Molecular Cell

CellPress

Perspective

Toward interpretable foundation models for molecular biology

Mukund Varma[1,*] and Trey Ideker[1,2]
[1]Division of Human Genomics and Precision Medicine, Department of Medicine, University of California at San Diego, La Jolla, CA 92093, USA
[2]Big Data Institute, Nuffield Departments of Medicine and Population Health, University of Oxford, Oxford OX3 7LF, UK
*Correspondence: mvarma@health.ucsd.edu
https://doi.org/10.1016/j.molcel.2026.09.030

SUMMARY

For machine learning models in molecular biology, explanation is as important as prediction. Encoding prior knowledge and treating interpretability as a first-class design objective accomplish both. We provide design elements and a checklist for the next generation of foundation models in molecular biology that are as interpretable as they are powerful.

INTRODUCTION

The first draft human genome, assembled and released in 2001, achieved data collection and analysis at a scale unprecedented for biology.[1] In the two decades since, our ability to generate large-scale datasets capturing biology at every resolution has only increased. Where Mendel reasoned about seven discrete traits across a garden of pea plants, the modern biologist must make sense of whole genomes in giga base pairs, tens of thousands of genes across millions of single cells, and millions of pixels per image. The features that carry the signal are numerous, continuous, and entangled and few of them have names. Machine learning and AI provide an avenue for analyzing these data.

Outside the life sciences, the reflexive answer to escalating data complexity has been to match the increase in data with a corresponding increase in model size. We add more parameters, more pretraining, more compute, and we let capacity absorb what we do not know how to specify. A language model trained on text in a hundred languages will translate between a pair it was never explicitly trained to translate. Biological foundation models are built on the same idea. A foundation model is trained once at scale on broad data and then adapted to many specific tasks, which lets a downstream lab fine-tune on a cohort far smaller than training from scratch would require (Box 1). Train on enough examples, and the model will learn something general about biology, enough to transfer to a new assay, tissue, or disease from a small number of labeled cases.

This approach is especially valuable in biology because, in many important settings, dataset sizes are fixed and cannot easily be expanded, either because of cost or fear of introducing batch effects. A simple model trained on a single dataset may lack power or may fit noise before it fits biology. A foundation model mitigates these specific risks because it already "knows" some biology by virtue of having seen hundreds of thousands of samples and so the small dataset only has to specialize that knowledge rather than build it.

But the use of foundation models as a backbone introduces other risks—the risk to interpretability and the risk of propagating unstated biases. Many groups will typically reuse a published foundation model, so a wrong prior or an opaque representation stays inside and propagates to every group that adapts it. Interpretability therefore has to be a construction-time property. The consequences of these risks are amplified in biology because verifying by experimental validation carries a much higher temporal, monetary, and moral cost than many other disciplines. A reader can check LLM-generated text by reading it in seconds, and a bad paragraph costs a rewrite. Checking a biological model's claim that a molecular pathway drives resistance in a patient costs an experiment, a cohort, or a trial. The answer arrives months later, and the cost includes misformulated animal studies and suboptimal patient outcomes.

Importantly, very often in biology and medicine, the explanation itself is a key part of the question and what the model has to deliver. Understanding which component of a machine-designed antibody makes it bind is important for development of future medicines. Understanding which genes are implicated in disease state prediction informs the selection of novel targets and understanding which pathways are implicated in drug response prediction helps define mechanisms of resistance and response.

Accordingly, this piece is a *call to action* for a new landscape of foundation models in molecular biology that are interpretable, not just predictive. We describe design elements to guide the development of such models, grouped into four core principles. First, encode biological priors so that they constrain a model rather than decorate it. Second, make a constrained model interpretable rather than merely labeled. Third, interrogate and simplify a trained model. Finally, validate the model and its predictions. Few foundation models yet apply these principles at scale, so many of our examples are established, interpretable models built for a single task. We suggest resources for encoding prior knowledge (Figure 1), design elements to make foundation models interpretable by construction (Table 1), and a

All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Box 1. Terms used in this article

Ablation: removal or corruption of one component of a model, followed by re-evaluation or retraining, in order to measure the contribution of that component.
Attention head: one of several parallel operations within a transformer layer, each computing the influence that every element of the input exerts on every other.
Attribution: a method that apportions a prediction among the input features. Saliency maps, Integrated Gradients, and SHAP are attribution methods.
Backbone: the pretrained component of a foundation model, which downstream tasks reuse without modification.
Concept bottleneck: a layer of human-specified variables through which every prediction must pass. Because the prediction depends on the input only through these concepts, a concept can be set to a chosen value and its effect observed in the output.
Contrastive objective: a training objective that draws pairs designated as similar together in the embedding space and separates other pairs.
Embedding: a vector representation that a model computes for an input such as a cell, a molecule, or a protein. Distances within this space encode the similarity the model has learned.
Equivariance and invariance: properties that specify how a model's output must respond to a transformation of its input. An invariant output is unchanged by the transformation. An equivariant output undergoes the corresponding transformation. SE(3) denotes the group of rigid rotations and translations in three dimensions. An SE(3)-equivariant model therefore returns an equivalent structure under any choice of coordinate frame.
Fine-tuning: adaptation of a backbone to a downstream task by continued training on labeled examples. The parameters of the backbone are updated.
Foundation model: a model trained once at scale on broad data and subsequently adapted to multiple downstream tasks. The initial training is termed pretraining.
Frozen embeddings: embeddings obtained from a backbone whose parameters are held fixed.
Graph-Laplacian penalty: a training penalty that increases with the difference in value between connected nodes. It discourages departure from a curated graph without prohibiting it.
Latent space: the space of embeddings a model can produce. A latent dimension is a single coordinate of that space.
Likelihood: the probability a model assigns to the observed data under its assumptions. A negative binomial likelihood accommodates discrete counts that exhibit greater dispersion than a Poisson model permits, as observed in single-cell data.
Linear probing: adaptation of a backbone by training a separate, typically linear model on its embeddings. The parameters of the backbone remain unchanged.
Masked connectivity: a binary matrix that fixes selected weights at zero, restricting the network to the connections the prior permits.
Message passing and graph convolution: an operation that updates each node from the values of its neighbors in a graph. Repeated application propagates information along paths in the graph.
Pruning: removal of parameters, nodes, or edges from a trained model while performance is preserved. Structured pruning removes entire units rather than individual weights.
Token and token vocabulary: the discrete units over which a model operates. In biological models a token may denote a gene, a residue, or a named transcriptional regulator. The vocabulary is the fixed set of tokens available to the model.

checklist for authors and reviewers to consider when evaluating foundation models (Table 2). As the field moves from isolated biomarkers toward richly featured, multimodal representations of biology, interpretability, and visibility must be intrinsic to the models themselves for their outputs to be actionable.

ENCODING PRIOR KNOWLEDGE

Molecular biology is unusually rich in structured knowledge, curated over decades of experimental work. Much of it now exists in machine-readable form, from ontologies such as Gene Ontology (GO)[21] to interaction networks such as STRING.[22] A model that ignores these resources must recover known biology from data that are finite and expensive to produce. Where the data do not determine the answer, the model is free to fit a technical bias, a batch effect, or a shortcut feature instead. An encoded prior removes that freedom. It supplies the part of the answer that is already known, so the data are spent on the part that is not.

A prior earns its place in a model by excluding or penalizing something that goes against it. If the encoded structure does not narrow or reweight the set of hypotheses the model can express, it is decoration rather than a constraint. Encoded priors carry the biases, omissions, and errors of the resources that produced them. Functional gene-gene or protein-protein interactions measured in a systematic assay, from literature curation or from computational inference all enter the model in the same form, so the choice of resource dictates the level of evidence behind the prior. Encodings should therefore consider inclusion of a mechanism to overrule the prior when the data disagree. One mechanism softens the prior. A hard prior permits no deviations, whereas a soft prior acts as a guide. A masked weight matrix is a hard prior that forbids a connection, and its soft analog is a graph-Laplacian penalty that discourages departures from known interactions but

TOPOLOGY	BIOLOGICAL PRIOR	HOW IT ENTERS THE MODEL
1 Graphs of relations		
	**Undirected graph** nodes, unsigned edges **PRIOR** Physical and functional interaction; co-membership	Graph convolution or message passing on the adjacency matrix; graph-Laplacian regularization of weights; masked connectivity between layers; one attention head per network **SOURCES** STRING (Szklarczyk 2023); BioGRID (Oughtred 2021); IntAct (Del Toro 2022); HuRI (Luck 2020); OmniPath (Türei 2021)
	**Directed, signed graph** **PRIOR** Regulation and causal influence	Directed message passing; signed linear operator applied to expression; constrained transition or propagation matrix in a dynamical model **SOURCES** DoRothEA (Garcia-Alonso 2019); CollecTRI (Müller-Dott 2023); TRRUST (Han 2018); ENCODE cCREs (2020)
2 Hierarchies and sets		
	**Rooted DAG** multiple inheritance, nested containment **PRIOR** Hierarchical organization of function and structure	Layer-by-layer masked connectivity mirroring the DAG, so each hidden unit is a named system ("visible" architectures) **SOURCES** Gene Ontology (Ashburner 2000); Reactome hierarchy (Milacic 2024); NeST (Zheng 2021); Cell Ontology (Diehl 2016)
	**Tree** single inheritance, ultrametric or additive distances **PRIOR** Shared ancestry and descent	Column- and row-wise attention over an alignment; coevolutionary coupling terms; tree-structured priors on parameters or latent states **SOURCES** Multiple sequence alignments and species phylogenies; Zoonomia and phyloP constraint (Genereux 2020); CRISPR lineage-tracing data
	**Set membership** flat cover; overlapping sets **PRIOR** Program, pathway, and complex membership	Binary mask on a decoder or first hidden layer, so one latent unit maps to one annotated program; sparse factor loadings **SOURCES** MSigDB Hallmark and C2 (Liberzon 2015); KEGG (Kanehisa & Goto 2000); WikiPathways (Pico 2008); CORUM (Tsitsiridis 2023)
3 Order along a chain		
	**1D chain** ordered positions, translation structure **PRIOR** Sequence order, motif syntax, genomic distance	Convolution as translation-equivariant motif scanning; dilated receptive fields; distance-aware attention; thermodynamic binding parameterization **SOURCES** JASPAR (Castro-Mondragon 2022); HOCOMOCO (Kulakovskiy 2018); CIS-BP (Weirauch 2014)
4 Geometry and space		
	**Euclidean geometry** SE(3): rigid rotations and translations **PRIOR** Rigid-body symmetry, distance geometry, physical realizability	Equivariant or invariant attention and message passing on coordinates; triangle-inequality-consistent pair updates; frame-based coordinate prediction **SOURCES** PDB (Berman 2000)
	**Spatial neighborhood graph** 2D/3D tissue coordinates **PRIOR** Physical proximity and local microenvironment	Graph convolution over the neighborhood graph; niche-level latent variables; latent dimensions tied to annotated ligand-receptor programs, with de novo programs learned alongside **SOURCES** *k*-nearest-neighbor or Delaunay graphs built on measured coordinates; ligand-receptor priors from OmniPath or CellPhoneDB

Figure 1. Structured priors in molecular biology organized by topology
Two priors with the same topology are encoded the same way, whatever their biology: a signaling network and a spatial neighborhood graph both become message passing, and an ontology and a protein-complex hierarchy both become masked, layer-wise connectivity. Resources are identified by name, first author, and year.

allows a new interaction when backed by the data. Another way to permit the model to learn novel biology while incorporating priors is to allow for *de novo* terms beside the annotated ones. Spectra, which factorizes single-cell expression under a gene-set prior, shows what such a term can yield.[7] Of 197 factors inferred to describe breast tumor immune cells, 26 depended little on the prior gene-set knowledge base. One of these, a macrophage program, matched no input gene set, expanded only in patients who did not respond to anti-PD-1, and reappeared in an independent cohort. A *de novo* term earns its place by capturing what the supplied annotation cannot express, whether or not the field already knows it.

Table 1. Design elements for interpretable foundation models, the models that exemplify them, and where each could be extended

Design element	Exemplar and what the prior buys	A model that could be extended	The extension and what it would provide
Match the encoding to the topology; the mathematical object should represent the biological structure faithfully	Novae (Blampey et al.),[2] a spatial transcriptomics foundation model; builds a Delaunay graph over cells and encodes it with a graph attention network, so a niche is a neighborhood rather than a coordinate (Figure 1, spatial neighborhood graph); a cell's representation is built from its neighbors, which is the unit a spatial domain is defined over	TANGLE (Jaume et al.),[3] a slide representation model; encodes the transcriptome as a flat 1,000-gene vector through a three-layer perceptron	enter the transcriptome as named pathway tokens, as SurvPath does with several hundred Reactome[4] and Hallmark[5] terms (Jaume et al.)[6]; a slide-level prediction could then be attributed to named pathways rather than to one pooled vector
Choose a hard or a soft constraint deliberately; prefer soft where the prior is a curated resource that may be wrong or incomplete	spectra (Kunes et al.),[7] a factor model of single-cell expression; represents input gene sets as a gene-gene graph and estimates, for every factor, how far it departs from that graph (Figure 1, set membership, encoded as an undirected graph); the departure is therefore reported: of 197 factors, 171 stayed close to the graph and 26 did not	MutationProjector (Kong et al.),[8] a tumor-genotype foundation model; confines each encoder unit's eight attention heads to eight curated interaction networks with a binary mask	replace the masks with a graph-Laplacian penalty or add a ninth unmasked head and report which *de novo* edges appear; both these options provide a path for the model to learn new interactions, while still leveraging the structure provided by the input knowledge graphs
Make the prior narrow or reweight the hypothesis space; a structure that does neither is decoration rather than a constraint	AlphaFold2 (Jumper et al.),[9] a protein structure predictor; encodes a physical symmetry in frame-based attention, so a predicted fold does not depend on the coordinate frame (Figure 1, Euclidean geometry); orientation is removed from what the network must learn, and the symmetry holds for every input	CellVQ (Wang et al.),[10] a single-cell foundation model; adds a trainable ontology embedding to the gene and position embeddings, ahead of twelve standard transformer layers	freeze the ontology embedding or penalize departure from it, route its edges through restricted attention, and report how predictions change when the prior is removed, randomized, or contradicted; the ontology might then narrow what the model can express, and a reader could tell how much of a prediction it carries
Base the objective function on a biological target; a model should be trained to recover real biology, not an arbitrary target	Orthrus (Fradkin et al.),[11] an RNA foundation model; positive pairs are splice isoforms of one gene and orthologs across 400+ mammalian species (Figure 1, tree); the objective therefore states an explicit hypothesis that evolutionary relatedness predicts shared function; pairs the trained model holds apart become candidates for isoform-specific or lineage-specific function	scGPT (Cui et al.),[12] a single-cell foundation model; masks a subset of gene-expression values in each cell and trains the model to reconstruct them	draw the mask from biology, for example from a perturbation experiment or an interaction network, so biology sets what the model must recover; the objective might then state a hypothesis, as the Orthrus objective does; the scGPT authors themselves name perturbation and temporal data in pretraining as a direction of interest
Draw the names of visible units from a trusted resource, not from intuition	ChromBERT (Yu et al.),[13] a foundation model of transcriptional regulation; takes about 1,000 named transcription regulators from a curated chromatin immunoprecipitation sequencing (ChIP-seq) compendium as its token vocabulary, fixed before training rather than labeled afterwards; each token a user reads is a named regulator	CellWhisperer (Schaefer et al.),[14] a chat interface to transcriptomes; aligns each transcriptome with a free-text summary that a language model writes from the sample metadata	anchor the text side on human-curated terms such as MSigDB[5] gene sets; each direction in the joint space might then refer to a term a reader can look up

(*Continued on next page*)

Table 1. *Continued*

Design element	Exemplar and what the prior buys	A model that could be extended	The extension and what it would provide
Pass the prediction through named concepts and admit *de novo* terms alongside them; the concepts carry the annotated explanation, and the *de novo* terms expose the biology the annotation leaves out	CB-pLM (Ismail et al.),[15] a protein language model; every prediction passes through named Swiss-Prot concepts alongside an explicit unknown embedding for information the curated vocabulary does not capture; a concept can be edited and the effect read from the output, and the unknown embedding shows where the vocabulary ends	Novae (Blampey et al.),[2] a spatial transcriptomics foundation model; assigns every cell to a learned prototype, under a constraint that spreads assignments evenly	add an unknown prototype exempt from that balancing and report what fraction of tissue lands there; every cell now receives one of the learned domains, including tissue unlike any seen in training; an unknown prototype might flag such tissue, as the unknown embedding does in CB-pLM
Parameterize the model mechanistically; each learned parameter or each term of the output should be a quantity with its own meaning, so an explanation is read from the model rather than estimated from it	MACE-OFF (Kovács et al.),[16] a force field for organic molecules; decomposes total energy into a sum of per-atom terms, with forces as the analytic gradient (Figure 1, Euclidean geometry); the energy assigned to an atom is exact, because the per-atom terms sum to the prediction; the decomposition is a property of the output; the weights that produce each term carry no individual meaning	MoLFormer (Ross et al.),[17] a chemical language model; predicts a molecular property from the mean of the last hidden state	predict a decomposable quantity, a sum over atoms or fragments, so a property can be attributed to the atoms that carry it; the published attention analysis describes the representation, and a decomposable output might describe the prediction as well; not every property is additive over atoms, so the extension applies only where one is
Incorporate the process that produced the data; the model should encode how the measurement was made and not only what it measured	SaProt (Su et al.),[18] a protein language model; pairs each residue with a Foldseek 3Di state in a 441-token vocabulary, carrying AlphaFold structure only where it was called above pLDDT 70, so the token records how far the structure can be trusted	Geneformer (Theodoris et al.),[19] a single-cell foundation model; represents each cell as a rank ordering of its genes, with no assay or batch covariate and no likelihood over counts	add a count likelihood with library size and batch as covariates, as scVI does,[20] and report whether rank ordering still helps; technical variation might then be modeled rather than left to the embedding to absorb

Each design element is a construction-time choice. The checks that verify a model once it is built are given in Table 2. The models are examples rather than a complete set, and several satisfy elements beyond the one they appear under. Where a topology in Figure 1 applies to an exemplar, column 2 names it. Columns 3 and 4 describe one model per element that could be extended to follow it and state the extension. These models were chosen because their designs are strong and widely adopted, which is what makes the extension worth making. Each extension is a possibility whose benefits and trade-offs remain to be assessed. Models are identified by name, first author, and year, and each is entered in the numbered reference list.

A prior is a hypothesis about biology, and each kind carries a different one. Such assumptions are distinct from the errors and biases of a resource. The hypothesis is carried by the encoding rather than by the resource, so a prior can fail where the resource behind it is accurate. An interaction network asserts that genes joined by an edge behave similarly and that assertion fails wherever an edge carries the opposite sign in a different tissue. A hard mask asserts that the curated resource is complete because the model cannot express an interaction that the resource omits.

Choosing a prior involves two additional decisions: which biological structure to trust and which mathematical object represents that structure faithfully inside a differentiable model (Table 1). Priors can be used to provide the model with knowledge about the biology being studied, physical constraints driving these systems, the data generation process and, for foundation models, the relationships a pretraining objective must capture.

Our DCell approach[23] and successors such as P-NET,[24] KPNN,[25] GenNet,[26] and NeST-VNN[27] are examples of neural networks that provide a direct biological scaffolding to the model by restricting the connections in a multilayer perceptron to a curated ontology, turning the neural network into a "visible" representation of the cell, where each node corresponds to a gene, assembly, or system and each edge denotes membership. A very different kind of prior is one of those used by AlphaFold2, whose frame-based attention encodes a physical symmetry, so a folded shape does not depend on its coordinate frame, and orientation is removed from what the network must learn.[9] AlphaFold3 supplies the same prior by a different route. Its diffusion module is not equivariant, and its authors state that no invariance or equivariance with respect to global rotation and translation is required in the architecture, so they omit it.[28] The symmetry is supplied during training instead by rotating and translating each structure at random. The two routes differ in what a reader may assume. An enforced

Table 2. A checklist for interpretable-by-construction models

Question	What a satisfactory answer states
Checks of explanation quality	
Does removing or isolating a visible unit change the prediction?	a unit was removed, set to a fixed value, or used alone, in the trained model or in a model trained again, and the change in performance is reported for each task
Is every path that departs from the prior or bypasses the visible units reported?	each route by which signal reaches the output without passing through a red unit is named, and the share of the prediction it carries is quantified
Is each unit's contribution reported across training runs and not from a single seed?	the contribution was measured in runs that differ only by seed, and the variation across runs is reported
Is each element that carries an assumption removed or scrambled and its contribution reported?	every prior, objective, and architectural constraint was removed or randomized in turn, and the change in performance is reported for each task whose explanations are read
Do different attribution methods agree?	a second method from a different family was applied to the same model, and the two rankings are compared
Is the reduced structure reported and scoped to one task and dataset?	the reduced model is reported with the task and dataset it was scoped to and with its performance against the full model
Checks of predictive performance	
Is utility shown on more than one task?	performance is reported on tasks that differ in label and in biology and not only in cohort
Is accuracy compared against a strong single-task baseline?	a single-task model trained on the same data is reported, and it is one a practitioner would use
Do frozen embeddings place biological and technical replicates close together?	replicates of one biological state, and one sample measured twice, occupy the same region of the space
Do frozen embeddings use the dimensionality of the space efficiently?	the number of dimensions that carry variation is reported, rather than the nominal width of the layer
Is accuracy reported for every regime interpreted?	accuracy is reported separately for every cohort, task, and condition whose explanations are read
Checks of biological validity	
Has a prediction been tested outside the training data, and was the test computational or experimental?	the prediction was tested on data withheld from training, and the report states which of the two kinds of test was made

symmetry holds for every input. A learned one holds where training covered it, and that is a claim to be checked. Such symmetry can come from biology rather than physics. Caduceus holds a DNA strand and its reverse complement equivalent, so a genome model is not left to discover that they are the same molecule.[29] A prior can also describe the data generation process, as in scVI, which encodes discrete, overdispersed single-cell counts in a negative binomial likelihood with batch and library-size covariates.[20] In a recent benchmark of clustering and batch integration, scVI matched or outperformed transformer-based models carrying more than an order of magnitude more parameters, none of which encode the measurement process.[30] Finally, Orthrus is a foundation model trained contrastively on transcript pairs, either splice isoforms of one gene or orthologs across species, so the objective states a hypothesis that evolutionary relatedness predicts shared function.[11] Transcript pairs that the trained model holds apart are then candidates for isoform-specific or lineage-specific function.

In a foundation model, the reach of a prior is limited in two ways. A prior in the architecture or the pretraining objective is fixed for everyone who adapts the backbone, so a downstream lab inherits it and cannot correct it. Only a prior added at adaptation is still open to that lab. SurvPath is a transcriptomics and histopathology-based multimodal model that leaves its foundation model-derived histopathology encoder frozen and encodes the transcriptome as several hundred Reactome and Hallmark pathway tokens.[6] The prior thus sits only in the task model. Experiment-guided AlphaFold3 adds a prior one stage later still. Restraints from nuclear magnetic resonance, crystallography, and cryogenic electron microscopy are imposed at inference on a released backbone, by a group that never trained it, recovering conformational ensembles the backbone does not otherwise produce.[31] The authors of that backbone had removed the priors their own predecessor carried, and a downstream group supplied one from outside. A prior can also be bypassed inside the model. A skip connection, a parallel path, or an extra embedding carries signal around the constrained structure, so the model can predict well for a reason the prior does not carry. Where such a route is left open, the interpretability claim covers only the share of the signal that still passes through.

When choosing priors, authors should state which resource each came from, which mathematical object encodes it, whether it is held hard or soft, and what route the data has to overrule it.

INTERPRETABILITY BY CONSTRUCTION

Embedding prior biological knowledge in a network is an effective way to constrain what it should and should not learn. But a constrained model is not necessarily an interpretable one.

Interpretability is a separate design objective. An ontology or an interaction graph narrows the hypothesis space, yet the

variables inside the resulting network may remain distributed across many units, correlated with one another, and entangled in ways that no biological label resolves. We turn again to scVI's example. It encodes a genuine prior, the negative-binomial noise model of the assay (see above, encoding prior knowledge). It is accurate, but its latent space is not directly interpretable: the nonlinear decoder spreads each biological signal across many dimensions, which are read out only by clustering. Replacing that decoder with a linear one, as in LDVAE, holds the prior and the encoder fixed and turns each latent dimension into a gene-loading vector that names a transcriptional program.[32]

A model is interpretable by construction when it asserts, before training, that a given internal variable stands for a given piece of biology. We call such a variable a visible unit, after the visible representation of the cell and tissues above, whether it is a subsystem node, a biochemical concept, an attention head, or a latent dimension. Several design elements make a unit readable (Table 1). No model follows every element. Novae presents both as an exemplar and as a model that could be extended, and our own MutationProjector is among the models that could be extended. Some architectural families deliver several of those elements at once, and the four below are representative rather than exhaustive. First, concept bottlenecks route a prediction through a layer of human-specified concepts. In CB-pLM, that representation combines several hundred human-specified Swiss-Prot concepts with an explicit unknown component for information outside the vocabulary. A named concept can be inspected and edited, and the effect of that edit read from the output, while the unknown component makes the limits of the concept vocabulary explicit.[15] Second, visible neural networks encode a hierarchy, so each hidden unit is one subsystem that can be perturbed on its own.[27] Third, modular architectures partition the computation along known biological boundaries: MutationProjector assigns each attention head to one molecular interaction network, so message passing is confined to relationships specified in advance.[8] Fourth, mechanistic parameterizations give each learned parameter its own meaning: CellBox fits ordinary differential equations to perturbation data, so every parameter is a signed, directed interaction between molecular species.[33]

A foundation model offers two stages at which a user can extract an explanation, and the builder decides what is available at each. At the pretraining stage, the object of study is the backbone, the part trained once on broad data and reused unchanged by every downstream task. At the fine-tuning stage, it is the task model that a downstream lab trains on top of the backbone, on its own labels. Visible units make both readable.

MutationProjector shows the result at both stages. Its pretraining projects a tumor's alterations onto interaction networks drawn from curated resources, so the pretrained model reports the affected networks of a tumor before any downstream task exists, and the same projection reconstructs held-out mutations.[8] Applied to immunotherapy and chemotherapy resistance across several cancer types and cohorts, it returns named biomarkers, including KMT2D alteration in immunotherapy sensitivity and joint alteration of SMARCA4 and STK11 in resistance. The explanation at each stage arrives in the units a biologist already uses, networks and genes, because those units were built into the model.

How much of the explanation ascribed to a unit survives adaptation depends on whether the weights of the backbone change. A frozen backbone keeps every visible unit and every check made on it. MutationProjector is used this way, since each clinical task trains a random forest on the fixed embedding.[8] Fine-tuning updates the weights. A hard constraint such as a mask survives, because it is part of the structure, but the weights inside it change, so a unit keeps its name and can change its contribution. A soft penalty or a learned property carries no such guarantee, so the checks must be repeated on the fine-tuned model.

Omitting visible units from the architecture moves the burden of interpretability to the downstream user of the foundation model. Recovering biological meaning from the model's embedding dimensions takes a separate analysis of the backbone, whose success is discovered after the fact. Authors should state what a user can extract at each stage and ship whatever makes extraction possible.

Interpretability of this kind is a property of the model itself, established at construction, and it is distinct from post hoc explanations (considered in the next section below), which are properties of an analysis applied to a model already trained. The labeled internal components or visible units that permit a model to be checked are the same components that permit controlled intervention on it, in which a concept is set and its consequence recorded, and this capacity allows a model to be used as an instrument for generating hypotheses rather than only as a predictor. Setting a visible unit and recording the consequence produces a hypothesis. The unit's label is a claim about the model that this variable stands for that piece of biology. Whether the biology behaves as the label implies is a separate claim, and validating it requires an experiment (see below, computational and experimental validation).

INTERROGATING AND SIMPLIFYING MODELS

Once a model is trained, two operations ask what it has learned. Attribution describes the trained model as it stands, while ablation changes it and reads the consequence. Attribution is therefore a description, and ablation is a test.

Attribution methods divide into two families. Gradient and propagation methods read the model's sensitivity to each input from its internals, as saliency maps, Integrated Gradients, DeepLIFT, and the Shapley-value methods SHAP and DeepSHAP do. *In silico* perturbation methods change the input and record the response, as saturation mutagenesis and occlusion do.[34] Both families report which features drive model performance, which is often conflated with those features driving biology. An attribution can rank a feature highly when it only covaries with a true driver that was not measured, so the biological claim needs external experimental validation.[35] Sparse autoencoders are a recent development and offer another related way to interrogate a distributed representation by splitting it into a much larger set of sparsely active features.[36]

A typical biological model has too many parts to remove one at a time, so the ablations worth running target each logical unit that

carries an assumption, such as a biological prior (see above, encoding prior knowledge). The ablation method has to match the shape of the encoded prior. Rerunning with a degree-preserving random graph, or a shuffled assignment of genes to sets, separates a model that made real use of the biology from one that exploited only the shape of the constraint.

A smaller model is easier to interrogate, and in biology parsimony is often a first-class concern rather than a modeling preference. A model that retains fewer parameters, nodes, or edges is easier to interrogate because there are fewer units to name and fewer candidate explanations to enumerate. It is also easier to operationalize: a five-gene biomarker panel deploys in clinics worldwide on cheap, simple assays, whereas a signature that demands whole-genome sequencing does not. Sparsification and structured pruning reduce both model complexity and assay burden. When pruning removes every predictive path from an input gene, it also performs feature selection and can reduce a 500-gene signature to a deployable five-gene panel. The pruned model, not the fully parameterized one, is then the object under study.

Pruning a trained model is not without risk, especially for foundation models. A part that contributes nothing to the pretraining objective may be exactly what a later fine-tuning task requires and that task is not visible at the time of pruning. More generally, pruning finds the parts sufficient for the prediction task but not the parts sufficient for the biology. Parsimony is therefore a claim about one task and one dataset, and it should be scoped to them rather than read as a statement about the biology.

COMPUTATIONAL AND EXPERIMENTAL VALIDATION

Foundation models are not built for a single task. They are backbones for fine-tuning, which updates their weights on a few labeled examples, or linear probing, which trains a separate model on their frozen embeddings. Computational assessment of a foundation model should therefore take a different form than for a single-task model. Accuracy has to be reported across the range of downstream tasks the backbone might plausibly be adopted for and against a strong single-task model on each. The representation itself must also be assessed. Inputs that a biologist would treat as equivalent should sit close together in that space. A chemical foundation model that places canonical and non-canonical SMILES for one molecule far apart is representing the notation rather than the chemistry, a test that many current models fail.[37] The latent space can also be assessed without labels. RankMe reads the effective rank of the embedding matrix from its singular values and shows how many of its dimensions a model actually uses.[38]

In molecular biology, a machine learning model is rarely the final product. Its output feeds a decision to commit bench time, reagents, and biological material: a target to pursue, a molecule to synthesize, and a condition to screen. The standard is no different for a foundation model, whose deliverable is the hypothesis that reaches the bench rather than its test-set accuracy. A model that does not make that path straightforward moves the work onto the downstream user. Visible units shorten it, because a unit that stands for a pathway, a concept, or a regulator names something that can be manipulated in the physical world.

Geneformer demonstrates this path. Fine-tuned to separate cardiomyocytes from failing and non-failing hearts, it nominated GSN and PLN as targets whose inhibition should move diseased cells toward the healthy state. CRISPR knockout of both in TTN+/− cardiac microtissues improved their contractile stress.[19]

A CHECKLIST FOR INTERPRETABLE-BY-CONSTRUCTION MODELS

The requirements above reduce to a set of questions an author can answer in the methods of a paper and a reviewer can check (Table 2). They fall into three kinds. Checks of explanation quality ask whether the visible units carry the prediction and whether what the model reports is what it used. Checks of predictive performance ask whether the model is accurate enough, and across enough regimes, for its explanations to be worth reading. Checks of biological validity ask what has been tested outside the data the model was trained on. Authors should also state what a user can extract at each stage, since a released backbone, a fine-tuned model, and a frozen embedding do not offer the same access. We provide a worked example, applying this checklist to a model of our own (Box 2).

OUTLOOK: DESIGNING BETTER FOUNDATION MODELS

Many foundation models in biology have copied the transformer from the domain of human language but not the reason it works well there: masked-token pretraining learns the structure of language, but it does not on its own capture complex biological interactions within a cell. This is most apparent in the field of single-cell transcriptomics, where in several recent benchmarks, scale-only single-cell foundation models have failed to outperform strong simple baselines consistently.[30,39]

Each earlier section in this article supplies a piece of what designing interpretable foundation models that properly leverage the structure of biology could look like. The priors and structure that constrain a single-task model (Table 1) can be built into a foundation model, so that part of what it knows is specified rather than learned (see above, encoding prior knowledge). Parsimony, linearity, and modularity can be selectively used to make models interpretable by construction (see above, interpretability by construction). Task-specific models can be interrogated, pruned, and simplified to generate *in silico* hypotheses about new biology (see above, interrogating and simplifying models). And finally, these interpretable components can be extracted to feed the next iteration of experimentation (see above, computational and experimental validation).

Each of these goals requires that interpretability be built into the model. This requirement does not by itself deliver any of the goals themselves, which is why the checks above accompany the design elements. The backbone of a released foundation model may feed clinical and other high-stakes decisions, and adoption will require predictions that can be checked at the point of use. A design that is interpretable by construction supplies that check to a user who did not build the model. Post hoc analysis complements such a design, and it is the only option for a model

Box 2. Applying the checklist to MutationProjector

We apply the checklist to a model of our own, so that every answer can be given in full, including the answers that record a check we did not run. MutationProjector is pretrained on 30,328 solid tumors across ten cancer types, representing alterations in the 468 genes of a clinical sequencing panel. Each encoder unit carries eight attention heads, each confined to one of eight curated molecular networks, covering physical binding, transcriptional regulation, phosphorylation, ubiquitination, genetic interaction, DNA damage repair, STRING, and PCNet. Two encoder units give sixteen heads in total (Kong et al.).[8]

Each row restates one question of Table 2 as a requirement, in the same order. Satisfied means the published work meets it. Partially reported means the work supplies part of what the requirement asks for. Not reported means the check was available and was not made.

Requirement	Verdict	What the published work provides
Checks of explanation quality		
Removing or isolating a visible unit is shown to change the prediction	satisfied	each of the eight networks was used alone, which sets the other heads to zero, and performance is reported per task and cohort; the variants show that the eight networks differ in content
Every path that departs from the prior or bypasses the visible units is reported	partially reported	covariate nodes connect to every gene in every head, so a gene-to-covariate-to-gene path runs outside the curated networks; the mask forbids a direct departure from the prior, and this path permits an indirect one; the share of the prediction that belongs to this path is not quantified
Each unit's contribution is reported across training runs and not from a single seed	not reported	the masks fix which network each head reads, so a head has the same meaning in every training run; the contribution of each head was measured in one run; a repeat of the readout across seeds is a direct extension
Each element that carries an assumption is removed or scrambled and its contribution reported	satisfied	three families, reported per task and per cohort across two supplementary tables: each pretraining task removed in turn, the graph removed, and the graph randomized; shuffling node labels costs almost nothing on cancer type, 0.838 against 0.837, and a great deal on drug response, 0.68 falling to 0.53 for bladder chemotherapy
Attribution is shown to agree across methods	not reported	feature importance is read from attention weights; occlusion is native to this model, which is trained by masking genes and would supply a second method at no architectural cost
The reduced structure is reported and scoped to one task and dataset	satisfied	eight single-network variants are reported per task and cohort; one carries 359 interactions against 19,789 and matches the full model on several tasks
Checks of predictive performance		
Utility is shown on more than one task	satisfied	masked alteration recovery, cancer type, lymphocyte status, immunotherapy in four cohorts, chemotherapy in two, metastasis, and tissue of origin
Accuracy is compared against a strong single-task baseline	satisfied	logistic regression, random forest, and a matched graph network trained without pretraining and the clinical biomarkers in current use
Frozen embeddings place biological and technical replicates close together	satisfied	tumors of one type from MSK-IMPACT and The Cancer Genome Atlas (TCGA) share latent-space regions; prediction held on a cohort sequenced with a different panel
Frozen embeddings use the dimensionality of the space efficiently	not reported	the transfer embedding has 180 dimensions and the final layer 10; neither was assessed for effective rank
Accuracy is reported for every regime interpreted	satisfied	reported per cohort with confidence intervals for every task whose biomarkers are interpreted
Checks of biological validity		
A prediction is tested outside the training data, and the kind of test is stated	satisfied	immunotherapy response was tested in three independent cohorts and chemotherapy response in one, with metastasis and tissue of origin tested in further held-out cohorts; one immunotherapy cohort was sequenced on a different gene panel from the pretraining data; every test is computational

Eight requirements are satisfied, one is partially reported, and three are not reported. The satisfied requirements cover predictive performance, validation, and the removal of each prior. The checks that remain act on the trained model, and a report of them would provide additional context to users of the released backbone.

that was not built for inspection. Reporting standards for predictive models already exist,[40] and they cover code, data, and split release. We would extend them. Authors should make explicit their assessment and inclusion of interpretable priors and design elements at every stage of modeling. Reviewers and editors should use the included checklist to evaluate new models on interpretability. Which questions bear on a model depends on its design and its claims, and some will not apply. The requirement we would set is that the answers be reported, in the way journals now ask for data availability.

One open direction is to describe how a visible unit behaves across biological contexts. In current practice, construction fixes what a unit stands for, so a unit that maps to a pathway maps to that pathway in every tissue. Its contribution to a prediction can differ between tissues, however. Methods that report this contextual variation would show a user where an explanation transfers and where it does not.

Two further directions are open. The first is incorporating new knowledge. Curated resources are revised continually, and a prior built into a released backbone stays as it was on the day of training. The field needs ways to revise an encoded prior without retraining the backbone, and soft priors that report where a model departs from its resource, so that what the model learns can return to that resource. The second direction is to develop methods that measure how far a new input lies from the data a backbone was trained on. A foundation model is adapted to assays, tissues, and diseases it did not see in training, and its explanations can be trusted only where its predictions are trustworthy also. A model can be trusted only on inputs that resemble its training data, so a user needs a way to measure that resemblance for each new input.

When building models, molecular biology can rely on more than just larger scale data because it already knows a great deal. Our view is that the field is optimizing the wrong quantity. A predictive accuracy benchmark rewards a model for the right answer but does not ask about the right explanation. This is not a fair trade when the next step is a synthesis, a screen, or a therapeutic action in a patient. A model that cannot be interrogated cannot be corrected, and a shared backbone that cannot be corrected passes its errors to every group that adapts it. The biological priors and design elements that make a foundation model interpretable exist, and the work is for domain experts to adapt them to their use cases and implement them.

We would therefore set a different bar for the next generation of foundation models. A model should earn its place by naming the components behind a prediction, by surviving the tests those names imply, and by producing at least one hypothesis that someone carries to the bench. The foundation models that impact biology and medicine will be the ones built to that bar, not the ones that top an accuracy leaderboard.

ACKNOWLEDGMENTS

We gratefully acknowledge funding support from the ARPA-H Dynamic Digital Tumors for Precision Oncology project (ADAPT OT 140D042590013) and the CRUK Strategic Fund (SEBSTF-2025/100007). This research was funded in part by the US government. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US government.

DECLARATION OF INTERESTS

T.I. is a co-founder, advisor, and holder of equity for Data4Cure and Serinus Biosciences, and he is an advisor and shareholder for Ideaya BioSciences and Eikon Therapeutics. The terms of these arrangements have been reviewed and approved by UC San Diego in accordance with its competing interest policies.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Claude (Anthropic) and ChatGPT (OpenAI) to refine sections of human-authored text, assemble display items, and verify reference details. Afterwards, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

2. Blampey, Q., Benkirane, H., Bercovici, N., Mulder, K., Gessain, G., Ginhoux, F., André, F., and Cournède, P.-H. (2025). Novae: a graph-based foundation model for spatial transcriptomics data. Nat. Methods *22*, 2539–2550. https://doi.org/10.1038/s41592-025-02899-6.

3. Jaume, G., Oldenburg, L., Vaidya, A., Chen, R.J., Williamson, D.F.K., Peeters, T., Song, A.H., and Mahmood, F. (2024a). Transcriptomics-guided slide representation learning in computational pathology. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (Institute of Electrical and Electronics Engineers), pp. 9632–9644. https://doi.org/10.1109/CVPR52733.2024.00920.

4. Milacic, M., Beavers, D., Conley, P., Gong, C., Gillespie, M., Griss, J., Haw, R., Jassal, B., Matthews, L., May, B., et al. (2024). The Reactome Pathway Knowledgebase 2024. Nucleic Acids Res. *52*, D672–D678. https://doi.org/10.1093/nar/gkad1025.

5. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

6. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F.K., Liang, P.P., and Mahmood, F. (2024). Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (Institute of Electrical and Electronics Engineers), pp. 11579–11590. https://doi.org/10.1109/CVPR52733.2024.01100.

7. Kunes, R.Z., Walle, T., Land, M., Nawy, T., and Pe'er, D. (2024). Supervised discovery of interpretable gene programs from single-cell data. Nat. Biotechnol. *42*, 1084–1095. https://doi.org/10.1038/s41587-023-01940-3.

8. Kong, J., Lee, I., Boecher, D., Singhal, A., Kelly, M.R., Moon, J., Ahn, C.H., Ock, C.-Y., Pratt, D., Kumar, T., et al. (2026). A foundation model of cancer genotype enables precise predictions of therapeutic response. Published online May 26, 2026. Cancer Discov. https://doi.org/10.1158/2159-8290.CD-25-1735.

9. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

10. Wang, J., Tan, C., Gao, Z., Shao, S., Liu, S., and Li, S.Z. (2026). Illuminating cell states by a comprehensive and interpretable single cell foundation model. Nat. Commun. *17*, 4037. https://doi.org/10.1038/s41467-026-70071-5.

11. Fradkin, P., Shi, R.I., Dalal, T., Isaev, K., Frey, B.J., Lee, L.J., Morris, Q., and Wang, B. (2026). Orthrus: toward evolutionary and functional RNA foundation models. Nat. Methods *23*, 935–945. https://doi.org/10.1038/s41592-026-03064-3.

12. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. (2024). scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat. Methods *21*, 1470–1480. https://doi.org/10.1038/s41592-024-02201-0.

13. Yu, Z., Yang, D., Chen, Q., Zhang, Y., Li, Z., Wang, Y., Wang, C., and Zhang, Y. (2026). ChromBERT: a foundation model for learning interpretable representations for context-specific transcriptional regulatory networks. Cell Genom. *6*, 101130. https://doi.org/10.1016/j.xgen.2025.101130.

14. Schaefer, M., Peneder, P., Malzl, D., Lombardo, S.D., Peycheva, M., Burton, J., Hakobyan, A., Sharma, V., Krausgruber, T., Sin, C., et al. (2025). Multimodal learning enables chat-based exploration of single-cell data. Published online November 11, 2025. Nat. Biotechnol. https://doi.org/10.1038/s41587-025-02857-9.

15. Ismail, A.A., Oikarinen, T., Wang, A., Adebayo, J., Stanton, S.D., Corrada Bravo, H., Cho, K., and Frey, N.C. (2025). Concept bottleneck language models for protein design. https://doi.org/10.48550/arXiv.2411.06090

16. Kovács, D.P., Moore, J.H., Browning, N.J., Batatia, I., Horton, J.T., Pu, Y., Kapil, V., Witt, W.C., Magdău, I.-B., Cole, D.J., et al. (2025). MACE-OFF: short-range transferable machine learning force fields for organic molecules. J. Am. Chem. Soc. *147*, 17598–17611. https://doi.org/10.1021/jacs.4c07099.

17. Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. Nat. Mach. Intell. *4*, 1256–1264. https://doi.org/10.1038/s42256-022-00580-7.

18. Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. (2024). SaProt: protein language modeling with structure-aware vocabulary. https://openreview.net/forum?id=6MRm3G4NiU.

19. Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H., Brydon, E.M., Zeng, Z., Liu, X.S., et al. (2023). Transfer learning enables predictions in network biology. Nature *618*, 616–624. https://doi.org/10.1038/s41586-023-06139-9.

20. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.

21. Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics *224*. iyad031. https://doi.org/10.1093/genetics/iyad031.

22. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646. https://doi.org/10.1093/nar/gkac1000.

23. Ma, J., Yu, M.K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. Nat. Methods *15*, 290–298. https://doi.org/10.1038/nmeth.4627.

24. Elmarakeby, H.A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., Al-Dubayan, S.H., Salari, K., Kregel, S., Richter, C., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. Nature *598*, 348–352. https://doi.org/10.1038/s41586-021-03922-4.

25. Fortelny, N., and Bock, C. (2020). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biol. *21*, 190. https://doi.org/10.1186/s13059-020-02100-5.

26. van Hilten, A., Kushner, S.A., Kayser, M., Ikram, M.A., Adams, H.H.H., Klaver, C.C.W., Niessen, W.J., and Roshchupkin, G.V. (2021). GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. Commun. Biol. *4*, 1094. https://doi.org/10.1038/s42003-021-02622-z.

27. Park, S., Silva, E., Singhal, A., Kelly, M.R., Licon, K., Panagiotou, I., Fogg, C., Fong, S., Lee, J.J.Y., Zhao, X., et al. (2024). A deep learning model of tumor cell architecture elucidates response and resistance to CDK4/6 inhibitors. Nat. Cancer *5*, 996–1009. https://doi.org/10.1038/s43018-024-00740-1.

28. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500. https://doi.org/10.1038/s41586-024-07487-w.

29. Schiff, Y., Kao, C.-H., Gokaslan, A., Dao, T., Gu, A., and Kuleshov, V. (2024). Caduceus: bi-directional equivariant long-range DNA sequence modeling. In Proceedings of the Forty-first International Conference on Machine Learning (ICML 2024), *235* (PMLR), pp. 43632–43648.

30. Kedzierska, K.Z., Crawford, L., Amini, A.P., and Lu, A.X. (2025). Zero-shot evaluation reveals limitations of single-cell foundation models. Genome Biol. *26*, 101. https://doi.org/10.1186/s13059-025-03574-x.

31. Maddipatla, A., Sellam Bojan, N., Bojan, M., Masalitin, V., Vedula, S., Schanda, P., Marx, A., and Bronstein, A.M. (2026). Experiment-guided AlphaFold3 resolves measurement-consistent protein ensembles. Published online June 29, 2026. Nat. Biotechnol. https://doi.org/10.1038/s41587-026-03166-5.

32. Svensson, V., Gayoso, A., Yosef, N., and Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. Bioinformatics *36*, 3418–3421. https://doi.org/10.1093/bioinformatics/btaa169.

33. Yuan, B., Shen, C., Luna, A., Korkut, A., Marks, D.S., Ingraham, J., and Sander, C. (2021). CellBox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. Cell Syst. *12*, 128–140.e4. https://doi.org/10.1016/j.cels.2020.11.013.

34. Novakovsky, G., Dexter, N., Libbrecht, M.W., Wasserman, W.W., and Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat. Rev. Genet. *24*, 125–137. https://doi.org/10.1038/s41576-022-00532-2.

35. Chen, V., Yang, M., Cui, W., Kim, J.S., Talwalkar, A., and Ma, J. (2024). Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments. Nat. Methods *21*, 1454–1461. https://doi.org/10.1038/s41592-024-02359-7.

36. Gujral, O., Bafna, M., Alm, E., and Berger, B. (2025). Sparse autoencoders uncover biologically interpretable features in protein language model representations. Proc. Natl. Acad. Sci. USA *122*, e2506316122. https://doi.org/10.1073/pnas.2506316122.

37. Ganeeva, V., Khrabrov, K., Kadurin, A., and Tutubalina, E. (2025). Measuring chemical LLM robustness to molecular representations: a SMILES variation-based framework. J. Cheminform. *17*, 164. https://doi.org/10.1186/s13321-025-01079-0.

38. Garrido, Q., Balestriero, R., Najman, L., and LeCun, Y. (2023). RankMe: assessing the downstream performance of pretrained self-supervised representations by their rank. In Proceedings of the Fortieth International Conference on Machine Learning (ICML 2023), *202* (PMLR), pp. 10929–10974.

39. Ahlmann-Eltze, C., Huber, W., and Anders, S. (2025). Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nat. Methods *22*, 1657–1661. https://doi.org/10.1038/s41592-025-02772-6.

40. Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J.B., van Smeden, M., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ *385*, e078378. https://doi.org/10.1136/bmj-2023-078378.