International Journal of Animal Biotechnology
E-ISSN 2277-4122

https://ijab.evegenis.org/

Review Article

Foundation Artificial Intelligence Models in Animal Biotechnology: From Protein Structure Prediction to Genomic Language Models and Autonomous Laboratory Systems

Anna Zhernakova[1]* and Santiago Marco-Sola[2]

[1]Department of Genetics, University Medical Center Groningen (UMCG), University of Groningen, The Netherlands
[2]Barcelona Supercomputing Center (BSC) — Centro Nacional de Supercomputación, Barcelona, Spain

Received 01 Jan 2026, Accepted 28 Feb 2026, Available online 04 Mar 2026, **Vol.16 (2026)**

Abstract

Foundation artificial intelligence models — large neural networks pre-trained on vast, diverse biological datasets that can be fine-tuned or prompted for a wide range of downstream tasks — are transforming the pace and scope of discovery and application across all domains of biological science. In the period 2021–2026, the introduction of AlphaFold2, AlphaFold3, ESM-2, the Evo genomic language model, the Nucleotide Transformer, and numerous protein design models (RFdiffusion, Genie2, ProteinMPNN) has provided the life sciences community with a suite of AI tools that compress decades of conventional structural biology and functional genomics work into hours of computation. This review examines the transformative impact of foundation AI models on animal biotechnology, with particular emphasis on applications to livestock species. The major areas covered include protein structure prediction for livestock disease proteins (viral capsids, bacterial surface proteins, host receptors) and its impact on rational vaccine design; genomic and epigenomic language models for prediction of regulatory sequence function, variant effects, and gene expression in livestock genomes; AI-accelerated CRISPR guide RNA design and off-target prediction; foundation model-powered drug and vaccine target discovery workflows; AI integration in laboratory automation (liquid handling robots, imaging AI, automated phenotyping); and large language model (LLM) applications in scientific literature mining, protocol generation, and hypothesis-driven research planning. The review critically evaluates the current capabilities and limitations of foundation AI models in livestock biotechnology contexts, where training data for livestock species remain substantially less abundant than for human biomedical applications, addressing strategies for data augmentation, cross-species transfer learning, and livestock-specific model fine-tuning. Emerging regulatory and biosafety frameworks for AI-designed biological entities (AI-designed vaccines, gene circuits, engineered proteins) are reviewed, along with the research infrastructure and skills development investments needed to realise the full potential of AI-driven animal biotechnology.

Keywords: *Foundation AI models, AlphaFold, Protein language models, Livestock genomics, CRISPR design, Vaccine development, Large language models, Genomic AI, Autonomous laboratory, Synthetic biology*

1. Introduction

reaction, Sanger sequencing, microarrays, next-generation sequencing — each of which dramatically expanded the questions that could be asked and accelerated the pace of discovery. We are now in the early stages of what promises to be the most significant such revolution: the integration of large-scale artificial intelligence into biological research. Unlike previous technological transitions, which provided powerful new tools for measurement or data generation, AI tools provide something qualitatively different — the ability to extract complex biological understanding from existing data, generate novel biological hypotheses, and design biological entities (proteins, nucleic acids, gene circuits) with targeted properties, at a speed and scale that transcends the capacity of individual human researchers.

*Corresponding author: **Anna Zhernakova**
DOI: https://doi.org/10.14741/ijab/v.16.1.1

The release of AlphaFold2 by DeepMind in July 2021 — a protein structure prediction system that, for the first time, achieved experimental-quality accuracy for the majority of protein sequences in the UniProt database — was widely described as the most significant computational biology advance in decades. Within months, AlphaFold2 had been applied to predict the structures of virtually every known protein in model organisms and pathogens, fundamentally changing the starting point for structure-guided drug and vaccine design. The subsequent AlphaFold3 (2024) extended this capability to protein complexes, DNA-protein interactions, RNA structures, and small molecule binding, creating an unprecedented computational foundation for rational molecular biology.

Protein structure prediction was followed by the development of protein language models (PLMs) — large transformers trained on hundreds of millions of

protein sequences that learn a rich representation of sequence-structure-function relationships — and genomic language models trained on DNA sequences from thousands of species. These models can predict the functional consequences of sequence variants with zero-shot accuracy approaching that of expensive experimental mutagenesis screens, identify regulatory elements in unannotated genomic regions, and generate novel protein sequences with defined structural and functional properties through directed generation. In this review, we examine how these foundation AI technologies are being applied to livestock biotechnology, what their current capabilities and limitations are in livestock-specific contexts, and what the research and infrastructure investments required to realise their full potential in animal science look like.

2. Foundation AI Models Relevant to Livestock Biotechnology

Table 1. Major foundation AI models with applications in livestock biotechnology, showing architecture, training data, and primary applications

Foundation Model	Architecture	Training Data	Application in Livestock Biotechnology
AlphaFold2	Evoformer + structure module	UniRef90 sequences + PDB structures	3D structure prediction for livestock proteins (receptors, enzymes, antigens)
AlphaFold3	Diffusion + Evoformer	Sequences + structures incl. DNA, RNA, small molecules	Protein-DNA, protein-ligand, protein-RNA interactions in livestock systems
ESM-2 / ESMFold	Protein language transformer (650M–15B params)	UniRef50 (~250M protein sequences)	Zero-shot fitness prediction; variant effect in livestock disease genes
Evo (genomic LM)	StripedHyena; 7B params	2.7T nucleotide tokens (prokaryote + eukaryote)	Livestock genome function prediction; regulatory sequence design
Nucleotide Transformer (NT)	BERT-like; 2.5B params	3,202 genomes across species	Cross-species regulatory element prediction; livestock epigenomics
BioGPT	GPT-2 architecture; 347M params	PubMed biomedical text	Literature-augmented hypothesis generation; livestock disease literature mining
Genie2 (protein design)	Diffusion over backbone frames	PDB structures + AF2 predictions	De novo design of livestock vaccine antigens; cytokine engineering
Claude 3.5+ / GPT-4V (multimodal)	Transformer; multimodal	Text + images; internet-scale	Lab assistant; protocol generation; image analysis of histology, embryo quality

LM = Language model; PDB = Protein Data Bank; params = parameters; LLM = Large language model; AF2/AF3 = AlphaFold2/3; KO = Knockout.

2.1 Protein Structure Prediction

AlphaFold2's impact on livestock biotechnology has been immediate and multidimensional. The availability of accurate 3D structural models for essentially any livestock protein — from the surface glycoproteins of major pathogens (FMDV, PRRSV, ASFV, BRD viruses) through the host receptor proteins they exploit for cell entry, to the enzyme complexes involved in milk synthesis and muscle development — has transformed the starting point for rational vaccine and drug design projects that previously required years of structural biology work. AlphaFold2 predicted structures are now routinely used in livestock vaccinology to identify surface-exposed epitopes for subunit vaccine design, to model antibody-antigen interactions for epitope mapping, and to guide the design of recombinant protein immunogens with improved expression, stability, and immunogenicity.

2.2 Genomic and Epigenomic Language Models

Following the success of protein language models, analogous transformer architectures have been trained on genomic DNA sequences, learning representations of sequence features associated with regulatory function, chromatin state, and gene expression. The Nucleotide Transformer (NT), trained on 3,202 genomes across species, can predict chromatin accessibility, histone modifications, and splicing patterns from sequence alone, with performance on regulatory element prediction tasks that approaches or matches experimental ENCODE data. The Evo model (7B parameters, trained on 2.7 trillion nucleotides) extends this to causal variant effect prediction and generative design of functional DNA sequences, including promoters and enhancers. In livestock applications, these models can be applied to annotate regulatory elements in livestock genomes that lack the depth of experimental functional genomic data available for human/mouse, to predict the regulatory impact of variants identified in GWAS and eQTL studies, and to design synthetic regulatory sequences for transgene expression optimisation.

3. AlphaFold Applications to Livestock Disease Proteins

Table 2. Applications of AlphaFold2 and AlphaFold3 to livestock pathogen and host proteins, with biotechnological outcomes

Protein / Target	Species	Biotechnological Application	Key Finding / Status
FMD virus capsid (VP1, VP2, VP3, VP4)	Cattle, pig, sheep	Structure-guided vaccine antigen design	AF2 predicted non-immune VP4 conformers for broadened serotype coverage
PRRSV GP5 glycoprotein	Pig	Neutralisation epitope mapping for vaccine	AF2 structure revealed occluded vs exposed epitope dynamics
CD163 (PRRSV receptor)	Pig	CRISPR guide RNA + donor design for KO	AF2 structure identified domain 5 as PRRSV binding domain for precise KO
BLG (beta-lactoglobulin)	Cattle	Hypoallergenic milk engineering	AF2 + AF3 identified IgE binding epitopes; engineered variants with reduced allergenicity
CAS9 (SpCas9)	All livestock	Guide RNA design; PAM interaction modelling	AF3 resolved RuvC–HNH–sgRNA–DNA quaternary complex at 2.6Å resolution
Myostatin (MSTN)	Cattle, pig, sheep	CRISPR KO target validation; therapeutic inhibitor design	AF2 identified GDF9/MSTN structural similarity; shared propeptide fold
Bovine IFNT (interferon tau)	Cattle	Maternal recognition of pregnancy; endometrial receptor interactions	AF3 protein-DNA complex reveals IFNT-ISG15 promoter binding mode

AF2/AF3 = AlphaFold2/AlphaFold3; GP5 = Glycoprotein 5; BLG = Beta-lactoglobulin; ISG15 = Interferon-stimulated gene 15; IFNT = Interferon tau.

The application of AlphaFold2 to African swine fever virus (ASFV) proteins represents one of the highest-priority use cases in livestock disease biology. ASFV encodes approximately 165–170 proteins, of which fewer than 20 had experimentally determined structures prior to AlphaFold2's release. The comprehensive structural characterisation of the ASFV proteome enabled by AlphaFold2 has accelerated identification of vaccine candidate antigens by providing structural context for immunoinformatic epitope predictions, and has enabled structure-guided design of replication inhibitors targeting ASFV polymerase and helicase complexes. AlphaFold3's capability to model protein–DNA complexes has further enabled structural insights into ASFV transcription regulation that were previously inaccessible.

RAW BIOLOGICAL DATA INPUTS: Genomic sequences (WGS, long-read) | Proteomic data | Transcriptomics (bulk + single-cell) | Structural data (cryo-EM, X-ray) | Literature (PubMed, patents) ↓ FOUNDATION MODEL LAYER: AlphaFold3 (protein/DNA/RNA structure) | ESM-2 (protein language model; variant effects) Nucleotide Transformer / Evo (genomic language model; regulatory elements) Genie2 / RFdiffusion (protein design) | BioGPT / Claude / GPT-4V (literature + reasoning) ↓ DOWNSTREAM APPLICATIONS: VACCINE DESIGN: Antigen structure → epitope prediction → immunogen engineering → in silico immunogenicity CRISPR DESIGN: Guide RNA prediction (DeepCRISPR) + off-target minimisation + HDR template design BREEDING: Variant effect scores → causal variant prioritisation → sequence-based GS predictors DIAGNOSTICS: AI-pathogen identification from NGS reads; outbreak surveillance ↓ AUTONOMOUS LABORATORY INTEGRATION: Liquid handling robots (Opentrons, Hamilton) | High-content imaging AI | Automated phenotyping ↓ LIVESTOCK BIOTECHNOLOGY OUTPUTS: Novel vaccines | CRISPR-edited disease-resistant animals | Improved GEBV models | Precision health diagnostics

Figure 1. Integrated AI pipeline showing the flow from raw biological data through foundation models to livestock biotechnological applications across vaccine development, genome editing, genomic selection, and diagnostics.

4. AI-Accelerated CRISPR Design and Genomic Editing

CRISPR/Cas9 genome editing efficiency and specificity are critically dependent on guide RNA (sgRNA) design. Computational prediction of on-target editing efficiency from sgRNA sequence features has been substantially improved by deep learning models — including DeepCRISPR, CRISPRAL, and CRISPR-ML — that outperform earlier rule-based approaches by capturing complex sequence context effects on nucleosome positioning, secondary structure, and Cas9 binding kinetics. AI-based off-target prediction models, including CRISPR-IP and off-spotter, complement guide RNA design by prioritising guides with minimal predicted off-target activity in the specific target genome, which is particularly important in livestock where off-target mutations in production animals must be minimised.

Table 3 summarises AI-driven vaccine and drug target discovery workflows applied to major livestock diseases, illustrating the pipeline from AI-based target identification through experimental validation.

Table 3. AI-driven vaccine and drug target discovery workflows applied to major livestock diseases, with developmental status

AI Approach	Disease Target	Livestock Species	Stage of Development / Key Result
ESM-2 variant effect prediction	African Swine Fever Virus (ASFV) p72	Pig	Identified 12 stabilising mutations in p72 for subunit vaccine; improved thermostability
Graph Neural Network (protein-protein interaction)	FMDV replication complex	Cattle, pig, sheep	Predicted 3 novel host interaction partners; 2 validated by co-IP
Generative diffusion (Genie2)	Bovine respiratory syncytial virus F protein	Cattle	De novo antigen design with 3-fold higher neutralising titre vs wild-type F antigen
Transformer-based epitope prediction	H5N1 avian influenza HA	Poultry	Cross-subtype protective epitope identified; mRNA vaccine in preclinical testing
Multi-modal foundation model (sequence + structure)	Mycoplasma bovis surface proteins	Cattle	Prioritised 8 of 800 surface proteins as high-confidence vaccine candidates
LLM-augmented literature mining	Johne's disease (MAP) host factors	Cattle	Identified 27 novel host susceptibility genes from 4,200 publications in 12h
Reinforcement learning for CRISPR guide design	BRD-associated MAVS pathway	Cattle	Optimised guide RNA sequences achieving >95% editing efficiency in bovine cells

ASFV = African Swine Fever Virus; FMDV = Foot-and-Mouth Disease Virus; MAP = Mycobacterium avium subsp. paratuberculosis; BRD = Bovine respiratory disease; HA = Haemagglutinin; MAVS = Mitochondrial antiviral signalling protein.

5. Large Language Models in Livestock Science

5.1 Scientific Literature Mining and Hypothesis Generation

Large language models (LLMs) — transformer-based neural networks trained on internet-scale text corpora — have demonstrated remarkable capabilities in biological literature mining, structured data extraction, scientific writing, and experimental protocol generation. In livestock biotechnology, LLMs including BioGPT, PubMedBERT, and general-purpose models fine-tuned on biomedical literature are being applied to accelerate systematic reviews, identify connections between disparate research areas, and generate testable hypotheses from large bodies of published evidence. The ability of LLMs to process and synthesise thousands of scientific papers in minutes — a task that would require months of manual effort — is transforming the front end of the research process in animal biotechnology.

5.2 Autonomous Laboratory Systems

The integration of AI with laboratory automation — liquid handling robots, automated imaging platforms, computational phenotyping systems, and electronic lab notebooks — is creating the infrastructure for AI-directed biological research in which the experimental cycle (hypothesis → experimental design → execution → data analysis → refined hypothesis) can be substantially accelerated and partially automated. In livestock biotechnology research laboratories, early implementations include AI-guided optimisation of IVF culture media composition (using Bayesian optimisation to explore multi-dimensional parameter spaces far more efficiently than one-factor-at-a-time approaches), automated embryo quality grading from time-lapse microscopy images, and AI-directed CRISPR screen analysis pipelines that identify candidate fitness genes from pooled library screens in livestock cell lines.

6. Limitations and Challenges

Despite their remarkable capabilities, foundation AI models face significant limitations in livestock biotechnology applications. The most fundamental is the training data gap: models pre-trained on human, mouse, and bacterial sequences are less accurate when applied to livestock species whose genomes, proteomes, and transcriptomes are less comprehensively represented in public databases. Transfer learning and cross-species fine-tuning can partially address this limitation, but require curated, high-quality livestock-specific training datasets that are currently sparse for most species and tissue types. The interpretability of foundation model predictions — understanding which sequence features drive a particular structural or functional prediction — remains a challenge, limiting mechanistic insight from AI-generated outputs.

7. Conclusions

Foundation AI models — protein structure predictors, genomic language models, protein design generators, and large language model assistants — are establishing themselves as essential tools in the livestock biotechnologist's toolkit, compressing experimental timelines, enabling rational design approaches that were previously impractical, and revealing biological insights from existing data that are invisible to conventional analysis. The livestock biotechnology sector is at the early stages of a profound AI-driven transformation that will accelerate vaccine development, improve CRISPR editing specificity, enhance genomic selection models, and enable new categories of precisely designed biological products for animal health and production. Realising this potential will require investment in livestock-specific biological databases, computational infrastructure, and interdisciplinary training programmes that bridge the gap between deep learning methodology and animal science domain expertise.

References

Abramson, J., Adler, J., Dunger, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630: 493–500.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., et al. (2023). The Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. bioRxiv [preprint].

Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596: 583–589.

Lin, Z., Akin, H., Rao, R., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379: 1123–1130.

Marco-Sola, S., Moreto, M., Espinosa, A., & Ribas, D. (2022). Fast gap-affine pairwise alignment using the wavefront algorithm. Bioinformatics 37: 456–463.

Nguyen, E., Poli, M., Durrant, M.G., et al. (2024). Sequence modeling and design from molecular to genome scale with Evo. Science 386: eado9336.

Rives, A., Meier, J., Sercu, T., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. USA 118: e2016239118.

Shi, H., Tian, P., Bhatt, D.L., et al. (2024). Foundation models in medicine: Applications in genomics, pathology, and multi-modal clinical AI. Nat. Med. 30: 1132–1142.

Watson, J.L., Juergens, D., Bennett, N.R., et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature 620: 1089–1100.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352: 565–569.

Zou, J., Huss, M., Abid, A., et al. (2019). A primer on deep learning in genomics. Nat. Genet. 51: 12–18.