

SNAC-DB: The Hitchhiker's Guide to Building Better Predictive Models of Antibody & NANOBODY® VHH-Antigen Complexes

Abhinav Gupta*, Bryan M. Rivero*, Roel J.T., R. Li, N. Furtmann, Y.F. Nanfack, M. Wendt, Y. Qiu#

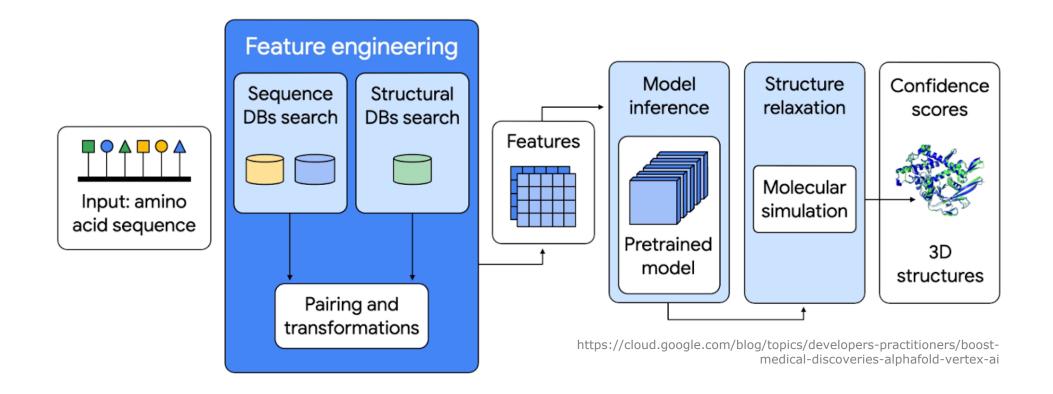
Large Molecule Research, R&D Data & Computational Science



*Equal Contribution #Currently at AstraZeneca

Protein Structure Prediction

AlphaFold-style models are revolutionizing drug-discovery!



Where is the Current Data Sourced

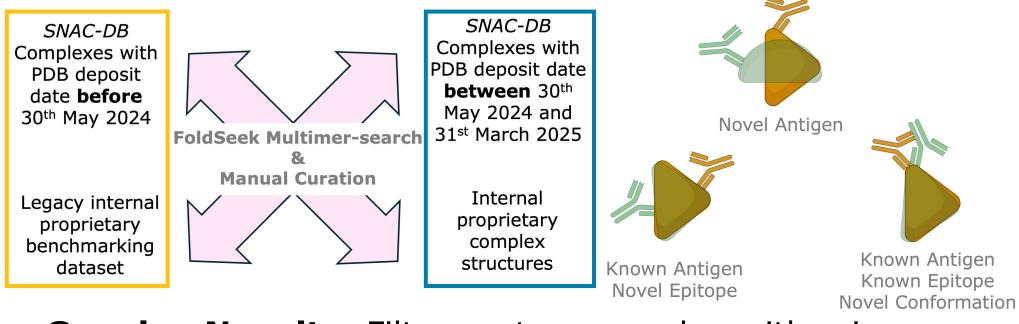
Protein Data Bank¹

- PDB the world's central archive of experimentally determined 3D macromolecular structures (X-ray crystallography, NMR, cryo-EM).
- Pros: Comprehensive coverage.
- **Cons**: Heterogeneous naming, missing residues, crystal-packing artifacts.

SAbDob Structural Antibody Database²

- SAbDab a curated PDB subset of Ab/Nb–Ag complexes from asymmetric units.
- Pros: VH/VL/VHH annotation, standardized variable domains.
- Cons: Issues with the accuracy of multi-chain antigens, excludes TCR and Ig-Ig interactions.

Towards Rigorous Benchmarking



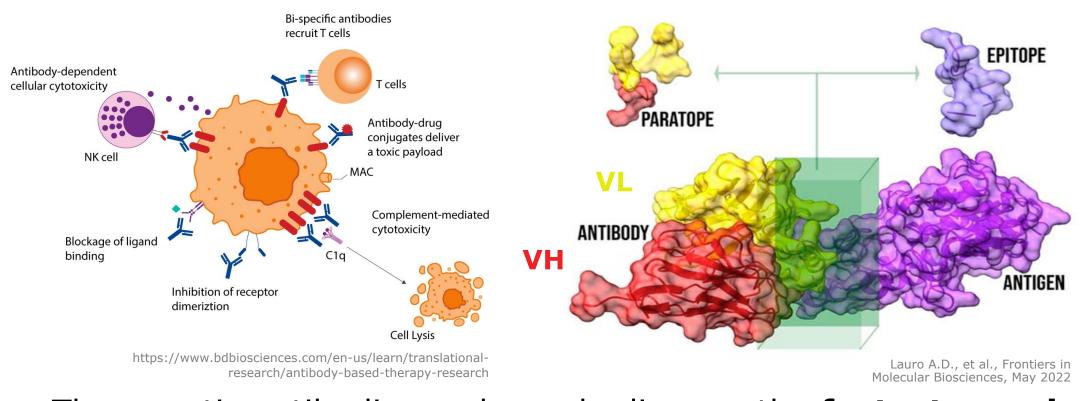
• **Genuine Novelty:** Filters out any overlap with prior training sets using TM-score to ensure evaluation on truly unseen targets, epitopes, and conformations.

https://github.com/Sanofi-Public/SNAC-DB

• Industry-relevant benchmark: Mix of public and proprietary structures.

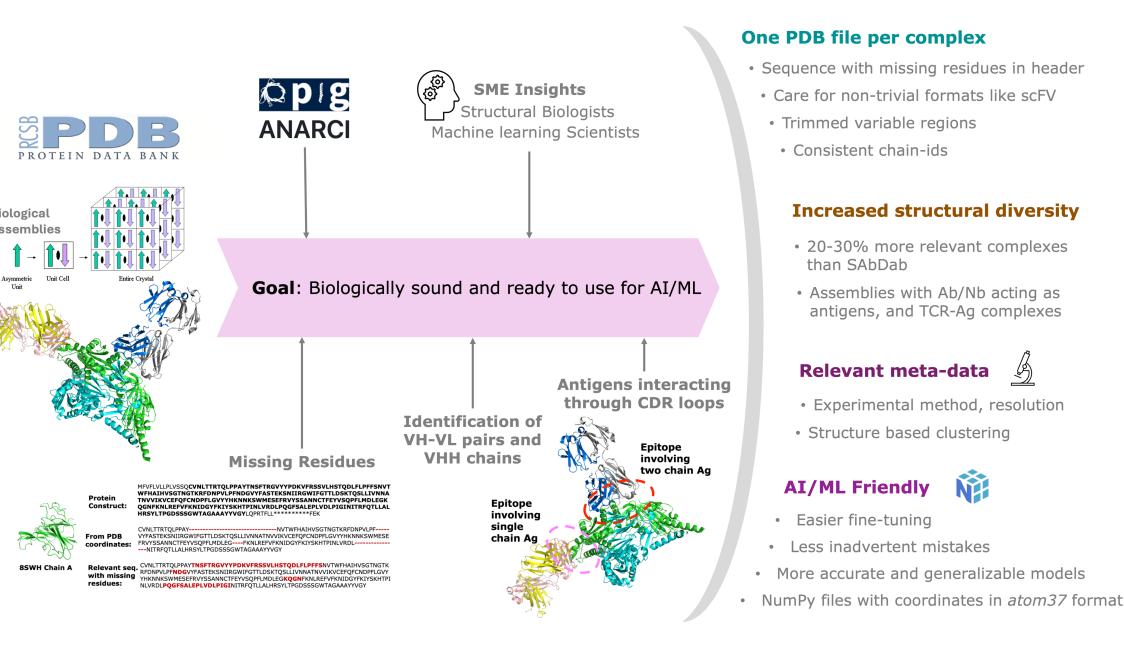
NANOBODY® VHH-Antigen Complexes Antibody-Antigen Complexes Acceptable Acceptable Medium Success Rate (%) Public (29%, N=31) 35%, N=49) **17.3**% 14.2% (71%, N=75) (65%, N=90) ____ - Medium 22.0% : Rate (%) ™ 17.3% 14.0% 14.3% 12.0% 12.5% 12.0% 10.7% 0.65 0.70 Confidence Score ¹Berman, H.M., et al (2000). Nucleic ²Dunbar, J., Krawczyk, K. et al (2014). References and Nucleic Acids Res. 42. D1140-D1146 Acids Res. 42. 28:235-242 **Data Availability:**

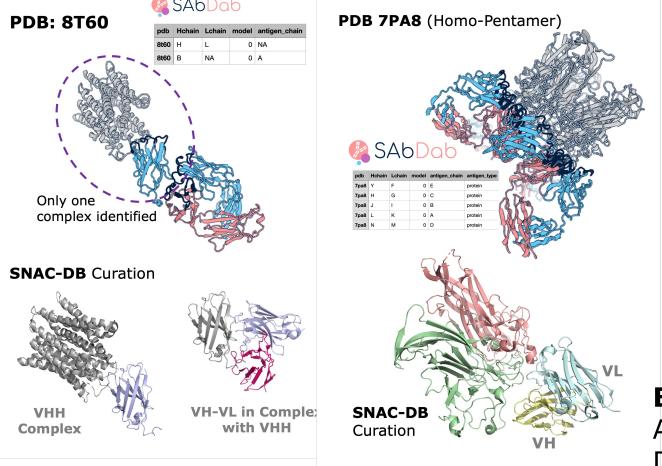
Importance of Antibody and NANOBODY® VHH

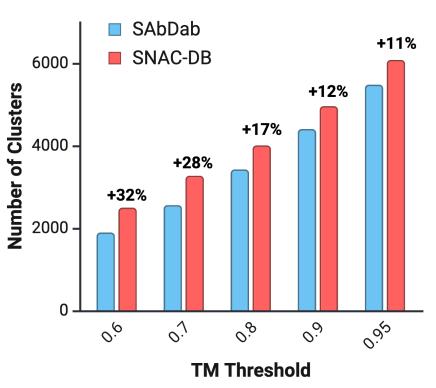


- Therapeutic antibodies and nanobodies are the fastest growing drug class.
- Highly variable and flexible nature of CDR loops make them a blind spot for AI/ML.
- Poor predictions risk wasted time and money in early-stage discovery.

Pipeline Overview







Broader Structural Coverage:
At different TM thresholds, SNAC-DB yields 11–32% more clusters—preserving complexes SAbDab filters out.

Conclusions & Future Directions

- **Data Matters:** Boltz-2's extra training on newer PDB entries yields noticeably better Ab-Ag and Nb-Ag predictions than purely architectural improvements.
- Poor generalization: All models struggle on novel epitopes, revealing a tendency to "remember" familiar binding sites instead of truly extrapolating.
- Unreliable confidence metrics: Even when the correct pose is sampled, internal scoring rarely ranks it first, limiting trust in topranked predictions. Boltz-2's internal confidence scores correlate weakly with actual DockQ quality (Spearman $\rho \approx 0.45$).
- SNAC-DB closes critical gaps. Immediate lift to structural coverage, better multi-chain antigens, missed complexes: TCR-Ag, Ab-Nb, Nb-Ab, weak-cognate interactions.
- Next steps: fine-tune models on SNAC-DB to evaluate how the performance is impacted.

https://zenodo.org/uploads/15870003