
# Conference Track Summary –

## Machine Learning for Protein Engineering

Brought to you by Cambridge Healthtech Institute, a division of Cambridge Innovation Institute

**250 1st Ave, Suite 300, Needham, MA 02494**
**T: 781.247.6262**

# Executive Overview

The PEGS Europe 2025 Machine Learning for Protein Engineering track demonstrated that artificial intelligence is moving beyond isolated prediction tools toward becoming an integrated component of biologics discovery. Across presentations, speakers emphasized that meaningful advances depend on combining machine learning with high-quality experimental data, iterative laboratory validation, and robust digital infrastructure. Rather than focusing solely on model accuracy, the discussions highlighted practical strategies for deploying AI within real-world discovery workflows, accelerating antibody engineering, improving developability assessment, and enabling more efficient design-build-test-learn cycles.

**Most Frequently Covered Issues**

**1. Integration of AI with experimental workflows**
Speakers consistently highlighted closed-loop design-build-test-learn strategies that combine computational prediction with continuous laboratory validation.

**2. High-quality data as the foundation for AI**
Large, standardized, experimentally validated datasets were repeatedly identified as the critical factor limiting model performance and broader generalization.

**3. Multi-objective optimization in protein engineering**
Rather than optimizing single properties, AI is increasingly used to balance affinity, stability, specificity, manufacturability, and other developability characteristics simultaneously.

**4. Operationalizing AI for routine research**
Successful implementation depends on scalable software platforms, standardized data infrastructure, model governance, and user-friendly tools that fit naturally into scientific workflows.

**5. Advances in protein language models and generative AI**
Presentations explored how transformer models, structural representations, and generative approaches are expanding capabilities in protein design, structure prediction, and antibody engineering.

**Recurring Takeaways**

- AI delivers the greatest value when tightly integrated with iterative experimental feedback rather than operating as a standalone prediction engine.
- Standardized, high-quality experimental data remain the most important prerequisite for developing reliable and broadly applicable machine learning models.
- Active learning strategies can significantly reduce laboratory workload while maintaining high-quality scientific decision making.
- Practical deployment requires equal attention to workflow engineering, infrastructure, uncertainty estimation, and scientist adoption alongside model development.
- The convergence of generative AI, protein language models, structural biology, and laboratory automation is enabling faster, more efficient discovery of next-generation biologics.

# Lab-in-the-Loop Application for Clinically Relevant Antigen Targets

Full Video Here

**Ji Won Park, PhD,** Principal ML Scientist, Prescient Design, Genentech

Machine learning creates greater value in antibody discovery when it operates as part of a continuous experimental cycle rather than a standalone prediction tool. A Lab-in-the-Loop approach combines generative models, laboratory testing, and iterative model refinement so each round of experimental data improves the quality of future designs while reducing unnecessary screening and accelerating optimisation. (00:05:39-00:09:10)

The workflow integrates generative sequence design, multi-property prediction, and active learning to evaluate millions of candidate antibodies before selecting a small group for laboratory validation. Rather than optimizing a single characteristic, the platform balances multiple therapeutic properties while incorporating prediction uncertainty, allowing researchers to explore both high-confidence candidates and promising unexplored sequence space. (00:09:10-00:15:56)

A central principle is treating antibody engineering as a multi-objective optimisation problem from the outset. Binding affinity, expression, specificity, stability, aggregation risk, and other developability attributes are considered together, helping identify liabilities earlier and producing candidates with a stronger overall balance of therapeutic properties. When experimental data are limited, diversity is intentionally maintained to maximise learning in subsequent design cycles. (00:11:03-00:17:30)

Applications across multiple clinically relevant antigen programmes demonstrated meaningful improvements in antibody affinity while requiring relatively modest experimental datasets to initiate optimisation. Structural analyses also indicated that sequence-based machine learning models captured biologically meaningful relationships without explicit structural inputs. The overall strategy illustrates how closely coupling computational design with iterative laboratory validation can improve sample efficiency, strengthen decision-making, and accelerate therapeutic antibody discovery. (00:17:30-00:27:49)

**Key Takeaways**

- Lab-in-the-Loop workflows continuously improve machine learning models through experimental feedback.
- Multi-property optimisation enables balanced selection across affinity, developability, and manufacturability metrics.
- Active learning uses prediction uncertainty to prioritise the most informative laboratory experiments.
- Integrating computational design with iterative validation improves sample efficiency and accelerates antibody discovery.

# Active Learning for Antibody Pairwise Epitope Binning

Full Video Here

**Akila I. Katuwawala, PhD,** Scientist II, Computational Biology, Adimab LLC

Experimental epitope binning provides critical information during antibody discovery by identifying which antibodies compete for similar binding sites, but generating complete pairwise competition datasets is expensive and resource intensive. The work presented an active learning strategy that predicts antibody competition using only a small subset of experimentally measured interactions. By combining targeted laboratory measurements with machine learning, the approach delivers actionable epitope diversity information while substantially reducing experimental effort. (00:00:49-00:08:46)

Early attempts to build a universal prediction model across many targets using sequence-derived features alone produced limited accuracy, highlighting the complexity of antibody competition. Instead, the model was trained separately for each antibody panel using a carefully selected set of experimental measurements. Randomly sampling a small number of pairwise interactions proved more effective than clustering-based selection methods, providing sufficient information for accurate predictions with only a fraction of the complete dataset. (00:05:46-00:19:35)

Gradient-boosted decision trees combined with fixed-length sequence representations consistently outperformed more complex neural network approaches in this low-data setting. The optimal workflow typically required measurements representing only 5% to 10% of the full competition matrix while maintaining strong predictive performance. Validation across multiple datasets and a prospective antibody discovery campaign demonstrated that the predicted competition maps closely matched experimentally generated reference datasets. (00:12:22-00:23:11)

The workflow has been fully integrated into antibody discovery operations, allowing scientists to request predictions as part of routine campaigns. Experimental measurements are automatically incorporated to train project-specific models before generating complete competition maps for downstream decision making. Compared with generating full experimental datasets, the approach substantially reduced antibody and antigen consumption, instrument time, and assay costs while maintaining approximately 90% predictive accuracy. Ongoing work is evaluating large language model embeddings to further improve performance. (00:23:11-00:28:06)

**Key Takeaways**

- Active learning can accurately predict antibody competition using only a small subset of experimental measurements.
- Project-specific models outperform universal sequence-only approaches for epitope binning.
- Random selection of training interactions provides better predictive performance than clustering-based sampling strategies.
- Integrating machine learning into laboratory workflows reduces resource requirements while preserving decision-quality results.

# Designing and Analysis of a Large Library-on-Library Dataset to Reveal Insights on Protein Stability across Different VHH:Antigen Complexes

Full Video Here

**Jurrian de Kanter, PhD,** Data Scientist, Genmab

Developing machine learning models that can design high-affinity antibodies requires more than increasingly sophisticated algorithms. Progress depends on generating larger, higher-quality datasets that capture meaningful antibody-antigen interactions, including both successful and unsuccessful binding events. Rather than sampling isolated interactions across the enormous design space, the proposed strategy builds dense clusters of related measurements around known complexes, creating richer datasets that better support model training and generalisation. (00:00:00-00:05:00)

The work combined experimentally generated library-on-library datasets with synthetic data to evaluate how different modelling approaches learn binding behaviour. Simulated datasets reproduced many broad experimental patterns and helped estimate the amount of data required for model development, although they could not fully capture the complexity of real biological interactions. This makes synthetic data valuable for planning experiments while reinforcing the need for extensive experimental validation. (00:05:00-00:11:03)

Comparisons between simulated and experimental datasets revealed an important distinction between protein stability effects and true interface-specific binding changes. Many mutations altered overall protein stability rather than directly affecting the binding interface, making affinity changes appear similar across unrelated antibodies. Careful experimental controls were therefore essential for separating genuine interface signals from broader effects on protein quality and folding. (00:11:03-00:18:56)

Existing structure-aware and language-based protein models successfully captured many stability-related patterns but remained less effective at predicting interface-specific interactions that determine binding affinity. The findings suggest that future advances will depend less on new architectures alone and more on systematically generated, standardised datasets that isolate biologically meaningful interaction signals. Such resources could provide the foundation needed for broadly generalisable antibody design models. (00:18:56-00:24:55)

**Key Takeaways**

- High-quality experimental data remain the primary constraint on developing generalisable antibody design models.
- Library-on-library datasets provide richer information than isolated mutational screening experiments.
- Separating protein stability effects from interface-specific binding changes is critical for accurate model training.
- Standardised, large-scale experimental datasets will be essential for improving next-generation protein engineering models.

# Maximise AI Potential in Biologics Discovery and Development: From Model Training to Consumption

Full Video Here

**Nicola Bonzanni,** CEO, ENPICOM

Successful AI adoption in biologics discovery depends on more than developing accurate models. Organizations also need reliable ways to operationalize those models so they become part of everyday scientific workflows and deliver measurable value. Achieving that goal requires close collaboration between laboratory scientists, machine learning teams, and digital infrastructure groups while overcoming fragmented data landscapes created by distributed research environments. (00:00:00-00:05:34)

Many organizations are progressing through three connected stages of AI maturity: establishing an AI-ready data foundation, automating analytical workflows, and integrating machine learning into routine decision making. Standardized data, consistent analytical pipelines, and accessible infrastructure help reduce variability while making experimental data available to both laboratory and computational teams. These capabilities create the foundation needed for scalable and trustworthy AI deployment. (00:05:34-00:10:42)

Operational success also depends on making machine learning accessible to scientists without requiring technical expertise. Model lifecycle management, version tracking, and straightforward deployment allow validated models to appear naturally within existing scientific workflows. Flexible software components can then be assembled into customized pipelines that accommodate different organizational processes while maintaining consistency, traceability, and scalability. (00:10:42-00:15:36)

Examples demonstrated how automated workflows can combine laboratory results with AI predictions while preserving key decision points for scientists. Integrating in vitro and in silico data supports more informed candidate selection without removing scientific oversight, helping build confidence in AI-assisted decisions. As open-source models continue to expand, organizations will increasingly benefit from platforms that enable rapid evaluation, validation, and deployment using their own experimental data. (00:15:36-00:27:37)

**Key Takeaways**

- AI adoption requires operational workflows as well as high-performing machine learning models.
- AI-ready data foundations and standardized pipelines are prerequisites for scalable deployment.
- Effective model lifecycle management improves accessibility, governance, and reproducibility.
- Human oversight remains essential when integrating AI predictions into scientific decision making.

# To What Extent Can Large-Language Models Represent 3D Information?

Full Video Here

**Isaac Ellmen,** Researcher, Oxford Protein Informatics Group, University of Oxford

Protein language models have traditionally treated proteins as linear amino acid sequences, while structural information has been handled separately using graph-based methods. This work explored whether standard transformer architectures can directly reason about three-dimensional protein structures when given atomic coordinates. Understanding this capability could simplify future model design by reducing the need for specialized geometric architectures while expanding the range of tasks that language models can address. (00:00:21-00:06:37)

The analysis examined how transformer attention mechanisms process spatial information and demonstrated that they can naturally focus on nearby structural features using relative distances between residues. Rather than requiring complex geometric operations, standard transformers were shown to develop internal representations that preserve meaningful three-dimensional relationships. These findings help explain why recent protein structure prediction models have achieved strong performance using increasingly conventional transformer architectures. (00:06:37-00:13:33)

Experimental comparisons showed that providing structural coordinates significantly improved protein language model performance across multiple downstream tasks. Models trained with coordinate information achieved better masked-token prediction, stronger molecular function prediction, and improved zero-shot stability prediction compared with sequence-only models. The greatest gains were observed for amino acids that play important roles in defining local protein structure, demonstrating the value of incorporating geometric information directly into model training. (00:13:33-00:18:17)

The results suggest that transformers are well suited for problems requiring reasoning across entire protein structures, including structure prediction and protein generation. However, specialized graph neural networks may still offer advantages when detailed local structural information is already available and highly precise property prediction is required. The same principles also enabled development of a highly efficient antibody structure prediction model capable of generating structures fast enough for real-time computational workflows. (00:18:17-00:21:43)

**Key Takeaways**

- Standard transformer architectures can learn meaningful three-dimensional protein representations from structural coordinates.
- Providing coordinate information substantially improves performance across multiple protein prediction tasks.
- Transformers excel at reasoning over whole-protein relationships, while graph neural networks remain valuable for detailed local structural analyses.
- Efficient transformer-based architectures can enable rapid antibody structure prediction suitable for high-throughput workflows.

# The Evolution of Biologics Engineering: Integrating AI into Biologics Discovery

Full Video Here

**Rebecca Croasdale-Wood, PhD,** Senior Director, Augmented Biologics Discovery & Design, Biologics Engineering, Oncology, AstraZeneca

Artificial intelligence is transforming biologics discovery by expanding the search space beyond the limits of traditional experimental approaches. An integrated strategy built around data generation, predictive modelling, and structured data management enables researchers to improve candidate diversity, accelerate optimisation cycles, and reduce development costs. Success depends on combining high-throughput laboratory platforms with machine learning in a continuous design-build-test workflow supported by high-quality, reusable data. (00:00:00-00:06:41)

Large-scale experimental data generation provides the foundation for effective AI model development. High-throughput screening platforms capture both positive and negative binding information across millions of variants, revealing rare but valuable candidates that conventional discovery methods may overlook. Automated production, characterisation, and functional testing then create rapid laboratory feedback loops that continuously improve computational models and expand their practical value. (00:03:25-00:13:00)

AI models are increasingly supporting lead generation, affinity optimisation, and developability assessment across active discovery programmes. Examples demonstrated how computational design reduced optimisation timelines while identifying candidates that retained desirable biological properties and resolved manufacturing liabilities. Predictive developability models also enabled early prioritisation of promising molecules, reducing unnecessary experimental work and improving the efficiency of downstream candidate selection. (00:05:02-00:19:50)

The long-term vision is an AI-first biologics discovery process built on harmonised laboratory automation, standardized data foundations, and integrated machine learning capabilities. Digital platforms that connect experimental workflows, predictive models, and user-friendly interfaces will allow scientists to make faster, more informed decisions while maintaining confidence in AI-assisted discovery. Equally important is investing in organisational change, training, and adoption so new technologies become part of routine scientific practice. (00:19:50-00:29:25)

**Key Takeaways**

- High-quality, large-scale experimental datasets are essential for effective biologics AI models.
- Integrated laboratory and AI workflows accelerate lead discovery and optimisation while improving candidate quality.
- Early developability prediction reduces experimental burden and supports better candidate selection.
- Successful AI transformation depends on robust digital infrastructure, standardized data, and scientist adoption.
- Simplified upstream production can lower costs while maintaining protein quality and functionality.

# From Strategy to Scale: Engineering the Future of Biologics with Generative AI

Full Video Here

**Stef van Grieken,** CoFounder & CEO, Cradle

Successfully applying generative AI in biologics requires far more than developing accurate machine learning models. The greatest challenge lies in building reliable software systems that integrate models into everyday scientific workflows while supporting collaboration between computational and laboratory teams. Lessons drawn from large-scale machine learning deployment show that engineering, usability, and continuous learning become increasingly important as AI moves from isolated projects into routine research. (00:00:00-00:08:09)

Robust AI platforms must account for the realities of biological data, including experimental variability, limited datasets, and uncertain predictions. Rather than relying solely on precise numerical predictions, resilient approaches emphasize comparative learning, confidence estimation, and continuous validation through laboratory experiments. Building systems that recognize when predictions fall outside their domain of expertise is essential for maintaining scientific reliability. (00:08:09-00:14:05)

Model performance also depends on the quality of the overall learning system rather than the choice of any single architecture. Integrating information across multiple assays, automatically incorporating relevant historical data, and training models to understand relationships among different molecular properties improves generalization. Continuous experimental testing allows models to evolve alongside laboratory programmes while reducing dependence on manually tuned workflows. (00:14:05-00:20:35)

Generative AI should optimize experimental design as well as molecular design by selecting experiments that both identify promising candidates and improve future model performance. Secure infrastructure, scalable computing resources, and intuitive user experiences are equally important for successful adoption in pharmaceutical organizations. As AI platforms mature, the greatest gains are expected to come from tightly coupling computation with rapid experimental feedback, enabling progressively shorter design-build-test-learn cycles. (00:20:35-00:29:04)

**Key Takeaways**

- Scalable AI adoption depends as much on software engineering and workflow design as on model accuracy.
- Reliable biological AI systems require uncertainty estimation, continuous validation, and robust handling of experimental variability.
- Learning across multiple assays and molecular properties improves model generalization more effectively than isolated predictors.
- AI platforms should optimize both molecule selection and experimental design to accelerate iterative discovery.

# Evaluation of Digital Protein Design Tools in an Industry Setting

Full Video Here

**Speakers: Karin Hrovatin,** Bioinformatic Scientist, Merck KGaA, and **Stephanie Linker, PhD,** Senior Computational Biochemist, Merck Group

Open-source protein AI models provide a strong foundation for protein engineering, but significant adaptation is required before they can support real-world research workflows. The presentation described how external innovations are evaluated, refined, and integrated into internal development pipelines while maintaining an active commitment to contributing improvements back to the open-source community. Building practical solutions depends on combining robust machine learning with high-quality experimental workflows and close collaboration across industry and academia. (00:00:00-00:11:07)

Developing reliable property prediction models requires careful attention to experimental data quality, uncertainty, and validation strategies. Rather than treating all measurements equally, the workflow accommodates noisy and non-numerical experimental outcomes while preventing information leakage during model evaluation. The presenters also emphasized that realistic data splitting, uncertainty estimation, and ensemble modelling are essential for generating predictions that generalize to new mutation sites instead of simply memorizing historical observations. (00:02:26-00:11:07)

The second part focused on structure prediction using open-source models and demonstrated how practical workflows can substantially improve performance without modifying the underlying algorithms. Benchmarking against internal datasets showed meaningful gains for challenging antibody-antigen structure prediction by combining diverse sampling strategies with confidence-based ranking. Similar workflow enhancements also improved small-molecule modelling by correcting chemical representations and refining predicted structures through automated post-processing. (00:11:23-00:22:22)

The broader objective is to make advanced AI capabilities directly accessible to laboratory scientists through intuitive interfaces that combine sequence modelling, structure prediction, and experimental knowledge in a single environment. Continued benchmarking, workflow development, and open-source collaboration were presented as key drivers for improving model robustness while accelerating adoption across protein engineering programmes. (00:22:22-00:31:23)

**Key Takeaways**

- Open-source AI models require substantial workflow engineering before they are ready for production research.
- Robust property prediction depends on careful handling of noisy data, realistic validation, and uncertainty estimation.
- Workflow improvements can significantly enhance structure prediction performance without retraining foundation models.
- Integrating AI tools into accessible scientific workflows and contributing improvements back to the community accelerates innovation.

# End-to-End AI Applications for Advanced Antibody Engineering and Multispecifics Design

Full Video Here

**Mary Ann Pohl,** Director, Alliance Management, Ailux Biologics by XtalPi

Artificial intelligence can deliver the greatest impact in biologics discovery when integrated across the entire design-build-test cycle rather than applied to isolated tasks. The platform presented combines curated experimental datasets with structure prediction, generative design, and predictive modelling to guide antibody engineering from target selection through experimental validation. This end-to-end workflow focuses on addressing challenging discovery problems where conventional approaches are often limited. (00:00:21-00:06:25)

Three primary application areas illustrated this strategy: difficult targets, multi-parameter antibody engineering, and multispecific antibody design. AI-guided antigen stabilization and epitope-focused design improve opportunities for generating functional binders against challenging targets, while repertoire analysis helps identify rare but valuable antibody clones that might otherwise be overlooked. Structure-guided analysis also expands candidate diversity by identifying functionally related antibodies with distinct sequences. (00:06:25-00:12:46)

The platform applies iterative active learning to optimize multiple molecular properties simultaneously rather than improving one characteristic at a time. Case studies demonstrated rapid improvements in affinity, selectivity, and developability by combining structure prediction, generative sequence design, predictive screening, and experimental feedback over successive design rounds. Similar workflows accelerated antibody affinity maturation while substantially reducing the number of variants requiring laboratory evaluation. (00:12:46-00:19:52)

Additional examples showed how AI can address common engineering bottlenecks such as Fab-to-scFv conversion and common light chain discovery for multispecific antibodies. By preserving antigen binding while improving expression, stability, and developability, these workflows reduced experimental burden and enabled further optimization toward defined target product profiles. The overall approach demonstrates how integrated AI systems can accelerate antibody engineering while supporting increasingly complex therapeutic formats. (00:19:52-00:28:26)

**Key Takeaways**

- End-to-end AI platforms combine structure prediction, generative design, predictive modelling, and laboratory validation within a unified workflow.
- Active learning enables simultaneous optimization of multiple therapeutic properties while reducing experimental workload.
- AI-guided engineering can accelerate affinity maturation, difficult target discovery, and multispecific antibody development.
- Iterative computational and experimental feedback improves candidate quality while shortening design cycles.