Evaluating Biology Foundation Models: Leakage and Validation

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · biology foundation models · genomic language models · protein language models · data leakage · benchmark design · single-cell foundation models · CASP · train test contamination · AI evaluation methodology

Executive Summary

Evaluating a biology foundation model, whether it processes DNA sequences, protein sequences, single-cell transcriptomes, small molecules, or biomedical text, requires distinguishing four increasingly rigorous forms of evidence: random held-out computational splits, leakage-controlled splits, independent or prospective blind assessment, and physical wet-lab confirmation. At the 2020 **Critical Assessment of Structure Prediction (CASP14)** round, **AlphaFold2** achieved a median backbone accuracy of 0.96 angstrom root mean square deviation on a blind, independent test set against which nearly 100 research groups submitted more than 67,000 models (nature.com) (predictioncenter.org). That result has held up because CASP is blind and prospective, not because it was a large leaderboard.

The same rigor does not automatically apply to genomic, single-cell, or general-purpose language model benchmarks. Random splitting of biological sequence databases "has been shown to produce dubious assessments of generalization" because of similarity between training and test samples (pmc.ncbi.nlm.nih.gov), a

magnitude of inflation consistent with a broader 2023 finding that leakage measurably affected 294 published papers across 17 scientific fields (hbiostat.org). Efficiency gains are real even where accuracy gains are contested: DNABERT-2 matched prior state-of-the-art genomic benchmark results using 21 times fewer parameters and roughly 92 times less GPU pretraining time (arxiv.org).

On general biomedical language models, zero-shot GPT-4 reached 86.65 to 86.7 percent on official USMLE self-assessment exams in 2023, against 53.61 to 58.78 percent for GPT-3.5 (arxiv.org). At the top of the evaluation ladder, a small number of protein design cases have cleared physical confirmation entirely: RFdiffusion's designed binder structures were confirmed by cryogenic electron microscopy to be "nearly identical to the design model" (nature.com), an example of a computational biology prediction verified by orthogonal experimental cryo-EM. Readers evaluating a biology foundation model in 2026 should therefore treat a leaderboard score as a starting hypothesis, check its splitting methodology and baseline comparisons, and reserve full confidence for claims that have cleared an independent, prospective, or experimental test outside the model's own training distribution.

Introduction and Background

Biology foundation models, large neural networks pretrained on genomic sequences, protein sequences, single-cell transcriptomes, small molecules, or biomedical text, are typically introduced to the field with a leaderboard score on a held-out test split. That number is necessary but, on its own, insufficient: a growing body of peer-reviewed work published between 2023 and 2026 shows that held-out accuracy on a randomly split biological dataset can systematically overstate what a model has actually learned, because biological sequences and cells are rarely as independent and identically distributed as the images or text corpora that inspired the pretrain-then-fine-tune recipe. A promoter sequence on one chromosome can closely resemble a promoter sequence in the test split; a single-cell transcriptome from one patient can resemble one from another; a protein family in a benchmark's test set may share a common ancestor with proteins the model saw during pretraining. Each of these forms of resemblance, generally termed **data leakage** or **train-test contamination**, can inflate an apparent benchmark win without any corresponding gain in biological insight.

This report is a companion to IntuitionLabs' existing overview of large language model (LLM) benchmarks in life sciences, which surveys biomedical question-answering, drug-discovery, and genomics leaderboards broadly (intuitionlabs.ai). Rather than re-covering that ground, this article scopes narrowly to the **methodology** of evaluation itself: how held-out computational benchmarks are constructed for DNA, protein, single-cell, and molecule-generation foundation models; where and how data leakage enters those benchmarks; and what separates a computational leaderboard entry from independent, prospective, or experimental (wet-lab) proof that a model's predictions hold up outside its training distribution. As of September 2026, no single benchmark score, however large the model or however many parameters it contains, has been shown to reliably predict success in a real drug-discovery program, and the strongest available evidence for that gap comes from published data, not opinion. The sections that follow walk through the evaluation ladder from computational splits to wet-lab confirmation, the specific and well-documented leakage failure modes in genomic and protein datasets, model-class-by-model-class benchmark practice, and the growing set of studies in which simple, non-foundation-model baselines have matched or beaten purpose-built biology foundation models.

The Evaluation Ladder: From Held-Out Splits to Wet-Lab Proof

Evaluation evidence for a biology foundation model falls into a rough hierarchy of increasing rigor and decreasing convenience. At the base sits the **random held-out split**: a dataset is partitioned into training and test subsets, typically by shuffling examples, and the model is scored on the portion it did not train on. This is the fastest and cheapest form of evaluation, and it is also the most exposed to leakage, because random shuffling does not account for similarity between biological sequences or samples. A 2026 Nature Methods review of biomedical foundation model benchmarking frames the underlying tension directly: model parameters "are supposed to capture the patterns underlying the data" (nature.com), yet a standard benchmark task cannot easily distinguish a model that has captured a generalizable biological pattern from one that has memorized near-duplicate training examples.

The next rung is the **leakage-controlled split**, in which sequences or samples are clustered by similarity (typically sequence homology for DNA and protein, or patient/batch identity for single-cell data) before partitioning, so that no test example closely resembles a training example. Tools built for this purpose include **CD-HIT**, **uCLUST**, **HHblits**, and **MMseqs**, all cited as standard sequence-clustering methods used to reduce train-test similarity before splitting biological databases (pmc.ncbi.nlm.nih.gov). Above that sits **independent or prospective evaluation**, in which a model is scored on data that did not exist, or was not disclosed, at the time predictions were made. A Nature Machine Intelligence comment on drug-discovery benchmarking states plainly that "the gold standard for unbiased evaluation is a blind, prospective benchmark, in which different methods are evaluated on a newly generated test set that will only be disclosed after the results have been announced" (nature.com). The **Critical Assessment of Structure Prediction (CASP)**, discussed in detail below, is the field's longest-running example of this design. At the top of the ladder sits **experimental, wet-lab validation**: synthesizing a designed protein, expressing it, and measuring whether it folds or functions as predicted, independent of any computational metric. Each rung answers a different question, and, as the sections below document, models frequently place differently depending on which rung is used to judge them.

Data Leakage and Contamination in Biological Training Data

Data leakage is not a hypothetical concern confined to biology. A widely cited 2023 analysis by Kapoor and Narayanan, published in the peer-reviewed journal Patterns, found that leakage "affects 294 papers across 17 scientific fields" and, in a re-analysis of a widely cited civil-war-prediction study, showed that correcting for leakage erased the apparent advantage of complex machine-learning models over decades-old regression methods (hbiostat.org). That general warning has a specific and well-documented analogue in genomics. A 2024 paper introducing the **SpanSeq** partitioning method notes that random splitting of biological sequence databases, "although standard, has been shown to produce dubious assessments of generalization due to the existing similarity between samples in the databases used" (pmc.ncbi.nlm.nih.gov). The same paper reproduced training of a protein subcellular-localization model and an RNA structure model under similarity-aware splitting, "not only confirming the consequences of randomly splitting databases on the model assessment, but expanding those repercussions to the model development" (pmc.ncbi.nlm.nih.gov), meaning leakage distorts not just the reported score but which architectures and hyperparameters researchers select in the first place.

A 2026 preprint from the University of British Columbia's de Boer lab formalizes this problem for genomic sequence models as **homology-based data leakage**, defined as the phenomenon by which test sequences resembling training sequences produce inflated performance estimates (biorxiv.org). Critically, the study found that the field's

default mitigation, splitting training and test sets by chromosome, does not solve the problem: chromosome-based splits retained sequence pairs with a maximum Smith-Waterman alignment score above 175 to 179, whereas the authors' purpose-built **hashFrag** tool reduced that figure to below 70 (biorxiv.org), with hashFrag itself detecting homology at 99.99 percent recall and roughly 0.17 percent false-positive rate (biorxiv.org). The same preprint documents a related and less intuitive leakage mode: many high-confidence, genome-wide association study (GWAS) variant sites have genomic "doppelgangers," near-identical sequence stretches elsewhere in the genome, and the authors show that a 41 base-pair window around a variant is specific enough (roughly 1 in 5 times 10 to the 24th power possible sequences) to indicate true shared ancestry rather than coincidence (biorxiv.org), meaning evaluation of variant-effect models must account for these hidden duplicates or risk crediting a model for memorizing a variant it saw during training under a different genomic coordinate.

Evaluating Genomic and Protein Foundation Models

Genomic (DNA) language models are typically benchmarked on curated task suites rather than a single leaderboard. **Genomic Benchmarks**, published in BMC Genomic Data in May 2023, is one such curated collection; at publication it contained "nine datasets that focus on regulatory elements" (doi.org). **BEND** (Benchmark for DNA language models), accepted to the International Conference on Learning Representations (ICLR) 2024, introduced "a collection" of biologically meaningful downstream tasks on the human genome (arxiv.org) and found that DNA-model "embeddings from current DNA LMs can approach performance of expert methods on some tasks, but only capture limited information about long-range features" (arxiv.org), a limitation directly relevant to any application requiring long-range regulatory context. **DNABERT-2**, introduced in a 2023 preprint, defined the **Genome Understanding Evaluation (GUE)** benchmark, described in its abstract as combining "36 distinct datasets across 9 tasks, with input lengths ranging from 70 to 10000" base pairs (arxiv.org), and the model reportedly matched prior state-of-the-art results with "21 times fewer parameters and approximately 92 times less GPU time" (arxiv.org). **HyenaDNA**, a 2023 long-context genomic model, extended single-nucleotide-resolution context to 1 million tokens, described as "an up to 500x increase over previous dense attention-based models" (arxiv.org), and reached the top score on "12 of 18 datasets" from the **Nucleotide Transformer** benchmark suite while using orders of magnitude fewer parameters (arxiv.org). That underlying Nucleotide Transformer suite spans promoter, enhancer, splice-site, and histone-modification prediction tasks across human, mouse, and multispecies data (biorxiv.org), on which the original Nucleotide Transformer's learned representations "match or outperform specialized methods on 11 of 18 prediction tasks, and up to 15 after fine-tuning" (biorxiv.org).

Protein foundation models have a substantially older and more rigorous evaluation tradition, anchored by **CASP**, in which participants submit models for structures whose experimental structures are not yet public (predictioncenter.org), using structures that "have not been deposited in the PDB [Protein Data Bank] or publicly disclosed so that it is a blind test" (nature.com), with results published in a dedicated special issue of the peer-reviewed journal Proteins after each round (predictioncenter.org). CASP14, held in 2020, drew "nearly 100 groups from around the world" submitting "more than 67,000 models on 90 modeling targets" (predictioncenter.org), and it was on this independent, blind test set that AlphaFold2 achieved "a median backbone accuracy of 0.96 angstrom" root mean square deviation (RMSD) across an 87-domain evaluation subset compared with the top 15 of 146 competing entries (nature.com) (nature.com). CASP organizes targets into **Template-Based Modeling (TBM)** categories, where a related protein structure already exists, and **Free Modeling (FM)** categories, covering novel folds without a detectable structural template and therefore testing ab initio, or de novo, prediction capability (predictioncenter.org). By CASP16 in 2024, a leading MULTICOM4-based predictor achieved a Template Modeling

(TM) score above 0.9, indicating high structural accuracy, for 73.8 percent of 84 evaluated domains, and a correct overall fold (TM-score above 0.5) for 97.6 percent of domains (pmc.ncbi.nlm.nih.gov), ranking among the top of 120 participating predictors (pmc.ncbi.nlm.nih.gov). Complementing CASP's biennial cadence, **CAMEO** (Continuous Automated Model EvaluatiOn) provides "weekly, automated benchmarking of structure prediction servers, complementing the biennial" CASP rounds (ncbi.nlm.nih.gov), and a 2026 comparison of AlphaFold2, AlphaFold3, and ESMFold constructed a leakage-controlled test set by "excluding those entries with close homologs in the structures released prior to 2022" (ncbi.nlm.nih.gov), a direct, model-agnostic application of the leakage-control principle described above to structure-prediction benchmarking itself.

A parallel, function-oriented blind assessment exists in **CAFA** (Critical Assessment of Functional Annotation). CAFA2 evaluated "126 methods from 56 research groups for their ability to predict biological functions" (ncbi.nlm.nih.gov), and CAFA3 went a step further than any structure-prediction benchmark discussed so far by pairing computational predictions with laboratory confirmation, "resulting in new functional annotations for more than 1000 genes" (ncbi.nlm.nih.gov), making CAFA3 one of the few large-scale benchmarks in this survey that closes the loop between computational prediction and wet-lab confirmation as part of the benchmark itself rather than as a separate downstream step. **ProteinGym**, a large-scale protein-fitness benchmark, standardizes evaluation across "over 250 standardized deep mutational scanning assays, spanning millions of mutated sequences" (ncbi.nlm.nih.gov), unifying results for "over 70 high-performing models from various subfields" on a public leaderboard (ncbi.nlm.nih.gov).

On efficiency rather than leakage, a benchmark of gene-fusion breakpoint classification found the Nucleotide Transformer needed only "approximately 2,600 samples to reach 95% of its peak performance" against more than 14,000 for a dedicated deep-learning baseline (ncbi.nlm.nih.gov), the kind of sample-efficiency claim that a leakage-controlled benchmark can support with more confidence than a randomly split one.

Beyond structure prediction, protein-specific benchmarks address other evaluation needs. **FLIP** (Fitness Landscape Inference for Proteins) targets protein-engineering fitness rather than structure or function classification, spanning "adeno-associated virus stability for gene therapy, protein domain B1 stability and immunoglobulin binding, and thermostability from multiple protein families" (biorxiv.org). **ProteinBench**, a 2024 to 2025 holistic evaluation framework, scores generative protein models along "quality, novelty, diversity, and robustness" (arxiv.org) and explicitly guards against leakage by restricting one structural test set to "82 proteins whose PDB entries were deposited after May 1, 2019 and are not part of the training or validation set" (arxiv.org) for any model, a direct methodological response to the leakage concerns documented above. **PFMBench**, a mid-2025 benchmark from BioMap and Westlake University researchers, evaluates protein foundation models "across 38 tasks spanning 8 key areas of protein science" (arxiv.org).

Evaluating Single-Cell Foundation Models and General Biomedical Language Models

Single-cell foundation models learn representations of gene expression across millions of individual cells. **scGPT**, published in Nature Methods in 2024, is a generative pretrained transformer trained "across a repository of over 33 million cells" (nature.com). **Geneformer**, published in Nature in 2023, is pretrained on "about 30 million single-cell transcriptomes" (nature.com). **scFoundation**, published in Nature Methods in 2024, is a 100-million-parameter model covering roughly 20,000 genes and pretrained "on over 50 million human single-cell

transcriptomic profiles" (nature.com). The **Universal Cell Embedding (UCE)** model was used to build an atlas embedding 36 million cells across "more than 1,000 uniquely named cell types, from hundreds of experiments, dozens of tissues and eight species," in work published in Nature in July 2026 (nature.com).

Independent re-evaluation of this model class has produced some of the field's most consequential findings. A separate NeurIPS 2024 workshop benchmark comparing scGPT, Geneformer, CellPLM, and UCE against classical baselines found that "scVI and PCA [principal component analysis] [are] far better suited models for understanding biological perturbations in comparison to existing foundation models" (arxiv.org). An earlier model, **scBERT**, was originally described in Nature Machine Intelligence in 2022 as validated by "extensive and rigorous benchmark studies" showing "superior performance" on cell-type annotation (nature.com). A 2026 harmonised benchmark evaluating six spatial and single-cell foundation models, including Nicheformer, CellPLM, and scGPT-spatial, similarly reported that "no model dominated across tasks, and rankings shifted with modality, preprocessing, tokenisation, biological prior, domain shift and metric choice" (arxiv.org).

A separate, independently authored preprint titled "A Deep Dive into Single-Cell RNA Sequencing Foundation Models" benchmarked scBERT and scGPT specifically against a non-foundation baseline, "L1-regularized logistic regression, including in the few-shot setting" (biorxiv.org), concluding that the field must prioritize "rigorously testing foundation models against well established baselines" before crediting pretraining with genuine gains (biorxiv.org). On the integration-benchmark side specifically, the widely used **scIB** (single-cell integration benchmark) toolkit documents that its underlying study evaluated "16 methods... with 4 combinations of preprocessing steps leading to 68 methods combinations" across dozens of batches (github.com), illustrating that rigorous single-cell benchmarking predates, and does not depend on, the foundation-model paradigm at all.

General-purpose LLMs and molecule-generation benchmarks occupy an adjacent evaluation lane. **Therapeutics Data Commons (TDC)**, introduced in a 2021 paper, "includes 66 AI-ready datasets spread across 22 learning tasks and spanning the discovery and development of safe and effective medicines" (arxiv.org), organizing tasks into benchmark groups on its public leaderboard (tdcommons.ai). **MoleculeNet**, an earlier and still widely used molecular-property benchmark, found that for quantum-mechanical and biophysical datasets, "the use of physics-aware featurizations can be more important than choice of particular learning algorithm" (arxiv.org), a caution against treating architecture choice as the dominant driver of benchmark performance. **GuacaMol** proposed "an evaluation framework... based on a suite of standardized benchmarks" for de novo molecule generation (arxiv.org). On biomedical text, **PubMedQA**, introduced in 2019, reported that its best fine-tuned model of the time reached "68.1% accuracy, compared to single human performance of 78.0% accuracy and majority-baseline of 55.2% accuracy" (arxiv.org), establishing a human baseline against which later general-purpose LLMs would be compared; the dataset's official release page confirms the underlying corpus "has 1k expert labeled, 61.2k unlabeled and 211.3k artificially generated QA instances" built from PubMed abstracts (pubmedqa.github.io). A related, longer-running challenge series, **BioASQ**, "organizes challenges on biomedical semantic indexing and question answering (QA)" and was running its fourteenth annual edition as of 2026 (bioasq.org), providing a longitudinal, multi-year track record against which a single leaderboard snapshot from any one model generation can be compared. By 2023, zero-shot GPT-4 reached an average of "86.65% and 86.7% on the Self-Assessment and Sample Exam of the USMLE [United States Medical Licensing Examination] tests, respectively, compared to 53.61% and 58.78% for GPT-3.5" (arxiv.org), and later that year Google's Med-PaLM 2 reported "up to 86.5% on the MedQA dataset, improving upon Med-PaLM by over 19%" using an ensemble prompting strategy on top of the base model (arxiv.org).

Data Analysis and Evidence

The quantitative record assembled above supports two complementary observations: first, that benchmark suite scope and scale vary enormously by model class, and second, that within nearly every model class, at least one credible, independently authored study has shown a simple baseline matching or beating a purpose-built foundation model on at least one task. *Table 1* below summarizes representative benchmark suites by model class, standardizing on the scope, task count, and best-documented finding for each; every figure it contains is drawn from, and cited at, the sources discussed in the two preceding sections.

Model Class	Benchmark / Suite	Scope (as reported)	Headline Finding
Genomic (DNA)	GUE (via DNABERT-2, 2023)	36 datasets, 9 tasks, 70 to 10,000 base-pair inputs	Matched prior state of the art with 21 times fewer parameters
Genomic (DNA)	Nucleotide Transformer suite (2023)	18 classification tasks, 4 categories	HyenaDNA led on 12 of 18 with far fewer parameters
Protein structure	CASP16 (2024)	84 evaluated domains, 120 predictor groups	Top predictor: TM-score above 0.9 on 73.8% of domains
Protein function/design	ProteinBench (2024 to 2025) and PFMBench (2025)	ProteinBench: 4 evaluation dimensions; PFMBench: 38 tasks across 8 areas	Multi-dimensional evaluation (quality, novelty, diversity, robustness); no single winner across all axes
Single-cell	Genome Biology zero-shot audit (2025)	scGPT and Geneformer	Zero-shot evaluation
Single-cell (spatial)	Harmonised benchmark (2026)	6 models across scRNA-seq, spatial, and Perturb-seq data	No model dominated across tasks or metrics
Molecule generation	GuacaMol / TDC	TDC: 66 datasets, 22 tasks	Featurization choice can outweigh algorithm choice on quantum and biophysical tasks
General LLM (biomedical QA)	USMLE-style exams (2023)	GPT-4 zero-shot vs. GPT-3.5	GPT-4: 86.65% to 86.7%; GPT-3.5: 53.61% to 58.78%

Table 1 illustrates that scope alone (datasets, tasks, participating models) is not a proxy for evaluation rigor: several of the largest and most cited suites are exactly the ones where independent replication subsequently found simpler baselines competitive with, or superior to, the foundation model being showcased. Reading down the "headline finding" column, a pattern recurs across genomic, protein, and single-cell benchmarks alike: parameter count and pretraining scale correlate weakly, at best, with the metric that ultimately matters for a given task. This does not mean foundation models add no value; DNABERT-2 and HyenaDNA both demonstrate large efficiency gains at comparable accuracy, which is a genuine and reproducible result. It means that a benchmark table entry should be read together with the baseline it was compared against, and, where available, with any subsequent independent replication.

A second quantitative thread concerns how benchmark scale differs by evaluation philosophy rather than by model class alone. Blind, community-run assessments tend to be smaller and slower to assemble than self-reported leaderboards: CAFA2 drew 126 methods from 56 groups (ncbi.nlm.nih.gov) and CASP14 drew close to 100 groups

(predictioncenter.org), both organized around a single blind evaluation event, whereas self-assembled academic suites such as ProteinGym unify results from more than 70 models without requiring a shared, simultaneous submission window (ncbi.nlm.nih.gov). Neither structure is strictly superior: blind assessments better control for leakage and post-hoc tuning, while open leaderboards accumulate breadth faster. Readers evaluating a specific benchmark claim should therefore note which structure produced it before comparing scores across benchmarks of different types.

Case Studies and Real-World Examples

Independent computational benchmarking is itself several rungs below experimental confirmation, and a small number of well-documented cases illustrate what that final rung looks like in practice. **ESM-2**, a protein language model scaled to 15 billion parameters, was shown to produce an "atomic-resolution picture of protein structure" purely from sequence-based learned representations, without an explicit structure module (science.org); the resulting ESMFold pipeline was used to predict structures for "over 617 million metagenomic protein sequences, including over 225 million" at high confidence, published in Science in March 2023 (science.org). **ProGen**, a generative protein language model trained on 280 million sequences, was validated by synthesizing designed lysozyme enzymes in the laboratory; the resulting proteins "showed similar catalytic efficiencies as natural lysozymes, with sequence identity to natural proteins as low as 31.4%" (nature.com), meaning the designed enzymes functioned despite being distant from any sequence the model would have memorized verbatim. **RFdiffusion**, a diffusion-based protein design method from the Baker laboratory, was validated by "experimentally characterizing the structures and functions of hundreds of designed symmetric assemblies, metal-binding proteins and protein binders" (nature.com), with one designed binder's structure confirmed by cryogenic electron microscopy (cryo-EM) to be "nearly identical to the design model" (nature.com), an instance in which orthogonal experimental cryo-EM confirmed a generative model's computational output at atomic resolution.

On the small-molecule side, a 2024 Nature Machine Intelligence review of generative molecular design catalogs "the theoretical, computational and empirical challenges in deploying generative machine learning" toward real drug-discovery endpoints (nature.com) while pointing to a small set of studies where generated molecules were carried through to laboratory synthesis and testing; it identifies one such 2019 study as "one of the first studies to experimentally validate ML-generated molecules," a kinase-inhibitor program that highlighted the potential for accelerated early-stage discovery (nature.com). Taken together, these cases share a common structure: a computational benchmark or generative pipeline produced a candidate, and a physical measurement (enzymatic assay, cryo-EM structure, or synthesized-compound testing) either confirmed or refuted it. That two-step pattern, compute then confirm, is the practical form the evaluation ladder takes once a biology foundation model's output is intended for real-world use rather than leaderboard placement.

Table 2 below condenses these four cases side by side, pairing each model's computational claim with a reported computational application or experimental measurement; citations for every figure appear in the discussion above.

Model / Method	Computational Claim	Reported Application or Experimental Measurement
ESM-2 / ESMFold (2023)	Atomic-resolution structure emerges from a 15-billion-parameter sequence-only language model	Computational application: predicted structures for over 617 million metagenomic protein sequences, including over 225 million predicted with high confidence

Model / Method	Computational Claim	Reported Application or Experimental Measurement
ProGen (2023)	Generative language model designs novel lysozyme sequences	Synthesized enzymes matched natural catalytic efficiency at sequence identities as low as 31.4%
RFdiffusion (2023)	Diffusion model generates symmetric assemblies, metal-binding proteins, and protein binders	Cryo-EM structure of a designed binder nearly identical to the computational design model
Generative molecular design (2019 case, per 2024 review)	Generative model proposes small-molecule kinase-inhibitor candidates	Candidates synthesized and experimentally tested, among the first such generative-design programs to be validated wet-lab

Table 2 underscores that experimental confirmation, when it has occurred, has been concentrated in protein design rather than genomic, single-cell, or general-LLM applications, reflecting both the relative tractability of expressing and assaying a single designed protein and the field's comparative youth in generating experimentally testable hypotheses from DNA, single-cell, or text-based biology models.

Implications and Future Directions

Several practical implications follow for teams selecting or building on a biology foundation model. First, a leaderboard number should always be read alongside its splitting methodology: was the test set constructed by simple random shuffling, or by similarity-aware clustering using tools such as CD-HIT, MMseqs, or the newer hashFrag (biorxiv.org)? Second, a claimed foundation-model advantage should be checked against a strong, non-learned or simple baseline before it informs a purchasing or research decision; across the genomic, single-cell, and general-LLM sections above, credible independent studies repeatedly found simple baselines competitive with purpose-built foundation models. Third, for any application where a prediction will drive a real, physical decision (which sequence to synthesize, which variant to flag, which compound to order), computational benchmark performance should be treated as a screening signal rather than as proof; the CASP model of blind, prospective, community-run assessment remains the closest analogue in biology to a genuinely unbiased evaluation (nature.com), and it exists specifically because computational accuracy and real-world utility are not the same claim.

For organizations building internal evaluation pipelines rather than relying solely on published leaderboards, the practical challenge shifts from finding a benchmark to governing how model outputs, citations, and evaluation results are tracked over time as new model versions ship. IntuitionLabs, a life-sciences and AI consultancy founded in 2023 whose stated focus is "AI Acceleration for Life Sciences" delivered "one department at a time" (intuitionlabs.ai), works with pharmaceutical and biotechnology organizations on connecting AI systems to internal, authoritative sources "with identity, permissions, retrieval, citations, evaluation, and accountable operation" (intuitionlabs.ai), a governance-layer problem distinct from, but complementary to, choosing which published model or benchmark to trust in the first place. As more biology foundation models move from academic leaderboards toward operational use in drug discovery and diagnostics workflows, the gap this report has documented, between held-out computational accuracy and independent, prospective, or wet-lab confirmed performance, is likely to remain the central open question for anyone evaluating a candidate model, more so than any single new architecture or parameter count.

Frequently Asked Questions (FAQs)

What is the difference between a held-out benchmark and independent validation of a biology foundation model? A held-out benchmark scores a model on data drawn from the same distribution and often the same collection process as its training data, simply excluded from training; independent or prospective validation, exemplified by CASP's blind assessment, scores a model on data that did not exist or was not disclosed at prediction time, which controls for both data leakage and unconscious dataset selection bias (predictioncenter.org).

How does data leakage happen in genomic language models specifically? The most common mechanism is sequence homology: near-duplicate or highly similar sequences appear in both training and test partitions because a standard random or chromosome-based split does not account for cross-chromosome sequence similarity, which a 2026 study showed can leave alignment scores well above the threshold that indicates shared ancestry (biorxiv.org).

How should protein language models be evaluated? Structure-prediction claims should be checked against CASP or CAMEO's weekly blind assessments where possible (ncbi.nlm.nih.gov); function-prediction claims can be checked against CAFA's blind, community-run design; and fitness- or design-oriented claims should be checked against benchmarks that explicitly filter for post-training-cutoff structures, such as recent leakage-controlled structure-prediction comparisons (ncbi.nlm.nih.gov).

What benchmark design practices reduce train-test contamination in biological datasets? Clustering sequences by similarity before splitting, using tools such as CD-HIT, uCLUST, HHblits, or MMseqs (pmc.ncbi.nlm.nih.gov), and dating test sets to a cutoff strictly after the training data's collection date, are the two most widely documented mitigations.

Do single-cell foundation models outperform simpler statistical methods? Results are task- and dataset-dependent.

What counts as experimental validation of an AI biology model's predictions? A physical, laboratory measurement performed independently of the computational pipeline, such as expressing and assaying a designed protein, solving its structure by cryo-EM or X-ray crystallography, or synthesizing and testing a generated molecule, as distinct from any in-silico metric; CAFA3's pairing of computational function predictions with new laboratory-confirmed annotations for over 1,000 genes is a benchmark-level example of this standard (ncbi.nlm.nih.gov).

Conclusion

Evaluating a biology foundation model well requires moving past a single leaderboard number and asking how that number was produced. This report has traced a consistent pattern across DNA, protein, single-cell, and molecule-generation foundation models: held-out computational benchmarks are necessary starting points, but random or chromosome-based splitting routinely leaves measurable sequence or sample similarity between training and test sets, and multiple independently authored studies across every model class examined have found that simple, non-foundation-model baselines can match or exceed purpose-built architectures on at least some tasks. The strongest available counterexample, and the field's clearest existing template for rigorous assessment, remains the blind, prospective, community-run structure of CASP, together with the handful of documented cases in which a computational prediction was subsequently assessed by an experimental physical measurement. Readers

evaluating a new biology foundation model in 2026 and beyond should treat leaderboard placement as a starting hypothesis, examine the underlying splitting methodology and baseline comparisons before crediting an architectural claim, and reserve full confidence for predictions that have cleared an independent, prospective, or experimental test outside the model's own training distribution.
