*.
40. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.
41. Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67.
42. Adebayo, J., et al. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31.
43. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of NAACL-HLT*, 3543-3556.
44. Wiegrefe, S., & Pinter, Y. (2019). Attention is not not explanation. *Proceedings of EMNLP-IJCNLP*, 11-20.
45. Elmarakeby, H. A., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880), 348-352.
46. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), 831-838.
47. Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10), 931-934.
48. Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7), 990-999.
49. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
50. Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. (2025). Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22, 287-297.
51. Gao, Y., Dong, K., Shan, C., et al. (2025). Causal disentanglement for single-cell representations and controllable counterfactual generation. *Nature Communications*, 16, 6775.
52. Kendiukhov, I. (2026). Systematic evaluation of single-cell foundation model interpretability: attention-derived edge scores add no incremental value over gene-level features for perturbation-target prediction. *BMC Genomics*, 27, 634.
53. Claye, C., Marschall, P., Ouerdane, W., Hudelot, C., & Duquesne, J. (2025). Discovering interpretable biological concepts in single-cell RNA-seq foundation models. *arXiv:2510.25807*.
54. Ghulam, A., Arif, M., Unar, A., Thafar, M. A., Albaradei, S., & Worachartcheewan, A. (2025). StackAHTPs: an explainable antihypertensive peptides identifier based on heterogeneous features. *IET Systems Biology*, 19(1), e70002.
55. Arif, M., Fang, G., Ghulam, A., Musleh, S., & Alam, T. (2024). DPI_CDF: druggable protein identifier using cascade deep forest. *BMC Bioinformatics*, 25(1), 1-18.
56. Vo, T. H., & Le, N. Q. K. (2026). Toward trustworthy artificial intelligence in multi-omics: a review of reproducibility, stability, and interpretability. *Briefings in Bioinformatics*, 27(3), bbag227.
57. Ahmed, F., Naz, N. S., Khan, S., Rehman, A. U., Ismael, W. M., et al. (2026). Explainable artificial intelligence (XAI) in medical imaging: a systematic

- review of techniques, applications, and challenges. *BMC Medical Imaging*, 26, 37.
58. Zhang, Y., Zhang, L., Zhang, Z., Lin, Y., Jiang, Z., et al. (2026). Explainable AI multiomics analysis reveals shared and divergent host responses in COVID-19 and influenza. *npj Digital Medicine*, 9(1), 111.
59. Kendiukhov, I. (2026). Sparse autoencoders reveal organized biological knowledge but minimal regulatory logic in single-cell foundation models: a comparative atlas of Geneformer and scGPT. *arXiv:2603.02952*.

