NextGenPLM: A Novel Structure-Infused Foundational Protein Language Model for Antibody Discovery and Optimization

Abhinav Gupta [*1] **Ruijiang Li** [*2] **Camila Leal** [1] **Madhumati Sevvana** [1] **Brian Christopher Mackness** [1] **Ryan Gavin Casner** [1] **Joseph D Batchelor** [1] **Anna Park** [1] **Michael Bailey** [3] **Lorenzo Kogler Anele** [3] **Sven Jager** [4] **Norbert Furtmann** [5] **Yves Fomekong Nanfack** [1] **Maria Wendt** [1] **Yu Qiu** [1 6]

Abstract

Sequence-only PLMs lack spatial context and miss critical folding, interface, and environment-dependent cues, while structure-prediction and docking methods are too slow and underperform on antibody complexes. *NextGenPLM* bridges this gap with a modular, scalable design that fuses pretrained PLMs with multimodal inputs—from raw sequences and functional assays to high-resolution structures—via spectral contact-map embeddings. It natively models multi-chain antigen structures and processes four complexes per second, enabling real-time, repertoire-scale insights such as epitope clustering. On a diverse benchmark of antibody–antigen complexes, *NextGenPLM* matches Chai-1 and Boltz-1x on contact-map and epitope accuracy at a fraction of the compute cost. In an internal affinity-maturation campaign, ranking mutants by predicted contact probabilities and masked-language-modeling (MLM) log-likelihoods helped achieve up to 17× affinity improvements—demonstrating its potential for rapid, data-driven biologics discovery.

[*]Equal contribution [1]Large Molecule Research, Sanofi, Cambridge, MA, United States [2]R&D Data & Computational Science, Sanofi, Cambridge, MA, United States [3]R&D Data & Computational Science, Sanofi, Toronto, ON, Canada [4]R&D Data & Computational Science, Sanofi, Frankfurt, Germany [5]Large Molecule Research, Sanofi, Frankfurt, Germany [6]Present address: Biologics Engineering, Oncology R&D, AstraZeneca, Waltham, MA, United States. Correspondence to: Abhinav Gupta <abhinav.gupta@sanofi.com>, Norbert Furtmann <norbert.furtmann@sanofi.com>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

1. Introduction

Target engagement is the defining requirement for antibody therapeutics, yet today's leading protein language models are either antibody specific (Burbach & Briney, 2024; Tobias H. Olsen & Deane, 2022) or trained on single chain proteins (Lin et al., 2023), so they cannot model an antibody and its antigen together. While these PLMs excel at predicting self-contained properties such as thermal stability (Chen et al., 2025), they struggle with target dependent functions like binding affinity and epitope specificity that drive early discovery. Common workarounds—building downstream predictors on costly, target-specific data or applying human-like CDR mutations that occasionally stabilize loops (Ewert et al., 2003)—either require expensive experiments or yield unpredictable improvements.

Structure prediction tools that handle multiple chains, including AlphaFold-Multimer (Evans et al., 2021), Chai-1 (Chai Discovery, 2024), and Boltz-1x (Wohlwend et al., 2024), can in principle capture both partners, but they are too slow and resource intensive for screening millions of antibody variants, and their ranking scores often fail to correlate with real binding measurements. *NextGenPLM* (Next Generation Protein Language Model) is a truly multimodal foundation model and overcomes these limitations. It seamlessly integrates diverse representations (sequence, structure, surface) and varied data sources (functional assays, epitope mapping, predicted or experimental 3D structure), without retraining from scratch or redundant computation, to reliably predict antibody–antigen interactions at scale.

Key Design Principles and Technical Contributions

- **Multimodal transformer backbone.** We ingest and jointly embed raw sequences, functional assays, and high-resolution structural data (via *spectral embeddings* and contact-maps) into a single transformer—avoiding costly Evoformer blocks.

- **Leverage and extend pretrained PLMs.** By building on off-the-shelf PLMs covering 10^8+ protein/antibody sequences, we maintain broad sequence coverage with-

out retraining, while explicitly injecting antigen contact maps so the model needn't reconstruct 3D geometry. Thus focusing compute on antibody–antigen recognition.

- **Structure and surface priors.** IMGT numbering (Manso et al., 2022), solvent-accessibility bins, and contact-map features are fused as input or as part of training tasks, improving interpretability and reducing learning overhead.
- **Robustness for real-world applicability** We handle practical structural variability—multi-chain antigens, missing residues, or incomplete coordinates—by incorporating low-resolution features like coarse contact maps. This reduces sensitivity to fine-grained structural details, making the model more applicable in scenarios where precise structural information is limited.
- **Epitope & contact-strength heads.** We introduce specialized heads for epitope prediction and per-residue contact-map regression that match state-of-the-art structure predictors (e.g. Chai-1, Boltz-1x) at ∼100× higher throughput.
- **Rapid affinity maturation.** By ranking mutants via summed high-confidence contact probabilities and filtering with MLM log-likelihood, we demonstrate immediate, rapid, and high-throughput leads optimization building on existing experimental data.
- **Scalable, extensible design.** The entire pipeline accepts antigen structure (experimental or predicted), skips full-atom antibody folding, and remains flexible for future data modalities or assay types, while preserving transformer-level efficiency for repertoire-scale screening.

1.1. Related Works

Beyond classic sequence-only PLMs [e.g. ESM (Lin et al., 2023), ProtTrans (Elnaggar et al., 2021), BALM (Burbach & Briney, 2024), AbLang (Tobias H. Olsen & Deane, 2022)], and dedicated structure-prediction or docking engines [Chai-1 (Chai Discovery, 2024), Boltz-1x (Wohlwend et al., 2024), GeoDock (Chu et al., 2023)], a variety of hybrid approaches have emerged—weakly embedding structural or interaction cues, or training epitope/paratope heads. Yet they fall short on multiple fronts. For example, PPI-focused frameworks like MINT (Ullanat et al., 2025) or D-SCRIPT (Sledzieski et al., 2021) leverage abundant co-evolutionary data but lack the specificity and structural detail required for antibody–antigen binding.

Graph- and CNN-based paratope/epitope predictors such as AsEP (Liu et al., 2024) and ParaSurf (Papadopoulos et al., 2025) require high-fidelity 3D structures, while purely sequence-based methods (AntiBinder; Zhang et al. 2025) or those using only limited pairwise antigen features (EpiScan; Wang et al. 2024) suffer from reduced accuracy and the same scalability challenges as Evoformer-style models. None approach the binding-site precision of leading multimeric predictors like AlphaFold-Multimer.

2. Methods

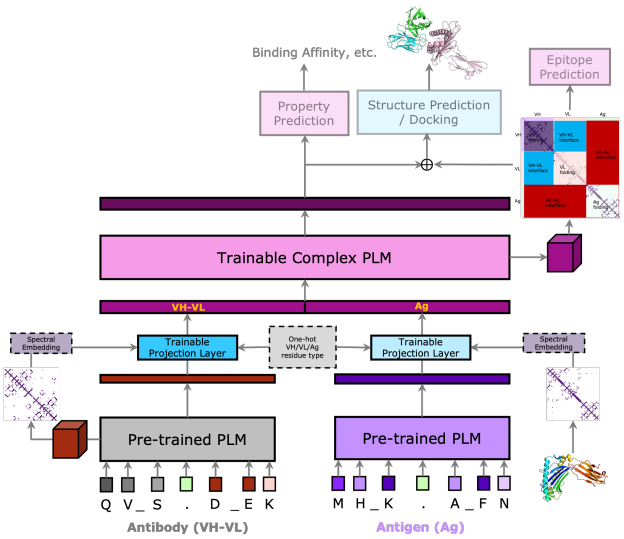

Figure 1. **Schematic of the *NextGenPLM* architecture**. Residue embeddings from frozen PLMs (VH–VL, Ag) are combined with chain IDs and spectral contact-map embeddings (using SignNet), projected onto a shared space, and processed by a transformer that models interactions for the full complex. Downstream heads predict masked tokens, RSA bins, and contacts; training uses structural and entropy-based masking for robustness. A zoomed replica of this schematic is provided in appendix Figure 8.

The overall architecture of *NextGenPLM* is illustrated in Figure 1.

2.1. Model Overview

NextGenPLM is a modular transformer-based architecture designed to model antibody–antigen complexes using a combination of frozen pretrained protein language models (PLMs), contact-map-derived spectral embeddings, and structural priors. Our design avoids full 3D antibody structure input while leveraging the antigen's structure (experimental or predicted), enabling high-throughput evaluation of thousands of antibodies against a small set of antigens—a typical regime in real-world discovery workflows.

Unlike Evoformer-based predictors (AlphaFold2; Jumper et al. 2021) with $\mathcal{O}(L^3 + sL^2)$ complexity, *NextGenPLM* maintains $\mathcal{O}(L^2)$ scaling by linearizing 3D context into

compact spectral encodings (top-k eigenvectors of the normalized Laplacian). Antibody geometry is captured by a low-resolution contact stencil, inferred from PLM attention heads via a lightweight classifier.

Residue representations are formed by concatenating PLM embeddings, spectral encodings (via SignNet; Lim et al. 2022), and one-hot chain IDs, then projecting into a shared latent space. A transformer encoder models cross-chain interactions, with specialized heads for masked token recovery, solvent accessible area classification, and contact-map regression.

All models referenced in this paper use `esm2_t33_650M_UR50D` (Lin et al., 2023) as the frozen pretrained PLM for both antibody and antigen inputs. For the trainable complex PLM, we initialize its architecture and weights from `esm2_t30_150M_UR50D` (Lin et al., 2023). Although an antibody-specific PLM could be justified, we did not perform that ablation—instead, we reuse the same general model for both Ab and Ag to reduce GPU-memory usage. However, the framework is fully capable of swapping in distinct pretrained PLMs for antibody versus antigen when desired.

2.2. Learning Objectives and Masking Strategies

NextGenPLM is trained using a multi-task objective that encourages the model to integrate both sequence-level and structure/surface-aware information. The training tasks include masked language modeling (MLM), contact map regression, and solvent accessible area classification.

Entropy-aware MLM. Antigen tokens are masked at random; for VH–VL we bias masking toward high-entropy IMGT positions (Complementarity Determining Regions; CDRs) to focus learning on key recognition regions.

Structural masking. We randomly remove rows/columns from input contact maps, recomputing spectral embeddings on these corrupted maps and filling masked positions with random vectors—training the model to handle incomplete structures.

Solvent accessible area classification. We bin Relative Solvent Accessibility (Tien et al., 2013) into five categories and train a per-residue classifier, helping the model distinguish buried versus exposed residues.

Weighted contact map loss. We predict binary contacts (intra-Ab, intra-Ag, inter-chain) and apply separate loss weights based on each category's class imbalance to emphasize true binding interactions.

Together, these objectives guide the model to align sequence, structure, surface, and interaction patterns effectively. Moreover, the architecture is designed for extensibility—additional heads (e.g., for any experimentally measured property, whether functional or structural) can be plugged in and fused seamlessly into the foundational model.

2.3. Training Data and Hierarchical Training

To overcome the scarcity of high-quality antibody–antigen complexes, we curate a multimodal dataset combining public Protein Data Bank (Berman et al., 2000) structures (crystallography, cryoEM released before March, 2024 (Gupta et al., 2025)), internal proprietary complexes, and binding sequence pairs from patents and assays. This diverse collection supplies sequence, functional, structural, and surface annotations that together capture the complex, context-dependent rules of Ab–Ag recognition.

We then train in successive rounds to build capability. First, we learn to interpret spectral embeddings on general multi-chain protein complexes with masked language modeling (MLM) and contact-map prediction. Next, we fine-tune on curated Ab–Ag structures to teach cross-chain interactions. We follow with binding sequence rounds—MLM on both chains plus antigen contact supervision—and finally interleave sequence-only and structure-based stages with an added solvent-accessibility classification task. Throughout, we apply structure-similarity weighting and use a 12 Å contact threshold to broaden interaction neighborhoods and ensure stable, progressive learning.

2.4. Epitope Prediction Head

To predict epitope residues with spatial precision (Figure 4), we add a graph-based head on top of the core *NextGenPLM* embeddings, keeping the trained foundational model frozen. We first pool antibody information into each antigen residue by combining contextual embeddings with learnable IMGT positional and chain-type (VH/VL) encodings, then concatenate and project this antibody-informed vector with the antigen's own embedding. We build an antigen residue graph using an 8 Å distance cutoff to capture fine structural neighborhoods and apply a lightweight GCN (Kipf & Welling, 2016) over this graph to refine node representations. A final per-node classifier then produces epitope probabilities, yielding interpretable, residue-level predictions that leverage both sequence context and local 3D structure. Further architectural details are provided in Appendix A.

3. Experimental Results

This section presents preliminary results from ongoing work. We are actively expanding our evaluation to include more public benchmark datasets and fully disclosed experimental assays and targets. The figures and metrics that follow reflect performance on our current internal and proprietary data.

3.1. Contact and Epitope Prediction Benchmark

Although our model lacks the intricate architecture and full-atom resolution of AlphaFold-style methods, we still compared *NextGenPLM* to state-of-the-art multimers Chai-1 and Boltz-1x by generating paired MSAs, sampling five conformations each, and evaluating the top-ranked model. On 112 diverse antibody–antigen complexes—selected for both sequence and structural variety (some public PDB entries released after May 30, 2024, plus proprietary structures (Gupta et al., 2025))—we defined contacts and epitopes via 12 Å and 8 Å distance cutoffs and report balanced accuracy on VH–VL versus antigen residues.

As shown in Figure 2, *NextGenPLM* matches Chai-1 and Boltz-1x on both tasks and exhibits thicker violins at higher accuracy—indicating a stronger tendency to partially capture true epitope residues. Because cross-attention is only weakly constrained by the contact-map objective, missing a few epitope positions still allows meaningful attention on neighboring residues.

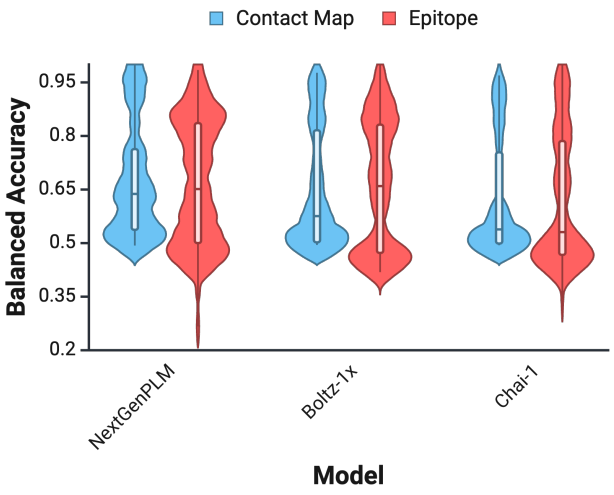

Figure 2. **Contact map and epitope prediction performance comparison.** The violin plots are based on the balance accuracy of each sample in the benchmark dataset. *NextGenPLM* matches the accuracy of state-of-the-art complex structure predictors and displays noticeably thicker violins at higher balanced-accuracy values—indicating a stronger tendency to partially capture true epitope residues.

In addition to accuracy gains, our model offers substantial efficiency improvements. On a single NVIDIA A10G GPU, our throughput reaches 4 complexes per second, compared to $\mathcal{O}$(1 min) per complex for Boltz-1x and Chai-1.

3.2. Antibody Affinity Maturation

Affinity maturation by structure-based design seeks mutations that increase both the number and strength of antibody–antigen contacts. Although *NextGenPLM* is trained on proxy objectives, we hypothesize its predicted contact maps—and other output features—encode real affinity information. Rather than building a separate regression or classification model on binding data, we tested ranking multi-point mutant libraries of a wild-type antibody against its antigen using only (a) summed high-confidence contact probabilities and (b) MLM log-likelihoods conditioned on the wild-type antibody and antigen. Details can be found in Appendix B.2.

To validate this approach, we applied it in an internal affinity-maturation campaign against an undisclosed target. *NextGenPLM* was trained on proprietary complex structures specific to this antigen. After two experimental rounds using a high-throughput screening of $\sim$1,140 variants, the team identified a top variant with three mutations and 3-fold improved binding over wild-type. Using that variant as the new parent, we generated about 350 mutants via our contact-map and likelihood ranking strategy (Figure 5) and benchmarked them against alternative, mutation-informed approaches. Top candidates—triaged again by high-throughput screening—were confirmed by the surface plasmon resonance (SPR) technique: NextGenPLM's designs, each bearing three to four non-intuitive additional mutations, achieved up to 17× affinity gains (several showing 10–15×) over wild-type (Figure 3). Future cycles can leverage the model's dynamic feedback loop to further refine rankings, pinpoint binding sites, and guide rational design.

3.3. Ablation Study

We performed an ablation study to isolate the contributions of key *NextGenPLM* components: spectral embeddings for 3D structure, the MLM objective on VH–VL/antigen pairs, and inclusion of proprietary internal data. Figure 7 reports AUC–PR and balanced accuracy for predicted contact probabilities between antibody and antigen residues. Spectral embeddings drive the largest gains in AUC–PR, while training on binding pairs most improves balanced accuracy—demonstrating that both structural encoding and interaction-specific data are essential to our model's success.

4. Conclusion and Future Works

In summary, *NextGenPLM* represents a new paradigm in multimodal foundational models for antibody–antigen modeling by integrating sequence, experimental 3D structure, and interaction data in a single, efficient transformer. Leveraging pretrained language models, explicit structural priors, and supervision across diverse modalities, it matches state-of-the-art multimer predictors on contact-map and epitope benchmarks while training at a small fraction of the cost and delivering orders-of-magnitude faster inference. Our affinity-maturation workflow further proves its practi-

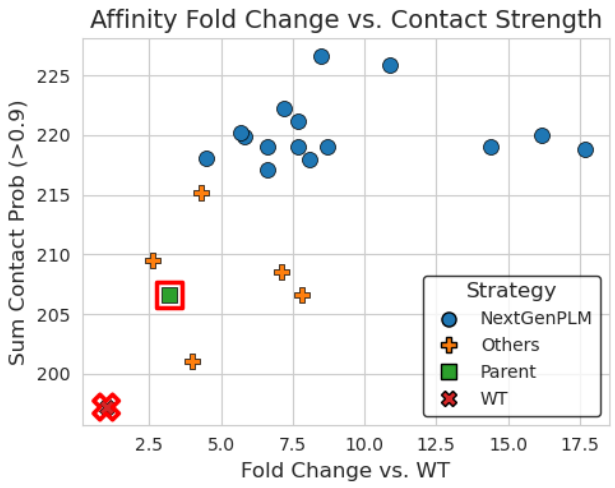

Figure 3. **Correlation of predicted contact strength with SPR-measured affinity gains.** Each point represents a mutant variant from the affinity-maturation campaign, plotted by its summed high-confidence contact probability and its fold-change in binding measured by surface plasmon resonance (SPR). The original three-mutation parent (*green*) shows a 3× improvement, while structure-infused NextGenPLM's top candidates—with three to four additional, non-intuitive mutations—achieve up to 17× gains over wild-type.

cal impact—achieving up to 17× gains measured by SPR in an internal campaign without any binding-data supervision. Looking ahead, this flexible framework can absorb additional data types—epitope binning, functional assays, even NANOBODY® VHH–antigen or RNA–protein interactions—and evolve through dynamic feedback loops to enable truly high-throughput, data-driven biologics discovery.

Acknowledgments

We thank Cendrine Lemoine, Paul Ferrari, and Dietmar Hoffmann and his team for their invaluable support during the affinity-maturation campaign; the BioAIM team at Sanofi—especially Hervé Minoux and Andrew Phillips—for providing the compute resources that made this work possible; and our leadership, Rebecca Sendak and Thorsten Schmidt, for their ongoing support and encouragement. Finally, we are grateful to the reviewers and the FM4LS workshop chairs for their constructive feedback.

Impact Statement

This paper contributes to advancing the intersection of machine learning and biologics-based therapeutic discovery. The approach has the potential to accelerate drug development, improve early candidate selection, and ultimately reduce costs associated with bringing life-saving treatments to patients. While we do not identify specific societal risks at this time, we recognize the importance of ensuring such technologies are used responsibly and equitably across healthcare and research contexts.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Burbach, S. M. and Briney, B. Improving antibody language models with native pairing. *Patterns*, 5(5), 2024.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

Chen, B., Cheng, X., Li, P., Geng, Y.-a., Gong, J., Li, S., Bei, Z., Tan, X., Wang, B., Zeng, X., et al. xtrimopglm: unified 100-billion-parameter pretrained transformer for deciphering the language of proteins. *Nature Methods*, pp. 1–12, 2025.

Chu, L.-S., Ruffolo, J. A., Harmalkar, A., and Gray, J. J. Flexible protein-protein docking with a multi-track iterative transformer. *Protein Science*, pp. e4862, 2023.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Ewert, S., Huber, T., Honegger, A., and Plückthun, A. Biophysical properties of human antibody variable domains. *Journal of molecular biology*, 325(3):531–553, 2003.

Gupta, A., Rivero, B. M., Roel-Touris, J., Li, R., Furtmann, N., Nanfack, Y. F., Wendt, M., and Qiu, Y. SNAC-DB: The hitchhiker's guide to building better predictive models of antibody & NANOBODY® VHH–antigen complexes. In *DataWorld: Unifying Data Curation Frameworks Across Domains, Workshop at the 42nd International Conference on Machine Learning (ICML 2025)*, Vancouver, Canada, 2025. URL https://openreview.net/forum?id=68DcIpDaHK.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Lim, D., Robinson, J., Zhao, L., Smidt, T., Sra, S., Maron, H., and Jegelka, S. Sign and basis invariant networks for spectral graph representation learning. *arXiv preprint arXiv:2202.13013*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. ESM-2 model family. Checkpoints used in this work: `esm2_t33_650M_UR50D` (https://huggingface.co/facebook/esm2_t33_650M_UR50D) and `esm2_t30_150M_UR50D` (https://huggingface.co/facebook/esm2_t30_150M_UR50D).

Liu, C., Denzler, L., Chen, Y., Martin, A., and Paige, B. Asep: Benchmarking deep learning methods for antibody-specific epitope prediction. *arXiv preprint arXiv:2407.18184*, 2024.

Manso, T., Folch, G., Giudicelli, V., Jabado-Michaloud, J., Kushwaha, A., Nguefack Ngoune, V., Georga, M., Papadaki, A., Debbagh, C., Pegorier, P., et al. Imgt® databases, related tools and web resources through three main axes of research and development. *Nucleic acids research*, 50(D1):D1262–D1272, 2022.

Papadopoulos, A.-M., Axenopoulos, A., Iatrou, A., Stamatopoulos, K., Alvarez, F., and Daras, P. Parasurf: A surface-based deep learning approach for paratope-antigen interaction prediction. *Bioinformatics*, pp. btaf062, 2025.

Sledzieski, S., Singh, R., Cowen, L., and Berger, B. Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model. *BioRxiv*, pp. 2021–01, 2021.

Tien, M., Meyer, A. G., Sydykova, D. K., Spielman, S. J., and Wilke, C. O. Maximum allowed solvent accessibilities of residues in proteins. *PLOS ONE*, 8(11):e80635, 2013. doi: 10.1371/journal.pone.0080635.

Tobias H. Olsen, I. H. M. and Deane, C. M. Ablang: An antibody language model for completing antibody sequences. *bioRxiv*, 2022. doi: https://doi.org/10.1101/2022.01.20.477061.

Ullanat, V., Jing, B., Sledzieski, S., and Berger, B. Learning the language of protein–protein interactions. *bioRxiv*, pp. 2025–03, 2025.

Wang, C., Wang, J., Song, W., Luo, G., and Jiang, T. Episcan: accurate high-throughput mapping of antibody-specific epitopes using sequence information. *NPJ Systems Biology and Applications*, 10(1):101, 2024.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Zhang, G., Li, X., Guo, Y., Wang, Y., and Yuan, S. A SARS-CoV-2 neutralizing antibody. https://doi.org/10.2210/pdb7F0X/pdb, 2021. PDB ID: 7F0X; X-ray diffraction at 2.80 Å resolution.

Zhang, K., Tao, Y., and Wang, F. Antibinder: utilizing bidirectional attention and hybrid encoding for precise antibody–antigen interaction prediction. *Briefings in Bioinformatics*, 26(1):bbaf008, 2025.

A. Epitope Prediction Head

Final antigen residue embeddings from the frozen *NextGenPLM* are enriched with antibody context. Specifically, for each antibody–antigen residue pair, we concatenate attention features with IMGT positional embeddings and one-hot VH/VL chain-type indicators corresponding to the antibody residue. These concatenated vectors are projected onto a latent space and pooled across antibody residues to obtain a contextualized representation for each antigen residue.

The resulting antibody-aware representations are further processed by a graph convolutional network (GCN) constructed over the antigen structure. In this graph, each node represents an antigen residue, and edges are defined based on an 8 Å threshold applied to the Cα–Cα distance matrix. The refined node embeddings are then passed to a residue-level classifier that produces epitope probability scores. This architecture enables interpretable and spatially localized epitope predictions, conditioned on both the antigen structure and the surrounding antibody context.

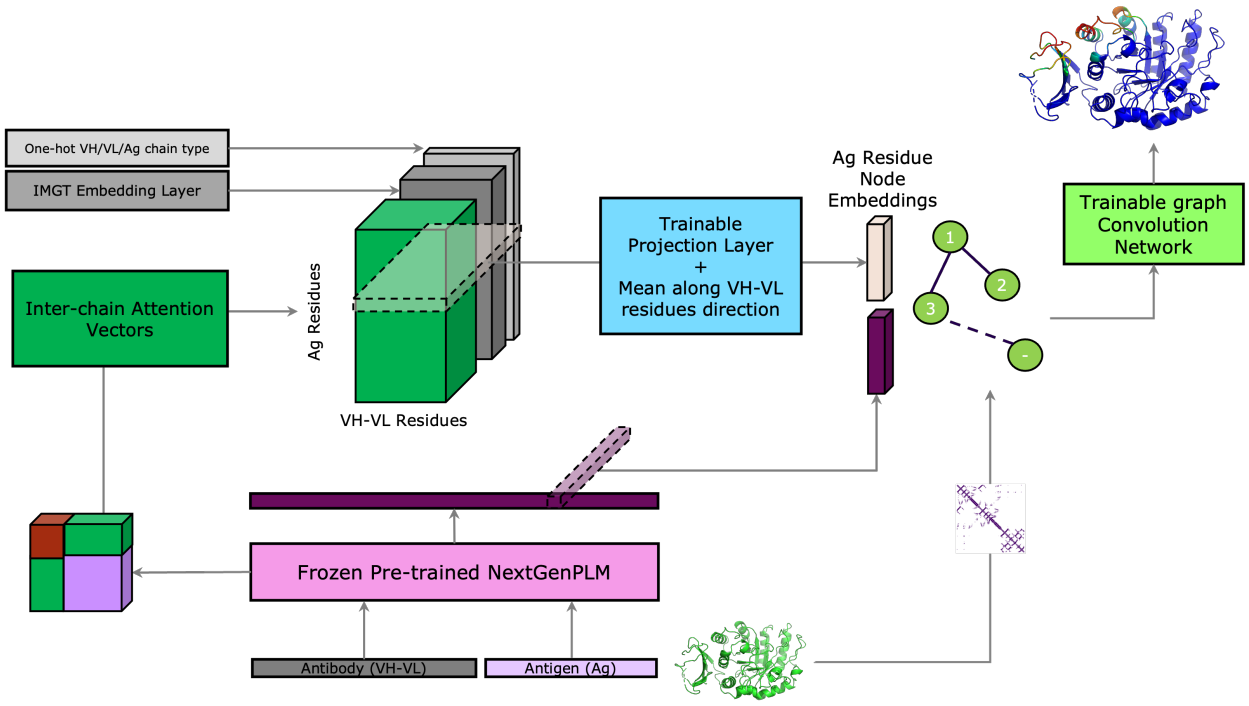

Figure 4. **Architecture of the epitope prediction head**. For details see Appendix A.

B. Affinity Maturation

To propose mutations for affinity maturation, we first use *NextGenPLM* to generate single-point mutation probabilities for each position in the VH and VL chains of parent antibody. For each position, candidate mutations are selected by applying a probability threshold to identify likely alternatives. These single-point mutations are then randomly combined to generate a pool of 100,000 mutant sequences.

To prioritize the proposed mutations, we compute two ranking metrics: joint mutation probability and contact strength. The overall workflow is illustrated in Figure 5.

B.1. Joint Mutation Probability Approximation

To score each multi-point mutant's plausibility given both the parent antibody and the antigen, one could approximate the joint probability of its k mutations in two ways as described below.

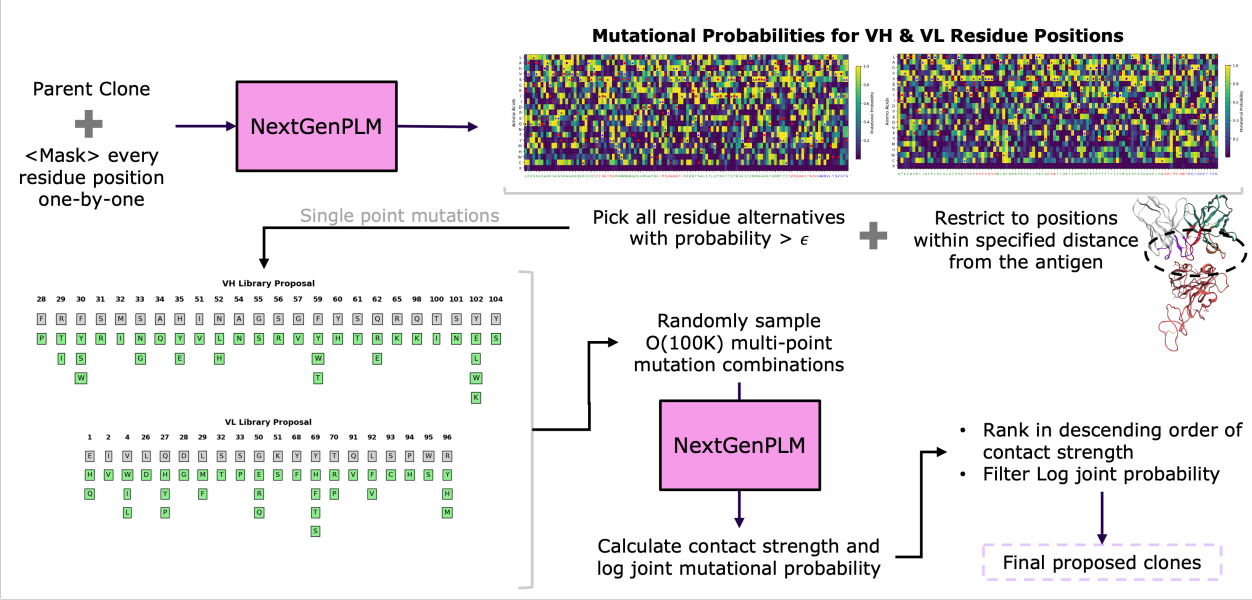

Figure 5. **Schematic a general affinity maturation workflow**. For details see Appendix B. If target-specific experimental 3D structures are available, they can be leveraged by adding it to the training data and performing a short fine-tuning round. Here, PDB 7F0X (Zhang et al., 2021) is shown as a representative complex.

First, the common "parent-masking" heuristic conditions on the antigen while masking each mutation in the parent sequence:

$$P(m_1, \ldots, m_k \mid \text{parent}, \text{Ag}) \approx \prod_{i=1}^{k} P\big(m_i \mid (\text{parent} \setminus m_i), \text{Ag}\big),$$

which requires at most L forward passes (with L the sequence length).

Second, the more accurate "variant-masking" approximation also conditions on the antigen but masks each mutation in the full mutant ("variant") sequence:

$$P(m_1, \ldots, m_k \mid \text{parent}, \text{Ag}) \approx \prod_{i=1}^{k} P\big(m_i \mid (\text{variant} \setminus m_i), \text{Ag}\big),$$

capturing inter-mutation dependencies at the cost of: number of variants $\times$ k forward passes.

In our workflow we adopt the variant-masking approximation to filter out implausible sequences, trading additional compute for more faithful likelihood estimates.

B.2. Contact Strength

We predict the binary contact map between the antibody and antigen, and compute a corresponding *contact strength* metric:

$$\text{contact\_strength} \;=\; \sum_{i \in \text{Ab},\, j \in \text{Ag}} \mathbf{1}\{p_{ij} > \tau\}\; p_{ij}\,,$$

where p_{ij} denotes the model-predicted contact probability between antibody residue i and antigen residue j, and τ (e.g., 0.9) is a high-confidence threshold. This summation captures both the quantity and intensity of predicted high-confidence contacts, reflecting the biophysical intuition that stronger and more numerous interactions typically correlate with higher binding affinity.

Figure 6 illustrates the relationship between observed fold changes and the approximated contact strengths for mutants from

a previous wave of affinity maturation. The results suggest a clear trend: higher contact strength is associated with a greater upper bound on binding affinity.

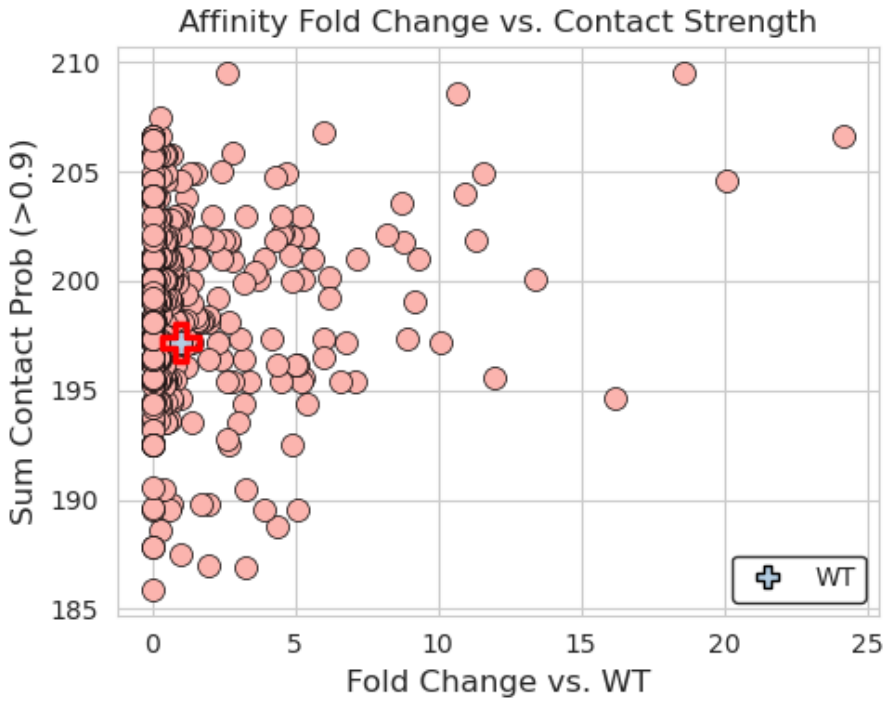

Figure 6. **Relationship between predicted contact strength and fold-change over wild-type measured by high-throughput screening.** Each point corresponds to a variant from earlier affinity-maturation rounds, with contact strength computed as the sum of high-confidence predicted inter-chain contacts and fold-change representing binding improvement relative to wild-type. Continuous enrichment of higher fold-change variants at greater contact strengths demonstrates the predictive value of our contact-map based metric.

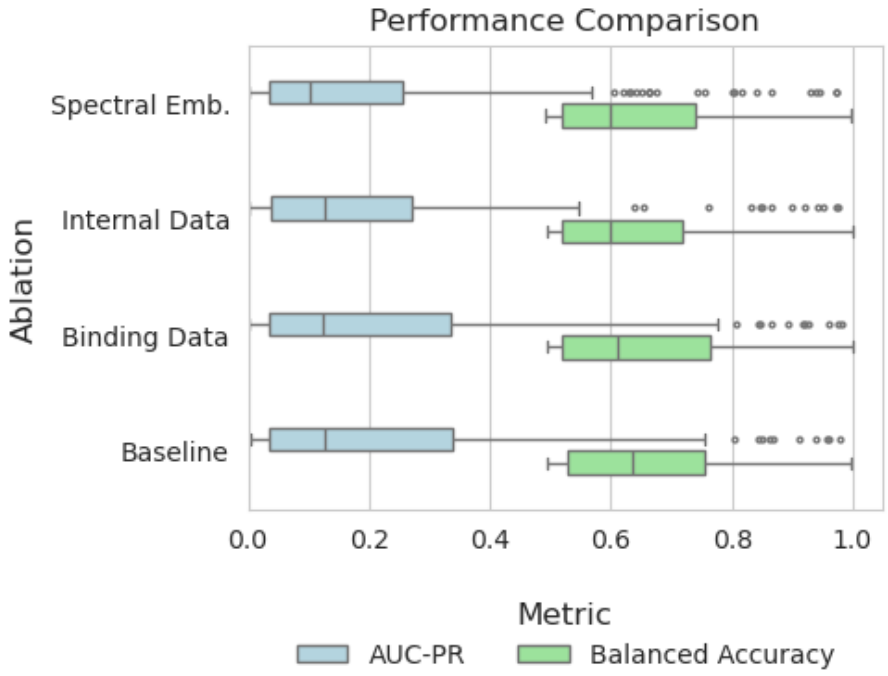

Figure 7. **Ablation study on *NextGenPLM* components.** Box-plots corresponds to AUC–PR and balanced accuracy for predicted inter-chain contact probabilities measured on benchmarking data used in Section 3.1. Results are plotted bottom to top for the full model and then after removing, in order: (1) the MLM task on VH–VL/antigen pairs, (2) proprietary internal training data, and (3) spectral embeddings.

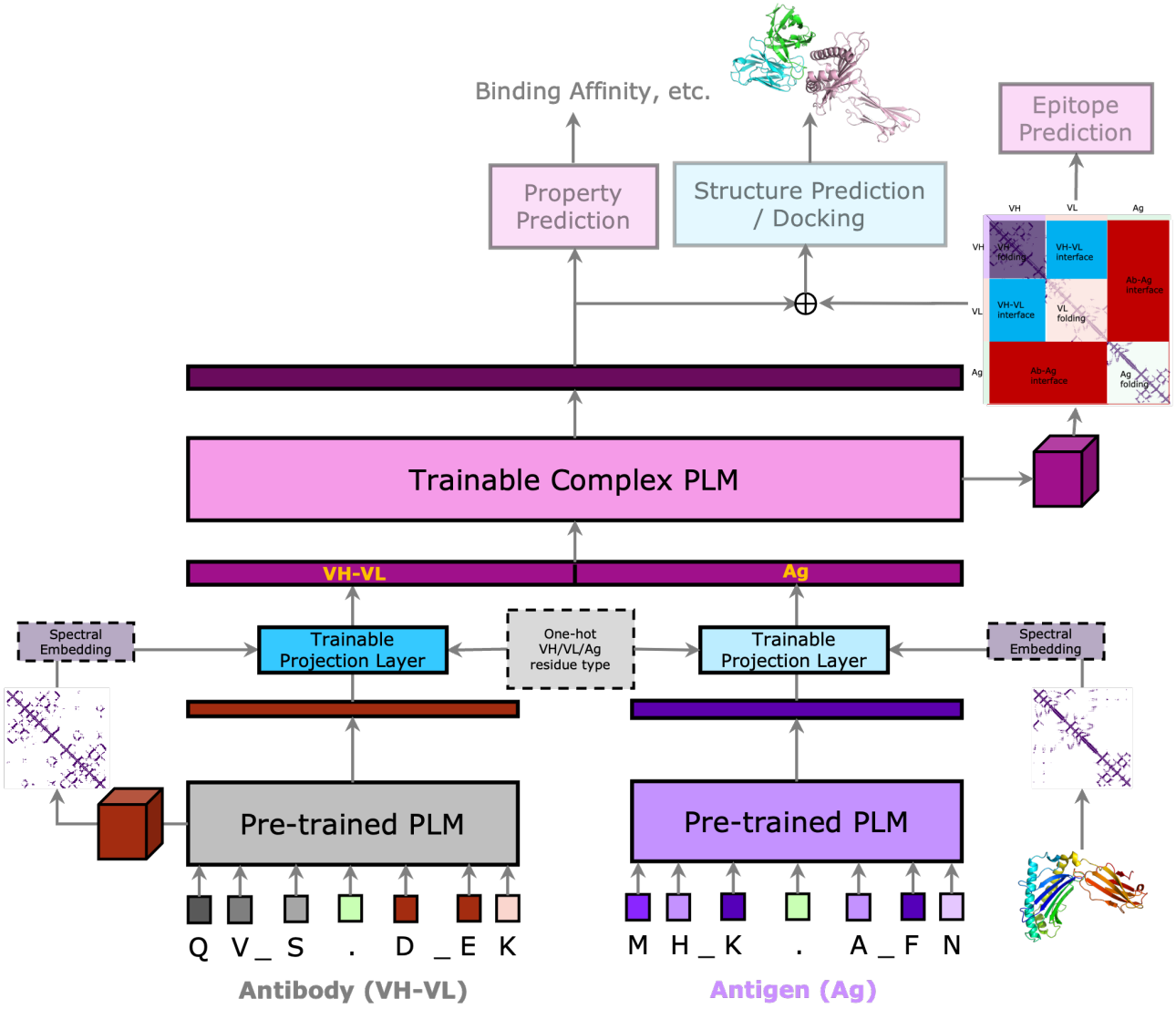

Figure 8. Same as Figure 1