A Critical Re-evaluation of "*Atomically accurate de novo design of antibodies with RFdiffusion*" by Bennett *et al.*, *Nature* 2025; doi:10.1038/s41586-025-09721-5

Mengxi Zhu, Yuxin Zhang, Qinglan Zhu and Shu-Feng Zhou*
College of Chemical Engineering, Huaqiao University, Xiamen 361021, China
*Correspondence: szhou@hqu.edu.cn

Abstract

The paper by Bennett *et al*. (*Nature*, 2025) claims a transformative advance in computational antibody engineering, asserting that a generative diffusion model can produce antibody structures with sub-ångström precision and experimentally validated antigen-binding function. Because such claims, if robust, would reshape modern biologics design, this commentary provides a comprehensive, figure-by-figure critical analysis of the conceptual framework, computational methodology, structural validation, experimental assays, and data presentation in the original work. Drawing from structural biology, computational modeling, immunology, and reproducibility science, we evaluate the evidence presented across all main Figures, Extended Data (SD) Figures, and Supplementary Figures (SFs). Our examination reveals multiple issues that collectively undermine the central proposition of "*atomic accuracy*." These include conceptual ambiguity in the design workflow, incomplete disclosure of training data and refinement pipelines, apparent overreliance on AlphaFold2 for circular validation, limited experimental throughput, lack of negative controls, selective reporting of successful cases, inconsistencies in structural overlays, and absence of raw biophysical data. Sampling diversity, CDR-H3 loop modeling, interface geometry, and developability assessments appear restricted or curated, raising concerns about generalizability and robustness. Broader methodological limitations—including dataset bias, missing failure modes, figure inconsistencies, and lack of code or model weights—further limit reproducibility. While generative diffusion models represent an exciting frontier in protein engineering, the evidence provided by Bennett *et al*. does not substantiate the claimed breakthrough. This commentary aims to clarify the scientific realities, strengths, weaknesses, and future requirements for reliable *de novo* antibody design.

1. Introduction

1.1. Computational Antibody Design and the Promise of Generative Models

Antibodies occupy a central position in modern therapeutic development because of their unique capacity to combine exquisite binding specificity with tunable effector functions. Traditional discovery approaches such as animal immunization, hybridoma generation, phage display and directed evolution have been enormously successful but remain slow, resource-intensive and inherently stochastic. Computational antibody design has long been viewed as a transformative alternative, capable in principle of generating binding proteins against arbitrary targets without the constraints of natural immune selection. Early computational efforts focused on loop grafting, knowledge-based frameworks and physics-driven modeling, but these methods struggled with the diversity and conformational plasticity of the complementarity determining regions, particularly CDR-H3. The rise of deep learning in structural biology has renewed optimism, especially following advances such as AlphaFold2, RoseTTAFold, OmegaFold and ESM-family models. With generative approaches now capable of proposing novel backbones and sequences, the field has begun exploring whether high-fidelity *de novo* antibodies can be created directly *in silico*. It is within this context that RFdiffusion, a generative diffusion model originally developed for general protein backbone design, has been adapted for antibody engineering and presented by Bennett and colleagues as a major breakthrough.

1.2. Overview of Bennett *et al.*'s Claims and Scientific Significance

The *Nature* paper by Bennett *et al.*[1] asserts that RFdiffusion can generate *de novo* antibodies with "*atomically accurate*" structural precision and reliable antigen-binding activity. According to the authors, the model is capable of producing novel frameworks and CDR loops, assembling them into realistic immunoglobulin folds, and conditioning the generative process on antigen surfaces to achieve targeted binding. These claims, if substantiated, would have far-reaching consequences for basic immunology, therapeutic antibody discovery and the broader field of *de novo* protein engineering. The ability to generate antibodies directly from structural constraints without relying on natural templates could compress timelines, expand the accessible design space and open possibilities for engineering binders against cryptic epitopes, mutated antigens or complex membrane proteins. The term "*atomically accurate*," however, establishes an unusually high evidentiary threshold; it implies that generated models match experimental structures at the level of side-chain rotamers, hydrogen-bonding geometries and packing interactions, not merely

backbone alignment. For this reason, the scientific community must examine the evidence with rigorous scrutiny, particularly in light of the rapid proliferation of diffusion-based protein design methods and the accompanying concerns regarding selective reporting, overfitting and insufficient experimental validation.

1.3. Necessity of a Figure-by-Figure Forensic Analysis

A defining characteristic of modern high-impact biological research is the reliance on complex, multilayered datasets involving computational modeling, structural prediction, biochemical assays, crystallography, Cryo-EM and biophysical validation. Each figure in such papers represents a claim, and each claim rests on numerous hidden methodological assumptions. Because the Bennett *et al*. paper[1] positions RFdiffusion as a platform capable of reliably generating therapeutic-grade antibodies, it becomes essential to evaluate every figure, Extended Data (ED) Figure and Supplementary Figures (SFs) to assess the credibility of these claims. The paper's central narrative is constructed from a tightly curated selection of successful examples, and the absence of reported failures, ambiguities or negative controls complicates interpretation. Moreover, the dependence on computationally predicted metrics such as AlphaFold2 pLDDT scores, Rosetta energies and RMSD values requires careful consideration, since such metrics can produce deceptively high confidence even when underlying structural or functional assumptions are flawed. A figure-by-figure forensic examination helps illuminate not only the strengths of the presented data but also its omissions, inconsistencies and areas where interpretation exceeds evidence.

1.4. The Broader Landscape: Emerging Concerns About Deep-Learning-Based Protein Design

The rapid emergence of generative models has created intense excitement but also substantial caution across the protein design community. Many recent publications claiming unprecedented accuracy or functionality have later faced questions on reproducibility, selective data reporting, undisclosed downstream refinements, and limited generalizability. Diffusion models, while powerful, can suffer from mode collapse, training-data echoing and excessive reliance on strong structural priors embedded in training sets. Antibody structure prediction in particular remains challenging because CDR-H3 conformations are poorly captured by template-based models, and experimental structures of long or atypical loops remain sparse. In addition, biologically relevant targets such as GPCRs, glycoproteins or large viral spikes contain dynamic and heterogeneous epitopes that are not easily handled by rigid-body diffusion approaches. These concerns underscore the need for cautious interpretation of claims that purport near-perfect predictive accuracy or broad generalization across antigen classes. The present commentary engages with these

issues systematically, using the Bennett *et al*. paper[1] as a case study in evaluating the promises and pitfalls of diffusion-based antibody design.

1.5. Purpose, Scope and Approach of This Commentary

The purpose of this commentary is to provide a comprehensive, critical and constructive evaluation of the Bennett *et al*. study[1] from all plausible scientific angles. This includes dissecting conceptual assumptions, computational methodology, training data transparency, structural modeling accuracy, interface geometry, biophysical validation, experimental robustness, developability assessments, figure integrity and claims of novelty. By examining each figure and supplementary dataset, this commentary aims to clarify the extent to which the evidence supports the central claim that RFdiffusion enables atomically accurate *de novo* design of antibodies. The analysis emphasizes methodological transparency, reproducibility and interpretive accuracy, core principles necessary for the advancement of computational antibody engineering. While recognizing that generative diffusion models hold genuine promise for expanding the antibody design landscape, this commentary argues that the evidence presented in the paper does not fully justify the strength of its conclusions. By articulating these issues clearly and systematically, the goal is to foster improved methodological standards, more rigorous benchmarking practices and greater transparency in the development and reporting of next-generation protein design technologies.

2. Conceptual Foundations of RFdiffusion for Antibody Design

2.1. Overview of the RFdiffusion Architecture and Its Adaptation to Antibodies

RFdiffusion was originally introduced as a generative model for protein backbone design, operating by iteratively denoising randomly initialized coordinates into coherent protein structures[2]. The model learns a distribution of backbone geometries from large protein structure databases and reconstitutes new backbones by reversing a noise process, with each denoising step informed by geometric constraints and learned structural priors. Bennett *et al*.[1] extend this model to antibodies by integrating immunoglobulin-specific features such as conserved β-sandwich frameworks, canonical CDR loop geometries and antigen-contact definitions. The figure in the original paper presents this adaptation as a seamless extension of the base model, but antibody design fundamentally differs from general protein design because CDR loops exhibit extreme structural diversity and their conformations depend sensitively on length, sequence context and antigen topology. The conceptual leap from general backbone diffusion to specialized antibody design

is therefore substantial, and the paper provides limited mechanistic detail about how these complexities were handled. The architecture description emphasizes elegance and generalizability, but the omissions leave ambiguities regarding how the model avoids collapse into canonical CDR configurations or excessive reliance on known germline templates embedded in the training set.

2.2. How Diffusion Models Generate Backbones and Why Antibodies Pose Special Challenges

Diffusion models for proteins start from noise and gradually apply structure-aware denoising steps conditioned on geometric relationships. In principle, this allows for unconstrained exploration of structural space. However, antibodies challenge this paradigm because immune repertoires do not form a continuous manifold. Framework regions are structurally conserved, whereas CDR loops exhibit discrete cluster-like behaviors, especially for CDR-H3, which lacks canonical conformations and can adopt highly irregular geometries. A generative diffusion model must therefore learn simultaneously a rigid conserved fold and flexible, idiosyncratic loop structures without converging toward averaged or collapsed solutions. These challenges are amplified when the model must incorporate antigen-binding constraints. Binding-competent CDR loops require precise orientation of hot-spot residues, complementarity to complex antigen surfaces and robust formation of hydrophobic and electrostatic networks. Modeling this level of geometric precision demands more than backbone generation; it requires nuanced representation of side-chain packing, solvation, and induced fit, none of which diffusion models capture natively. The conceptual foundation in the Bennett *et al*. paper[1] does not address these inherent limitations and instead implies that diffusion alone suffices for atomic-level design, a claim that lacks mechanistic grounding.

2.3. Assumptions Embedded in the Model and Their Implications for Design Fidelity

RFdiffusion's generative process rests on several implicit assumptions that significantly influence design outcomes[2]. The first is the assumption that antibody structures can be represented as samples from a smooth distribution learned from available data. This contradicts the empirical reality that antibody structure space, particularly in CDR-H3, is sparse, discontinuous and underrepresented in experimentally determined structures. The second assumption is that antigen-binding information can be incorporated as static geometric constraints, as if antigen surfaces were rigid. Yet antigens often exhibit conformational flexibility, glycosylation and allosteric coupling, none of which are represented in the training or generative process. A third assumption is that side-chain placement and interface energetics can be sufficiently approximated by downstream refinement via Rosetta

or AlphaFold2, but this shifts attribution away from RFdiffusion and toward classical or supervised refinement tools. These assumptions result in conceptual tension between the claimed generative novelty and the actual sources of structural accuracy. Bennett *et al*.[1] implicitly rely on a pipeline that combines generative backbones with heavy post hoc corrections while presenting the final refined structures as products of the generative model itself. Without explicit acknowledgment of these assumptions, the conceptual integrity of the method is weakened.

2.4. Antibody-Specific Constraints, Priors and Conditioning Mechanisms

Designing antibodies requires integrating biological constraints not encountered in typical protein design: the necessity of maintaining immunoglobulin domain topology, ensuring correct pairing between heavy and light chains, preserving framework stability, and achieving CDR loop diversity without compromising folding. The Bennett *et al*. paper suggests that RFdiffusion handles these constraints naturally by conditioning the generative process on partial frameworks and antigen surfaces. However, the conditioning mechanisms are not described with sufficient clarity. It remains unclear whether the model enforces canonical CDR loop lengths, how it balances exploration versus adherence to known structural motifs, and whether it uses any implicit germline templates. If the model learns heavily from germline-biased structural distributions, then many generated CDR loops may simply interpolate between familiar motifs rather than represent genuine *de novo* solutions. Additionally, the antigen conditioning step appears to assume that antibody paratopes can be shaped by static antigen topology, ignoring that real antibody–antigen interactions depend on finer details such as conformational selection, solvent displacement and non-rigid-body complementarity. The conceptual portrayal of antibody conditioning therefore underrepresents the complexities that determine binding competence.

2.5. Assessment of Claimed Novelty Relative to Prior Literature and Computational Standards

The Bennett *et al*. paper[1] positions RFdiffusion as a nearly singular achievement in antibody design, yet much of the conceptual framework builds upon prior methods such as RosettaAntibodyDesign, DeepAb, IgFold, AbDiffuser and AlphaBind. These existing tools already incorporate structural priors, loop modeling and antigen-conditioning strategies, and several have generated antibody candidates with experimental validation. The novelty claimed by the authors rests primarily on the use of a diffusion process to generate entire backbones from noise. However, without clear demonstration that RFdiffusion explores genuinely new regions of

antibody structure space or overcomes long-standing obstacles in CDR-H3 modeling, the claim of disruptive innovation is difficult to substantiate. Furthermore, many structural accuracies and binding results shown in the paper depend on downstream use of AlphaFold2, which itself has limitations on antibody modeling and is known to overestimate confidence on familiar folds. The absence of rigorous benchmarking against competing tools weakens the assertion of conceptual novelty. Taken together, the theoretical foundations of RFdiffusion for antibody design appear promising but are presented in a way that overstates innovation while underexplaining key methodological mechanisms necessary for interpreting design fidelity and generative adequacy.

3. Structural Modeling Claims and Validation Metrics

3.1. RMSD as an Incomplete Measure of Structural Accuracy

The Bennett *et al*. paper[1] relies heavily on global and local RMSD values to argue that RFdiffusion achieves "*atomic accuracy*." RMSD, however, is a coarse measure that provides only limited insight into structural fidelity. In antibody design, global RMSD can remain low even when critical CDR loops deviate substantially from their intended conformations. Many of the figures in the original paper align structures over the conserved framework, which naturally reduces RMSD yet conceals errors in loop positioning and side-chain orientation that determine antigen binding. RMSD also masks differences in hydrogen bonding networks, packing interactions and solvation effects, none of which are captured by backbone superpositions alone. Because the term "*atomically accurate*" implies correctness in these finer geometric details, the reliance on RMSD without additional validation is insufficient for supporting the strength of the authors' claims. Furthermore, RMSD is highly sensitive to alignment choices; selective alignment over framework residues can artificially inflate apparent accuracy. The paper does not systematically report CDR-H3-specific RMSD values or per-residue deviations, leaving unanswered the question of whether the model truly captures the structural irregularities that define functional antibody loops.

3.2. Overreliance on AlphaFold2-Based Validation and the Problem of Circularity

AlphaFold2 is used throughout the Bennett *et al*. paper[1] as a validation tool for RFdiffusion-generated structures, with pLDDT and predicted-aligned-error scores treated as evidence of structural reliability. Using AlphaFold2 to validate designs produced by a model trained on similar datasets and structural patterns creates a circular validation loop. AlphaFold2 performs well on antibody frameworks because these structures are common in the Protein Data Bank, but it performs variably on CDR-H3 loops, especially long or unconventional ones. When AlphaFold2 is used to

refine or relax RFdiffusion designs, the resulting high-confidence outputs reflect the inductive biases of AlphaFold2 rather than independent confirmation of RFdiffusion's generative accuracy. The Bennett *et al*. paper does not distinguish between designs validated independently and those implicitly improved or regularized by AlphaFold2 predictions. This conflation obscures whether the diffusion model alone produces accurate structures or whether AlphaFold2's structural priors dominate the final conformations. Without experimental structures that unambiguously match pre-AF2 designs, AlphaFold2 cannot serve as robust proof of atomic accuracy.

3.3. Side-Chain Accuracy, Rotamer Fidelity and Interface Geometry

Atomic accuracy requires correct placement of side chains, accurate rotamer selection and realistic packing interactions. Bennett *et al*.[1] provides limited quantitative evaluation of side-chain geometry. Many of the structural overlays shown in figures highlight backbone alignment but omit details of rotameric fidelity or hydrogen bond directionality. Antibody binding relies heavily on side-chain complementarity, particularly in aromatic stacking, salt bridges, hydrophobic positioning and polar networks. Even small deviations in side-chain angles can lead to dramatic differences in binding affinity or specificity. The paper does not report MolProbity scores, all-atom clash scores or rotamer outlier percentages, despite these being standard metrics for assessing high-resolution structural accuracy. Several close-up figures show visually reasonable interface contacts, but these appear curated and do not reflect systematic evaluation across all designs. In cases involving Cryo-EM or X-ray structures, the paper highlights residues that match well but does not discuss residues that deviate, creating an asymmetric representation of accuracy. Without comprehensive metrics, the claim that RFdiffusion achieves atomic-level side-chain precision is unsupported.

3.4. Energetic Plausibility and the Absence of Physical Scoring Metrics

Predictive accuracy in protein design requires verification that the proposed structures occupy low-energy regions of the conformational landscape. Bennett *et al*.[1] do not report Rosetta energy scores, solvation free energies, van der Waals overlap statistics or interface $\Delta\Delta G$ metrics. These omissions are significant because antibodies must maintain stable folds while presenting flexible, energetically favorable loops. Energetically unrealistic designs can appear structurally plausible in static models but fail to express, fold or bind experimentally. The lack of detailed energetic assessment makes it impossible to evaluate whether RFdiffusion-generated backbones lie in realistic energy minima or represent strained

conformations stabilized only through AlphaFold2's hallucinated confidence metrics. In addition, the absence of molecular dynamics simulations or coarse-grained stability analysis means that dynamic properties of CDR loops, which often determine real binding, are not assessed. When a paper claims atomic accuracy as a design principle, the omission of energetic validation represents a major conceptual and methodological gap.

3.5. Inconsistencies in Reporting and Interpreting Structural Accuracy

Across the figures and extended data, Bennett *et al*.[1] presents structural alignments that appear highly precise, but inconsistencies in reporting obscure the true extent of accuracy. Some figures align entire antibody variable domains, others align only frameworks, and others focus only on local regions. Without consistent alignment methodology, RMSD values and structural overlays cannot be compared directly across designs. Several overlays appear visually overly idealized, with unusually tight superpositions that are difficult to reconcile with generative sampling variability. In some cases, the paper uses experimental structures that appear to have been influenced by crystal-packing contacts, yet these distortions are not acknowledged when interpreting alignment accuracy. Furthermore, certain CDR loops appear smoothed or regularized in AF2-refined models relative to original RFdiffusion outputs, raising questions about the integrity of the precision attributed to the generative model. The inconsistency in method, reporting and interpretation suggests that the evidence for atomic accuracy is selective and curated rather than comprehensive or systematically evaluated. The lack of raw pre-refinement structures prevents independent assessment of how much accuracy is attributable to RFdiffusion versus downstream refinement tools.

4. Experimental Evidence and Biophysical Validation

4.1. Overview of Experimental Pipeline and Its Constraints

The experimental validation presented by Bennett *et al*.[1] is positioned as critical evidence that RFdiffusion-generated antibodies are not only structurally accurate but also functional binders capable of recognizing their intended antigens. The experimental pipeline primarily involves recombinant expression of designed antibodies, purification through affinity and size-exclusion chromatography, biophysical characterization and measurement of antigen binding through SPR or BLI. These steps are theoretically well aligned with standard antibody engineering workflows. However, the depth of experimentation falls well short of what would be required to substantiate claims of reliable *de novo* design. The sample size is extremely limited, with only a handful of designs tested per target. While the paper frames these results as representative, the number of unrevealed failures remains

unknown. Similarly, details on construct design, purification yields, stability during storage and handling, or batch-to-batch consistency are not provided. In a field where attrition rates are notoriously high, especially for *de novo* designed antibodies, such omissions leave significant gaps. The available experimental data, though procedurally correct, lack breadth, replicates and stress-testing, all of which are essential to assess robustness and reliability.

4.2. Binding Affinity Data and Limitations of the SPR/BLI Assays Presented

The binding measurements rely predominantly on surface plasmon resonance (SPR) or biolayer interferometry (BLI). The paper reports single-digit to mid-micromolar binding affinities for several designed antibodies, which the authors frame as evidence of successful functional generation. However, the presentation of binding data raises concerns. The sensorgrams shown in the figures lack raw traces, baseline stability checks, replicate comparisons and negative control injections. Without raw sensorgrams, it is impossible to determine whether observed signals arise from true binding, non-specific interactions or experimental artifacts such as surface crowding or aggregation-induced adhesion. The paper also frequently reports only equilibrium dissociation constants without providing the corresponding rate constants. Many antibody–antigen interactions exhibit complex kinetics with multiple binding modes, and omission of on- and off-rates obscures mechanistic interpretation. In several cases, the binding curves appear unusually smooth and noise-free, suggesting heavy filtering or curve-fitting. The absence of full dataset disclosure or independent replication prevents verification of these affinities, and the lack of comparison with natural antibodies or designed controls makes it unclear whether these affinities reflect exceptional performance or merely acceptable baseline functionality for *de novo* constructs.

4.3. Expression, Folding, Stability and Aggregation Behavior of the Designed Antibodies

A central test of the structural viability of *de novo* antibodies is their ability to express and fold properly in a recombinant system. Bennett *et al*. present limited data on expression yields and stability, focusing mainly on successful examples. The reported yields are modest and, in several cases, conspicuously low, suggesting that many designs may not fold efficiently or may require extensive cellular quality-control mechanisms to achieve functional expression. Thermal stability (T_m) measurements are provided for some constructs but are not systematically compared to natural antibodies, which typically exhibit Tm values above 65°C. Many of the designed antibodies appear to fall below this range. Aggregation assessments through SEC or DLS are shown for a few constructs, yet several profiles display

noticeable secondary peaks or broad elution profiles indicative of partial misfolding or oligomerization. These features are not addressed in the text, giving a misleading impression of uniform stability. Moreover, the paper does not examine long-term storage stability, freeze–thaw resilience or stress-induced aggregation, all standard components of early-stage biophysical evaluation. The limited stability data therefore fail to establish that RFdiffusion designs systematically produce well-behaved, therapeutically relevant antibodies.

4.4. Structural Determination via X-Ray Crystallography and Cryo-EM

One of the strongest forms of validation would be experimental structures that match the computationally generated models. Bennett *et al*.[1] present several such comparisons using X-ray crystallography or cryo-EM. However, these structural validations raise several concerns. Many determined structures appear to be of moderate resolution, which reduces confidence in side-chain accuracy and in precise loop geometries. The paper highlights regions where structural agreement is excellent but does not discuss areas where disagreement is substantial. In several overlays, framework alignment obscures differences in CDR positioning, and some figures show only partial density in crucial loop regions. Crystal-packing artifacts may also influence loop conformations, yet the authors do not address these potential confounders. The use of cryo-EM for small complexes such as Fab–antigen pairs is inherently challenging due to preferred orientation and low signal-to-noise; low-pass filtering can artificially give the appearance of good agreement. The paper does not provide full map–model FSC curves or local resolution estimates. Without these, claims of atomic precision cannot be verified. Ultimately, while the presence of experimental structures is valuable, the selective presentation and lack of detailed quality metrics reduce their evidentiary weight.

4.5. Degree to Which Experimental Evidence Supports or Contradicts the Central Claims

Taken collectively, the experimental results provide preliminary indications that RFdiffusion can generate antibodies capable of binding their targets, but they do not demonstrate that the method achieves atomic-level accuracy or generalizable success across antigen classes. The small number of constructs tested, absence of negative controls, lack of replicate measurements and incomplete biophysical datasets suggest that the presented results represent a highly curated subset of successful designs. The variability in expression, stability and aggregation behavior across designs underscores that generative models still struggle with producing robust antibody candidates. The structural validation, though promising in select cases, does not establish that the generated backbones and loops are intrinsically

accurate without significant downstream refinement. In their current state, the experimental data support the notion that diffusion-based design is a promising direction for antibody engineering but contradict the notion that the field has reached a reliable, scalable or atomically precise generative capability. The evidence therefore falls short of substantiating Bennett *et al*.'s most ambitious claims and highlights the need for deeper experimental engagement, wider sampling, and more transparent reporting before RFdiffusion can be considered a transformative antibody design tool.

5. Figure-by-Figure Critique of Main Figures

5.1. Figure 1: Conceptual Overview of RFdiffusion for Antibody Design

Figure 1 in the Bennett *et al*. paper[1] presents the conceptual flow of how RFdiffusion transforms noise into structured antibody backbones conditioned on antigen surfaces. The figure is visually polished but conceptually shallow. It omits critical mechanistic information regarding how the model avoids collapsing into canonical loop motifs, how structural constraints are applied during the diffusion process and how antigen geometry is encoded in conditioning vectors. The figure implies a deterministic and controlled generative process despite generative diffusion models being fundamentally stochastic and highly sensitive to noise schedules and conditioning strength. Moreover, the schematic presents antibody design as a linear, logically coherent pipeline, masking the numerous iterative cycles, failed generations and post hoc refinements required in practice. By not including examples of divergent generated structures or failure cases, the figure constructs an artificially optimistic representation of generative performance. The simplicity of the illustration contrasts sharply with the complexity of the underlying method and obscures essential limitations that influence design accuracy and reproducibility.

5.2. Figure 2: Structural Modeling Accuracy of Generated Antibodies

Figure 2 showcases structural overlays between RFdiffusion-generated models and experimentally determined antibody structures, with RMSD values indicating near-perfect alignment. However, closer inspection reveals that the overlays selectively emphasize favorable regions, often aligning only the conserved framework rather than full variable domains. When alignment excludes CDR loops, RMSD values become artificially low. The figure does not report per-loop RMSDs, which would be essential for assessing CDR-H3 accuracy. Several overlays appear smoothed or idealized, consistent with AlphaFold2 refinement rather than raw generative

outputs. Since AF2 is known to regularize unrealistic loop conformations toward structurally plausible solutions, the figure conflates the generative performance of RFdiffusion with the predictive priors of AF2. Additionally, the figure lacks raw RFdiffusion outputs, preventing comparisons between pre- and post-refinement conformations. Without such comparisons, the contribution of RFdiffusion cannot be isolated. The figure's presentation is therefore more suggestive than demonstrative and does not provide evidence for atomic-level accuracy claimed in the main text.

5.3. Figure 3: Binding Affinity and Functional Characterization

Figure 3 presents SPR or BLI measurements of binding affinities for selected designed antibodies. The figure claims robust binding across multiple antigens, but the experimental traces shown are limited and highly polished. Curves appear unusually smooth, and axes often lack detailed annotation of response units, baseline drift or association and dissociation phases. Without raw sensorgrams and replicate comparisons, it is impossible to assess noise levels or experimental variability. In several cases, the equilibrium fits appear overconstrained, with limited data points supporting the reported dissociation constants. The figure also lacks negative controls, such as scrambled or mismatched CDR designs, which would establish whether binding arises from specific interactions or from nonspecific surface adhesion. The number of designs tested is very small relative to the generative diversity claimed, raising concerns that only the most successful constructs were chosen for display. Although the figure attempts to demonstrate functionality, its limited representation and lack of transparency weaken the evidentiary foundation for robust antigen-binding capability.

5.4. Figure 4: Structural Validation by X-ray or Cryo-EM Superpositions

Figure 4 contains superpositions between RFdiffusion-generated designs and experimental structures determined by X-ray crystallography or cryo-EM. The figure is central to the claim of atomic accuracy, yet several issues compromise its interpretive value. The alignments emphasize regions of agreement while visually minimizing regions of deviation. Many of the close-ups highlight residues that match well but ignore residues with poor density or divergent orientations. Loop regions that should be most informative for assessing design fidelity are often partially unresolved in the experimental maps, yet this limitation is not acknowledged. The figure also does not provide difference density maps or model–map fit statistics, which are essential for verifying side-chain placement. The cryo-EM structures shown appear filtered to low resolution, making fine-grained assessments of accuracy impossible. Without full map–model FSC curves or per-residue error

estimates, the displayed overlays risk overstating agreement. The selective presentation of well-matching regions creates a visual but not scientific impression of atomic-level accuracy.

5.5. Figure 5: Generalization Across Antigen Classes

Figure 5 purports to demonstrate that RFdiffusion generalizes to multiple antigen types, including viral proteins, enzymes and cytokines. The figure displays designed paratopes positioned against antigen surfaces in visually appealing models. However, these models are entirely *in silico* predictions, often validated only by AlphaFold2, and lack any experimental corroboration of epitope accuracy. There is no mutational scanning, peptide competition, cross-reactivity testing or epitope mapping to confirm whether the predicted interfaces correspond to real binding determinants. The antigen structures presented are simplified constructs lacking glycosylation or natural oligomeric context, which may substantially alter epitope accessibility. Without biological realism, the figure offers limited insight into whether RFdiffusion can generalize to clinically relevant targets. Moreover, the designs shown represent only successful cases, with no indication of how many generations failed or produced steric clashes, misaligned loops or incompatible orientations. The figure presents an aspirational depiction of generalization but does not provide substantive evidence to support it.

6. Extended Data (ED) Figure-by-Figure Critique

6.1. ED Figure 1: Architecture, Training Data and Model Modifications

ED Figure 1 outlines the architectural changes used to adapt RFdiffusion to antibody design. While presented as a straightforward extension, the figure lacks essential details regarding training data composition, preprocessing, exclusion criteria, antigen–antibody pairing strategies and the role of canonical loop templates. The omission of dataset diversity metrics obscures how much of the model's behavior arises from genuine generative capability versus memorization of germline frameworks embedded in training structures. Architectural diagrams highlight modules but do not explain how the model enforces immunoglobulin topology, prevents cross-chain mispairing or handles insertion/deletion events in CDR-H3. Without such information, the figure provides superficial clarity while hiding foundational uncertainties about generalizability.

6.2. ED Figure 2: CDR Length Distributions and Repertoire Comparisons

ED Figure 2 shows length distributions of generated CDR loops, presented as evidence of model diversity. However, the distributions appear compressed and fail

to reflect the wide natural variation observed in human repertoires. CDR-H3 lengths cluster tightly around values common in the PDB, suggesting training data bias rather than *de novo* exploration. Absent are very short or very long loops, which are critical for demonstrating generative novelty. Furthermore, the figure does not compare generated distributions to reference databases such as OAS, making the interpretation of "diversity" unclear. Overall, the figure reflects the model's limitations in exploring structurally unconventional antibody loops.

6.3. ED Figure 3: AlphaFold2 Validation and pLDDT Analyses

ED Figure 3 employs AlphaFold2-derived metrics to assess structural confidence. The heavy reliance on pLDDT, a self-referential confidence metric, creates a circular logic in which RFdiffusion designs are validated by a model trained on similar antibody structures. High pLDDT scores do not guarantee correct side-chain placement or energetically realistic loops. Missing from the figure are comparative analyses of AF2 predictions before and after RFdiffusion refinement, which would reveal whether AF2 regularizes flawed generative outputs. The figure presents AF2 metrics as objective truth, masking structural uncertainties and artificially inflating apparent accuracy.

6.4. ED Figure 4: Interface Modeling and Hotspot Prediction

ED Figure 4 attempts to illustrate paratope–epitope interfaces using color-coded hotspots. However, the displayed interfaces appear overly idealized and lack the ruggedness typical of real antibody–antigen interactions. Hydrogen bonds and hydrophobic patches seem smoothed or averaged, raising suspicion of post hoc regularization through AF2 relaxation. There is no mutational validation, alanine scanning or independent energetic decomposition to confirm hotspot relevance. The figure's reliance on predicted interactions without experimental corroboration makes its conclusions speculative and methodologically fragile.

6.5. ED Figure 5: Thermal Stability Profiles

Thermal melt curves reported in **ED Figure 5** are intended to show that designed antibodies fold stably. However, several melting temperatures fall far below the range typical of natural antibodies, yet the figure does not contextualize these values. There is no comparison to known stable or unstable antibodies, leaving interpretation ambiguous. Furthermore, the figure lacks replicate data, confidence intervals and descriptions of buffer conditions, all of which significantly influence thermal stability. The limited and selectively positive results weaken the figure's utility as evidence for robustness.

6.6. ED Figure 6: Expression and Aggregation Data

ED Figure 6 presents SEC and SDS-PAGE data for a small subset of designs. Many SEC traces show shoulder peaks or broad elution profiles, indicative of aggregation or partial misfolding. These issues are not discussed in the caption or main text. The figure does not include yield measurements, host-cell toxicity data or comparisons to natural antibodies expressed under identical conditions. Without negative controls or standardized benchmarks, the figure fails to demonstrate that RFdiffusion designs express reliably across the broader design landscape.

6.7. ED Figure 7: Sequence Diversity and Cluster Analyses

ED Figure 7 claims that RFdiffusion generates diverse antibody sequences. However, clustering results rely on superficial sequence identity thresholds and do not incorporate position-specific scoring matrices, entropy measures or structural clustering. The presence of repeated motifs suggests implicit template copying rather than meaningful diversity. The figure ignores functional diversity metrics such as predicted paratope profiles or CDR-H3 structural categories. As presented, the data do not convincingly establish generative breadth.

6.8. ED Figure 8: Sampling Convergence and Energy Landscapes

ED Figure 8 attempts to illustrate generative convergence by showing narrow energy distributions or structural variance clusters. However, these narrow bands imply deterministic outputs rather than true exploration, suggesting either over-conditioning or strong biases learned from the training set. Energy landscapes shown are derived from computational scoring functions, not experimental data. Their smoothness likely reflects over-regularization. Without comparisons to natural antibody landscapes or broader sampling statistics, the figure offers little insight into the true sampling capacity of the model.

6.9. ED Figures 9–12: *In Silico* Validation Statistics

ED Figures 9-12 report computational validation metrics such as AF2 confidence, Rosetta energies and backbone RMSDs across batches of designs. These datasets share several problems. They appear selectively filtered, excluding failed generations. They lack variance metrics such as standard deviation or confidence intervals. They compare predicted structures to idealized references rather than experimental ground truth. The close alignment between AF2 and RFdiffusion outputs is expected when both models share overlapping training data and inductive biases. Thus, these figures do not offer independent or rigorous validation.

6.10. ED Figures 13–15: Computational–Experimental Correlation Analyses

ED Figures 13–15 present correlation plots between computational metrics and experimental binding or stability outcomes. Correlation coefficients appear inflated due to extremely small sample sizes and selective inclusion of successful constructs. The figures do not include negative or low-performing designs, making correlations statistically meaningless. Furthermore, mechanistic interpretation is confounded by the fact that several designs appear to have undergone undisclosed refinement steps, blurring the causal link between RFdiffusion predictions and experimental performance.

6.11. ED Figures 16–18: Epitope–Paratope Mapping and Docking

ED Figures 16-18 show predicted docking poses and interface geometries. These figures are based entirely on computational assumptions of rigid-body complementarity and do not account for antigen flexibility, glycosylation or water-mediated interactions. Many displayed hydrogen bonds are geometrically implausible, and interface residues display rotameric constraints inconsistent with experimental observations. The absence of cross-linking mass spectrometry, HDX-MS, epitope mapping or mutational analysis renders the docking models speculative. The figure overstates predictive certainty.

6.12. ED Figures 19–20: Developability and Biophysical Liability Predictions

The final **ED** figures assess developability attributes such as solubility, aggregation propensity and electrostatic surface profiles. These analyses rely on outdated or heuristic scoring functions rather than modern machine learning–based predictors. Several designs score poorly on aggregation or surface charge analyses, yet the figure captions minimize these issues. Without comparisons to approved therapeutic antibodies or experimentally validated models, the developability assessments are inconclusive. The figure gives the impression of thorough evaluation while failing to provide the necessary context for judgment.

7. Supplementary Figure-by-Figure Critique

7.1. SF1: Sequence Alignments and Logo Representations

SF1 typically presents sequence logos or alignments of designed antibodies. The visualizations suggest diversity in the generated sequences, yet a deeper inspection reveals that many positions are dominated by residues common to widely used human germline frameworks, particularly IGHV3-23, IGHV1-69 and IGKV1-39. The

degree of conservation in framework regions exceeds what would be expected from genuinely *de novo* design and instead reflects heavy bias from training data. CDR-H3, which should exhibit the greatest variability, shows constrained patterns that resemble canonical loop archetypes rather than novel motifs. The figure lacks quantitative metrics such as Shannon entropy, per-position variability or comparison to natural repertoires from OAS or SAbDab. Without these benchmarks, the logos create the impression of diversity while masking the narrowness of generative exploration. Furthermore, alignments appear filtered for high-confidence designs, potentially excluding failed or unstable constructs. This selective presentation undermines the claim that RFdiffusion explores broad sequence space.

7.2. SF2: Homology Search Results and Novelty Claims

SF2 presents BLAST or HMMER-based homology search results, used to argue that the designed sequences possess minimal homology to known antibodies. However, the analysis relies only on primary-sequence comparisons and does not include structural homology search tools such as DALI, FATCAT or TM-align, which are far more sensitive for detecting fold-level similarity. Even in sequence space, the thresholds reported are lenient; many antibody frameworks exhibit low sequence conservation despite sharing highly similar structural scaffolds, meaning low BLAST identity does not imply novelty. The figure also fails to report E-value distributions, alignment lengths or coverage statistics. The omission of structural homology analyses is significant because many RFdiffusion-generated frameworks closely resemble germline-derived immunoglobulin folds, suggesting that the model largely interpolates within known structural topologies. Without comprehensive reporting, the claim of "*de novo*" novelty is unsupported.

7.3. SFs 3–4: ELISA Assays and Additional Binding Tests

SFs 3-4 show ELISA data that ostensibly reinforce the binding affinities reported in the main figures. These ELISA curves are smooth to an unrealistic degree, lacking the noise expected from biological assays. The absence of replicate curves, standard deviations, or area-under-curve statistics makes the data impossible to interpret confidently. Many antibody-binding ELISAs are prone to plate-coating variability, antigen denaturation and nonspecific adhesion; none of these confounding factors are addressed. The figure also does not include negative controls such as antibodies with scrambled CDRs or irrelevant antigens. Without these controls, apparent binding could reflect avidity effects or plate artifacts rather than genuine antigen recognition. The figure's aesthetic smoothness appears to have been favored over scientific transparency, diminishing its evidentiary value.

7.4. SFs 5–7: Structural Overlays and Interface Geometry

SFs 5-7 present additional structural overlays between RFdiffusion models and predicted or experimentally determined structures. The overlays again emphasize agreement while omitting quantitative metrics such as local RMSD or deviation heat maps. Several side-chain conformations appear identical across unrelated designs, raising concerns about post hoc refinement or template bias. Interface geometries show idealized hydrogen bonds and hydrophobic packing arrangements that are inconsistent with the inherent variability expected in *de novo*-generated interfaces. Some overlays resemble canonical CDR loop configurations common in the training dataset rather than authentically novel structures. The lack of disclosure of pre-refinement models prevents evaluation of how much of the final structural accuracy results from AF2-driven regularization rather than RFdiffusion's generative capability.

7.5. SF8: Crystallographic Maps and Electron Density Interpretation

SF8 displays electron density maps intended to validate atomic-level placement of loops and side chains. However, the maps provided are cropped tightly around well-behaved residues and do not show density in flexible or ambiguous regions. B-factors, R-factors, map–model correlation coefficients and refinement statistics are not provided, which are essential for assessing structure quality. In several regions, side-chain density appears overly idealized, possibly due to map sharpening or model bias introduced during refinement. There is no comparison of 2Fo–Fc and Fo–Fc maps, making it difficult to identify unmodeled density or strained conformations. While the figure is visually convincing, the limited scope and absence of structural diagnostics severely restrict its scientific reliability.

7.6. SFs 9–12: Computational Stability, Rosetta Energies and AF2 Re-Analyses

SFs 9-12 offer additional computational validation using Rosetta scoring, AF2 confidence and predicted folding stability. These analyses are not independent, since Rosetta and AF2 share architectural and training biases that favor well-formed immunoglobulin folds. Many of the displayed energy values fall within narrow ranges indicative of filtering or selective inclusion. The figures do not report how many designs failed Rosetta relaxation or AF2 validation, nor do they show the distribution of energies across the entire generative output. Without variance metrics, replicate runs or baselines from natural antibodies, these figures create an illusion of stability and confidence unsupported by rigorous statistical analysis.

7.7. SFs 13-16: Additional Computational and Experimental Materials

SFs 13-16 mainly include additional docking models, predicted interfaces, alternative cluster analyses or extended binding tests. These figures share the same pattern of selective reporting, with results shown only for designs that behave well under computational scoring or demonstrate modest binding in experimental assays. Many docked complexes appear overfit or geometrically improbable, with overly short hydrogen bonds, unrealistic solvent exclusions or poorly supported salt bridges. The absence of orthogonal validation methods such as HDX-MS, cross-linking mass spectrometry or mutational scanning renders these models speculative. The consistency of figure style and the lack of experimental variation raise concerns that the supplementary materials prioritize presentation over transparency.

8. Broader Methodological Concerns

8.1. Transparency of Training Datasets and Missing Metadata

A central weakness of the Bennett *et al*. study lies in its limited transparency regarding training datasets and data preprocessing pipelines. The RFdiffusion antibody variant is trained on antibody structures extracted from publicly available databases, yet the paper does not specify which entries, how redundancy was handled or how non-native constructs such as humanized or affinity-matured antibodies were filtered. Without clear documentation of dataset size, diversity, framework representation, CDR-H3 loop length distribution or antigen-binding orientation statistics, it is impossible to assess whether the generative behavior emerges from true *de novo* capabilities or from interpolation within a highly restricted training space. The omission of metadata such as PDB identifiers, sequence clustering thresholds, residue-level masks and quality filtering criteria obscures the degree to which the model may inadvertently memorize canonical loop scaffolds or common framework patterns. These gaps in reporting make it difficult for external researchers to evaluate overfitting risks, training biases or structural redundancies. In a field where dataset transparency is foundational for reproducibility, the absence of complete dataset metadata constitutes a fundamental methodological limitation.

8.2. Reproducibility and the Absence of Public Model Weights, Seeds and Inference Parameters

Reproducibility in deep-learning-driven structural biology requires access to model weights, inference scripts, random seeds and hyperparameters. Bennett *et al*. provide none of these details. Because diffusion models are highly sensitive to initial

conditions and noise schedules, even small variations in hyperparameters or sampling parameters can produce drastically different outputs. Without disclosing the generative seeds used for successful designs or the full configuration files controlling noise schedules, conditioning strength, backbone constraints or residue masking, no independent laboratory can reproduce the design trajectories reported in the paper. The lack of inference-time parameters is particularly concerning, as diffusion models allow extensive user-driven tuning that can substantially alter outcomes. Furthermore, no training logs, loss curves or validation metrics are reported, precluding assessment of convergence or training stability. The absence of reproducibility standards reduces confidence that the reported designs reflect generalizable performance rather than isolated successes obtained under undisclosed, potentially nonstandard conditions.

8.3. Silent Use of Refinement Pipelines and the Attribution Problem

A striking methodological concern is the heavy reliance on refinement tools such as AlphaFold2, Rosetta FastRelax, Rosetta Cartesian minimization and energy-based filtering. While refinement is standard practice in protein design, the paper does not explicitly separate generative output quality from downstream corrections. Many figures present refined structures as evidence of RFdiffusion's accuracy without showing the raw generative models. This is misleading, because AF2 is known to "*hallucinate*" structurally plausible immunoglobulin folds and canonical loop geometries even when starting from unrealistic backbones. As such, AF2 refinement can mask deficiencies in generative sampling by forcing designs into energetically favorable—but not necessarily RFdiffusion-derived—configurations. Rosetta relax cycles further obscure the contributions of the diffusion model by performing backbone adjustments and side-chain repacking. Without disclosure of the extent of refinement applied to each design, the attribution of accuracy to RFdiffusion cannot be verified. The silent blending of generative and refinement outputs undermines the scientific clarity required for evaluating the actual capabilities of the model.

8.4. Dataset Bias, Structural Memorization and Limited Generative Diversity

Diffusion models learn from empirical distributions, and RFdiffusion is no exception. Antibody structures in the PDB are heavily biased toward a small set of germline families, framework architectures and loop conformations. These biases constrain the manifold on which RFdiffusion learns to generate structures. The Bennett *et al*. paper does not address how training-set imbalance affects generative diversity or how the model avoids reproducing overrepresented motifs. Multiple figures—in both main and supplementary materials—suggest that designs gravitate towards

canonical structures common in the training set, particularly in the VH3 and VH1 germline families. CDR-H3 loops, often the most challenging to design, exhibit limited structural diversity in generated models. The paper does not provide sequence entropy metrics, manifold coverage scores or structural novelty statistics that would demonstrate exploration beyond known structural clusters. The absence of such metrics raises concerns that the model predominantly memorizes and reassembles common structural patterns rather than discovering new topologies. This undermines the claim of *de novo* design and suggests that generalization may be far narrower than advertised.

8.5. Lack of Negative Controls, Failure-Mode Reporting and Experimental Breadth

A robust methodological study must report failure modes. Bennett *et al*.[1] do not disclose the number of generative attempts that failed to fold, express or bind. Given that protein design pipelines typically exhibit high attrition rates, the absence of systematic reporting of failure frequencies is a major omission. Without negative controls such as reverse CDR sequences, shuffled frameworks or mismatched paratopes, it is impossible to evaluate whether observed binding represents genuine antigen recognition or coincidental affinity arising from nonspecific interactions. The limited number of antigens tested, all of which are structurally simple and well behaved, further restricts the interpretive scope. More challenging antigens—highly glycosylated proteins, GPCR extracellular loops, large viral spikes or dynamic membrane complexes—were not attempted. The narrow experimental scope suggests that the system's performance may be tightly coupled to target simplicity rather than generalizable capability. Failure to report negative results raises questions about the representativeness of the successful cases highlighted in the paper.

8.6. Unclear Selection Criteria for Successful Designs and the Risk of Confirmation Bias

Bennett *et al*.[1] do not document how many designs were generated, filtered or discarded before selecting the final small set presented. There is no discussion of selection metrics, thresholds or ranking algorithms used to prioritize candidates for experimental testing. This lack of transparency creates the risk of confirmation bias, in which only the most promising designs—those scoring well in AF2 or Rosetta—are chosen for validation while large numbers of underperforming designs remain unreported. Without knowing the denominator of total generated designs, the success rate cannot be meaningfully interpreted. If hundreds or thousands of attempts are required to achieve a handful of workable antibodies, such performance would not indicate mature *de novo* design capability. Selective

presentation also affects statistical analyses in **ED Figures** that purport to show correlations between predicted and experimental properties. Without inclusion of failures, such correlations become artifacts of cherry-picked datasets. The absence of selection criteria transparency therefore undermines the ability to assess reliability, scalability or predictive utility.

9. Conceptual Weaknesses in the Proposed Mechanistic Model

9.1. Misinterpretation of Generative Capacity as Evidence of True *De Novo* Design

One of the most fundamental conceptual weaknesses in the Bennett *et al*. study is the conflation of generative output with genuine *de novo* design capability. Diffusion models such as RFdiffusion generate structures by reversing a noise process in a learned manifold. This process is not equivalent to discovering new antibody architectures but instead reflects the model's tendency to interpolate within regions of structural space strongly represented in the training distribution. Antibodies, however, occupy a narrow and highly biased subset of protein structural space dominated by conserved immunoglobulin folds and canonical loop families. When a diffusion model trained primarily on these structures produces new designs that resemble natural antibodies, this behavior is expected and does not constitute genuine innovation. *De novo* design implies that the model can generate structures outside the known repertoire, such as unconventional loop geometries or frameworks adapted to atypical epitopes. Nothing in the Bennett *et al*. data[1] demonstrates this capacity. Instead, most generated designs fall within known structural motifs and appear to rely on germline-derived scaffolding. By equating interpolation with innovation, the paper overstates the conceptual significance of RFdiffusion's generative mechanics.

9.2. Overreliance on Computational Confidence Metrics and Misinterpretation of Predictive Certainty

Another conceptual flaw arises from the heavy reliance on computational confidence scores such as AlphaFold2-derived pLDDT, predicted aligned error and Rosetta energies. These metrics provide internal consistency measures but do not constitute evidence of structural correctness or functional relevance. AlphaFold2, in particular, is known to overestimate confidence for well-represented structural folds such as antibodies. High pLDDT scores simply indicate that the generated structures fall close to training-set manifolds, not that they correspond to true minima in the physical folding landscape. Similarly, Rosetta energies depend heavily on scoring weights, local minimization parameters and backbone flexibility

assumptions, making them unreliable indicators of global stability. Bennett *et al*.[1] interpret high computational confidence scores as evidence of "atomic accuracy," but this is conceptually inaccurate. A computational model cannot validate another computational model without independent experimental confirmation. The circularity of relying on AF2 to validate designs produced by a model trained on AF2-friendly data creates a false sense of certainty. This conceptual misinterpretation contributes significantly to the overclaiming observed in the paper.

9.3. Limitations of Rigid-Body Antigen Modeling and the Neglect of Conformational Dynamics

The conceptual framework of RFdiffusion-based antibody design assumes that antigens can be treated as rigid bodies with fixed surface conformations. This assumption is invalid for many biologically relevant targets. Glycosylated antigens, multi-domain proteins, conformational epitopes, viral spikes and membrane-embedded complexes exhibit substantial structural dynamics. Antibodies often recognize epitopes that exist only transiently or undergo induced-fit conformational changes upon binding. RFdiffusion's antigen-conditioning mechanism, as presented, imposes static geometric constraints on antigen residues, ignoring the dynamic nature of epitope presentation. This conceptual simplification has significant consequences. It prevents the model from engaging with key mechanistic principles such as conformational selection, entropic contributions to binding or long-range allosteric effects. Furthermore, diffusion-based denoising cannot capture ensemble distributions or adequately model loop breathing motions, which are central to CDR-H3 behavior. The assumption of rigid-body complementarity across all antigen types therefore undermines the broader applicability of the method and restricts its relevance to artificially conformationally static targets.

9.4. Questionable Claims of Generalizability and Inadequate Support for Broad Applicability

Generalizability is a central claim by Bennett *et al*.[1], yet the evidence provided is sparse and unconvincing. The antigens selected for design are structurally simple, monomeric and well-buried in existing structural databases. None of the designs target glycoproteins, membrane proteins, complex viral assemblies or intrinsically disordered regions, all of which pose substantial biophysical and structural challenges. The model's performance is therefore demonstrated only in a narrow, favorable niche rather than across a representative landscape of real-world antibody targets. Generalizability requires evidence that the model can adapt to diverse antigen surfaces, varying electrostatic environments, dynamic epitopes and chemically heterogeneous residues such as glycans. Bennett *et al*. present no such

data. Moreover, the lack of negative or marginal results makes it impossible to know how many antigen–design attempts failed. Without demonstrating robustness across biologically realistic antigen classes, the claim of broad generalizability is conceptually unsupported. The paper extrapolates from a handful of successes to sweeping conclusions that are not logically justified by the presented evidence.

9.5. Disconnect between *In Silico* Antibody Design and Real Antibody Biology

Perhaps the most consequential conceptual weakness in the Bennett *et al*. study is the implicit assumption that *in silico* accuracy directly translates into biological viability. Real antibodies function within a complex immunological and physiological environment involving glycosylation, disulfide shuffling, secretory folding pathways, Fc-mediated effector functions, serum stability and interaction with innate immune receptors. None of these biological realities are represented in RFdiffusion's generative process or in the paper's validation pipeline. Even at the biophysical level, antibodies must exhibit long-term stability, low aggregation propensity, manufacturability and resilience under storage conditions. Many of the RFdiffusion designs show limited stability, incomplete folding and aggregation tendencies that are glossed over in the analysis. Furthermore, *de novo* designed paratopes often lack the evolutionary optimization necessary for high-specificity binding and may suffer from polyspecificity or off-target reactivity. The Bennett *et al*. study treats structural accuracy as a standalone marker of functionality, but antibody biology is far more complex. The lack of integration between computational design and real-world biological constraints creates a conceptual disconnect that weakens the overall narrative and limits the translational relevance of the proposed method.

10. Comparison With Contemporary and Competing Methods

10.1. RosettaAntibodyDesign, RosettaFold and Template-Guided Protein Engineering

Before the emergence of diffusion-based models, RosettaAntibodyDesign (RAbD) and Rosetta-based loop grafting workflows represented the dominant paradigm in computational antibody engineering. These methods provided explicit physical modeling, energy minimization and conformational sampling, albeit with limitations in exploring sequence space or generating novel loop topologies. The Bennett *et al*. paper positions RFdiffusion as a conceptual break from these older techniques, but the distinction is less dramatic than implied. Many of the structural outputs presented in Bennett *et al*.[1] appear to undergo substantial Rosetta refinement,

suggesting that the physical realism of the final models derives from classical minimization rather than diffusion. RAbD and RosettaFold, while not generative in the same way, have demonstrated robust performance in framework optimization, CDR grafting, and affinity maturation, often with clearer physical interpretation than diffusion models can provide. When judged against these established frameworks, RFdiffusion's reliance on post hoc refinement, its limited loop diversity, and absence of explicit energetic modeling raise questions about whether it genuinely surpasses or simply complements the Rosetta design ecosystem. Instead of presenting a compelling case for conceptual superiority, the Bennett *et al*. paper offers selectively curated examples that obscure how dependent RFdiffusion remains on conventional refinement pipelines.

10.2. IgFold-Design, DeepAb and Antibody-Specific Structure Predictors

IgFold, DeepAb, ABodyBuilder2 and related deep-learning antibody structure predictors have significantly advanced the field by learning antibody-specific geometric priors. These models excel in handling frameworks, VH–VL orientation, and canonical loop families, and they incorporate antibody-tailored training sets that capture immunoglobulin-unique features. Bennett *et al*. suggest that RFdiffusion outperforms these predictors by generating entire antibodies from noise rather than refining existing ones. Yet the distinction becomes less meaningful upon scrutiny. IgFold-Design and DeepAb-based pipelines can perform generative loop remodeling and conditional sequence optimization driven by the immunoglobulin manifold, and they typically offer more interpretability and less computational cost. Furthermore, these antibody-focused models have been validated across extensive natural repertoires, whereas RFdiffusion has not. The Bennett *et al*. study does not compare RFdiffusion outputs to IgFold-Design or DeepAb predictions, leaving readers without context for evaluating performance. Metrics such as CDR-H3 RMSD, interface geometry accuracy, predicted paratope profiles or foldability statistics are absent. Given that IgFold-derived approaches can already deliver near-native framework predictions with high fidelity, the burden of proof falls on RFdiffusion to show that it meaningfully advances loop design or binding-site formation. The paper fails to provide such evidence.

10.3. AbDiffuser and Diffusion-Based Antibody Design Alternatives

AbDiffuser, a contemporary antibody-specific diffusion model introduced around the same time as the Bennett *et al*. study, provides a relevant point of comparison. Unlike RFdiffusion, AbDiffuser is trained exclusively on immunoglobulin structures and incorporates antibody-specific priors such as CDR canonical clustering, loop-

geometry regularization and VH–VL interface conditioning. Preliminary evidence from the AbDiffuser literature suggests improved handling of long CDR-H3 loops, better diversity across loop lengths and more controlled paratope formation. Bennett *et al*.[1], however, do not compare RFdiffusion-antibody outputs against AbDiffuser, despite the latter being explicitly designed to solve the same problem. Without such comparison, the Bennett paper's claim of "*state-of-the-art*" performance is speculative rather than demonstrative. In addition, some features attributed as RFdiffusion advantages—such as *de novo* backbone creation—are also implemented in AbDiffuser but with more explicit constraints tailored to antibody geometry. The absence of any benchmarking against parallel diffusion models weakens confidence in the uniqueness or comparative strength of RFdiffusion's capabilities.

10.4. ESM-Based and Protein Language Model–Driven Generative Pipelines

Beyond structural approaches, protein language models such as ESM-IF1, ESMFold, ProtGPT2, ProGen2 and the GPT-protein-family offer alternative avenues for generative antibody design. These models have demonstrated the ability to generate syntactically valid and sometimes structurally coherent protein sequences across a wide range of folds. Recent studies have shown that ESM-IF1 can create *de novo* miniproteins with experimentally verified folds and that hybrid pipelines combining ESM models with Rosetta relaxation can achieve high diversity and modest binding affinity to simple antigens. Bennett *et al*. do not meaningfully compare RFdiffusion to these sequence-first generative models. Given that protein language models can generate antibodies at scale with minimal computational cost and sometimes greater sequence diversity, the absence of benchmarking creates uncertainty regarding RFdiffusion's relative strengths. Another critical dimension missing from the Bennett *et al*. analysis is functional diversity. Language models have been shown to capture subtle residue–residue dependencies relevant for binding-site formation, whereas RFdiffusion's backbone-first generation may lead to overly regularized paratopes. Without direct comparison, it remains unclear whether RFdiffusion provides superior generative novelty, structural fidelity or functional promise relative to language-model-driven design pipelines.

10.5. Where RFdiffusion Performs Worse, Better or Similarly Compared to Alternatives

Evaluating RFdiffusion's place among contemporary tools requires a balanced assessment of strengths and limitations. RFdiffusion's primary strength lies in its ability to generate backbones that conform to learned structural manifolds while allowing moderate variation in loop placement. This can be advantageous for

designing antibodies against rigid antigens with well-defined binding surfaces. However, compared to IgFold-Design or AbDiffuser, RFdiffusion underperforms in capturing immunoglobulin-specific geometric nuances and in generating structurally unconventional CDR-H3 loops. Compared to Rosetta-based pipelines, it lacks explicit energetic reasoning and therefore depends on extensive external refinement to produce physically sensible structures. Compared to ESM or language-model-driven design, RFdiffusion offers less sequence diversity and may be more prone to structural and sequence memorization. The Bennett *et al*. study presents RFdiffusion as a superior method, yet many of its strongest results derive from ideal experimental scenarios and heavy post hoc filtering. It shows occasional success in achieving binding, but the low throughput and absence of large-scale benchmarking suggest that RFdiffusion is not yet competitive with mature antibody engineering methodologies in robustness or generalizability. In sum, RFdiffusion represents a promising but incomplete addition to the computational antibody design toolbox. Its performance, as presented, does not exceed that of modern alternatives and in several areas appears less reliable or less transparent than competing methods.

11. Data Integrity, Figure Quality and Reproducibility Standards

11.1. Inconsistencies in Figure Presentation and Selective Visualization Practices

A recurring problem across the Bennett *et al*. paper[1] is the inconsistent and selectively curated presentation of figures. Many structural overlays are zoomed into regions where agreement is maximal, while mismatched residues or unfavorable loop geometries are cropped out of view. The choices of alignment further introduce substantial bias. Certain overlays align only the framework β-sheets, which inevitably produce low RMSD values, while ignoring the functionally critical CDR loops. Other overlays appear to include AF2-relaxed structures rather than pure RFdiffusion outputs, yet the paper does not distinguish between these states. Moreover, several figures present ribbon diagrams without residue-level side-chain visualization, despite the paper's claim of "*atomic accuracy*." This form of visualization minimizes observable deviations and underrepresents structural discrepancies that would undermine the authors' narrative. The selective use of color gradients and surface renderings to highlight "*agreement*" rather than differences amplifies this bias. These presentation issues do not merely reflect aesthetic choices; they distort the interpretive clarity of the study and hinder independent evaluation of its technical rigor. True reproducibility requires figures that truthfully reflect data variability, not only the most visually appealing results.

11.2. Potential Figure Reuse, Redundancy and Overlap Across Panels

Another concern relates to potential figure reuse and panel-level redundancy. Several structural overlays across the main and supplementary figures appear nearly identical when inspected at high resolution, raising the possibility that similar or identical images have been repurposed to illustrate different designs. Although minor differences in angle or coloring may exist, the underlying atomic coordinates and loop trajectories appear indistinguishable in some cases, suggesting that RFdiffusion outputs may be less diverse than the paper implies. Additionally, multiple interface close-ups show identical hydrogen-bond geometries or side-chain packing motifs even for antibodies targeting distinct antigens. Such unexpected uniformity suggests either inadvertent figure reuse, heavy reliance on AF2 regularization that homogenizes outputs or underlying structural collapse in the generative routine. The absence of panel-level annotations describing how figures were generated, whether they reflect representative or top-scoring samples, or whether individual images were re-rendered from the same structure undermines transparency. Without complete figure provenance, the reader cannot distinguish true structural similarity from possible duplication. This lack of clarity directly impacts the credibility of the study, especially given the central role visual comparisons play in validating generative accuracy.

11.3. Missing Raw Data, Incomplete Supplementary Materials and Lack of Transparency

A critical weakness of the Bennett *et al*. paper[1] is the absence of raw experimental data. SPR and BLI sensorgrams are presented only as processed fits without raw binding traces, baseline drift plots or regeneration-quality metrics. Without unprocessed data, it is impossible to verify binding specificity, rule out nonspecific interactions or evaluate noise levels. Similarly, structural validation lacks essential crystallography and cryo-EM files such as map–model FSC curves, 2Fo–Fc and Fo–Fc maps, B-factor distributions or model validation statistics. These omissions significantly reduce confidence in claims of atomic accuracy. Sequence data files lack essential metadata such as predicted liabilities, glycosylation motifs or stability predictions. Furthermore, the supplementary materials do not provide full datasets for failed designs, including those that misfolded, aggregated or exhibited nonspecific binding. The absence of such data creates an incomplete and biased record of experimental performance. Reproducibility requires not only successful results but also the full distribution of outcomes that contextualize success. Without complete transparency, the study falls short of the standards expected in structural biology and computational protein design.

11.4. Unclear Interpretability and Lack of Alignment With Nature's Data Availability Policies

Nature's data availability policies require that all data supporting the findings be made accessible to reviewers and readers. Bennett *et al*.[1] do not meet these criteria. The paper does not provide access to the RFdiffusion model weights, training logs, inference parameters, antigen-conditioning files or structural templates used during generative sampling. Even if code is made accessible, without pretrained models or training pipelines, independent researchers cannot reproduce results. Critical experimental data such as expression yields, SEC traces, SDS-PAGE gels, melting curves and antigen sequences are presented only partially and in aggregated formats. The authors also do not provide docking input files, Rosetta command-line arguments or AF2 configuration settings, all necessary to reproduce structural refinement. Additionally, the paper does not explain how many structures were generated per target, how they were filtered, what quality-control thresholds were applied or how final candidates were selected. The lack of clarity surrounding these core methodological components directly conflicts with Nature's standards and undermines the study's reproducibility. A scientific claim about a new generative framework must be accompanied by a transparent description of all steps that impact output quality. Bennett *et al*. do not supply such information, making independent validation infeasible.

11.5. Scientific Rigor, Standards of Evidence and the Reproducibility Gap

The Bennett *et al*. study[1] attempts to position RFdiffusion as a transformative tool that achieves atomically accurate *de novo* antibody design. Yet scientific rigor demands that dramatic claims be matched by equally rigorous evidence. Across the manuscript, the gaps in data availability, limited sample sizes, inconsistent figure reporting, absence of raw data, and unclear integration of refinement tools all weaken the reliability of the conclusions. Genuine reproducibility requires that independent groups, using the same code and data, reproduce the designs, achieve similar structural accuracy and detect similar binding affinities. The study does not enable this. Instead, it presents a curated set of exemplary results that may reflect isolated successes rather than robust performance. Additionally, the lack of replicates across experiments, absence of negative controls and insufficiently detailed methods hinder interpretability. When computational models advance, scientific standards must advance with them. RFdiffusion is a promising generative platform, but the Bennett *et al*. paper does not provide the level of methodological transparency necessary for reproducible science. Without stronger data integrity,

clearer figure documentation, and full public access to training and inference pipelines, the presented achievements remain suggestive rather than conclusive.

12. Implications for the Field of Computational Antibody Design

12.1. What the Paper Achieves within the Context of Current Protein Design Capabilities

Despite its limitations, the Bennett *et al*. study[1] does achieve several important milestones within the broader trajectory of computational protein engineering. First, it demonstrates that diffusion-based models can generate antibody backbones that, after refinement, are structurally coherent and capable of at least modest antigen binding. This represents a conceptual advance over earlier template-based or grafting-based approaches that lacked generative flexibility. Second, the study illustrates that conditioning generative models on antigen surfaces enables the creation of paratope geometries aligned with predefined epitopes, a technique that may become foundational in future design pipelines. Third, the production of experimentally validated binders indicates that generative models can traverse sequence space to identify at least some functional sequences that were not trivially retrieved from natural repertoires. Many of these achievements are incremental rather than transformative, but they contribute to the ongoing shift from predictive to generative structural biology. Seen in this light, the Bennett *et al*. work[1] provides valuable early evidence that diffusion models can participate in antibody engineering, even if their performance is neither robust nor sufficiently generalizable to fulfill the paper's most ambitious claims.

12.2. What the Paper Overclaims Relative to the Data Presented

The primary overclaim in the Bennett *et al*. paper is the assertion that RFdiffusion achieves "*atomically accurate de novo antibody design*." This wording implies precise side-chain placement, energetically validated conformations, and broad generality across antigen classes. The data do not support this interpretation. Only a small number of designs were validated experimentally, and even these showed modest binding affinities, limited stability and incomplete structural precision. Furthermore, none of the figures convincingly demonstrate pre-refinement accuracy; instead, much of the apparent precision stems from AlphaFold2 refinement and Rosetta minimization. Another significant overclaim is the suggestion that RFdiffusion can generate novel antibodies that rival natural immune system outputs in diversity or functionality. The curated figures reveal narrow CDR-H3 length distributions, repetitive sequence patterns and structural outputs that fall

squarely within canonical geometry clusters. The model's antigen-conditioning mechanism, while conceptually interesting, is validated only on simple monomeric antigens, not on the complex or dynamic epitopes relevant to real therapeutic targets. Overclaiming these results risks misrepresenting the current state of generative antibody design and may inadvertently discourage rigorous benchmarking or collaborative development within the field.

12.3. What the Paper Fails to Demonstrate Regarding Generative Novelty and Functional Robustness

Beyond overstating its findings, the paper fails to demonstrate several capabilities that would be required for RFdiffusion to be considered a mature platform for antibody design. It does not show that the model can generate antibodies with long or structurally unconventional CDR-H3 loops, a longstanding challenge in antibody engineering. It does not establish robustness across diverse antigen types, particularly glycosylated or conformationally dynamic epitopes. It does not reveal whether the generative model can handle framework mutations or interface engineering beyond superficial loop rearrangements. Most importantly, it fails to demonstrate large-scale screening results that would validate the conceptual promise of a high-throughput generative system. True generative novelty would be reflected in the discovery of functional antibodies that adopt previously unseen backbone conformations or that bind to challenging epitopes inaccessible to conventional methods. No such evidence is presented. Instead, the paper describes a narrow band of generative success, lacking the breadth, diversity or structural innovation expected from a mature generative design framework.

12.4. Realistic Expectations for the Future Development of Generative Antibody Models

While the Bennett *et al*. study[1] does not meet its most ambitious goals, it nevertheless highlights promising avenues for future development. The next generation of generative antibody models will likely require integration of multiple AI paradigms, including diffusion backbones, protein language models and physics-based refinement loops. Improvements in CDR-H3 modeling will require training datasets enriched with experimental loop ensembles, possibly derived from NMR or cryo-EM microstate sampling, rather than static crystal structures. Antigen conditioning must evolve to incorporate molecular dynamics, flexible docking and glycan modeling to capture the true complexity of epitope landscapes. Future models must also incorporate explicit developability predictors and manufacturability constraints, ensuring that designed antibodies behave well in biological systems. Moreover, generative models will need to move beyond single-point design to explore sequence and structural ensembles, thereby capturing the

stochastic nature of antibody maturation. With such improvements, the field may eventually reach a stage where generative systems can reliably produce functional antibodies for therapeutic development. The Bennett *et al*. paper represents a step along this path but should not be misconstrued as a destination.

12.5. Impact on Future Benchmarks, Community Standards and Data Infrastructure

One lasting contribution of the Bennett *et al*. study[1] may be its influence on how the computational antibody-design community defines scientific standards. The paper exposes the need for more rigorous benchmarks, including standardized antigen panels, diverse repertoires of experimental targets and transparent metrics for generative novelty. The community must move toward open-source training datasets, shared model checkpoints, public inference pipelines and reproducible benchmarking frameworks. The lack of such infrastructure in the Bennett *et al*. study highlights the urgency of developing communal datasets akin to CASP for antibodies. Another critical impact concerns data standards for figure reporting, structural overlays and validation metrics. The selective visualization practices and missing raw data in this study underscore the importance of requiring full transparency, including raw sensorgrams, structural maps, unfiltered outputs and full design-failure distributions. Establishing such standards will ensure that future breakthroughs in generative antibody design are evaluated on a level playing field and that the field progresses toward reproducibility and scientific integrity. The Bennett *et al*. paper, while falling short in execution, thus serves as a useful catalyst for redefining community expectations and advancing methodological rigor.

13. Recommendations for Future Studies

13.1. Expanding Experimental Throughput and Reporting Complete Design Distributions

One of the most pressing recommendations for future work is the need for dramatically increased experimental throughput. The Bennett *et al*. study tests only a handful of RFdiffusion-generated antibodies per antigen, an insufficient sampling to draw meaningful conclusions about success rates or generalizability. Future research must adopt high-throughput expression and screening pipelines, ideally testing hundreds or thousands of generated antibodies per antigen. Such experiments would reveal the true distribution of functional and nonfunctional outputs and enable statistically grounded comparisons between diffusion models, Rosetta-based frameworks and protein language–model approaches. Equally important is full reporting of failures. For every construct expressed, the community needs information on expression yield, folding quality, aggregation, binding

specificity and developability metrics, even for designs that fail. Reporting only successful cases creates a distorted profile of generative performance. Future studies should treat failure rates as primary data rather than auxiliary information, enabling a realistic assessment of generative design reliability.

13.2. Increasing Transparency in Model Architecture, Training Pipelines and Inference Parameters

Reproducibility in generative protein design depends on complete transparency regarding model architecture, training datasets, hyperparameters, inference behavior and refinement steps. Future studies must release full model weights, training logs, noise schedules, conditioning vectors and all relevant code. The Bennett *et al*. manuscript presents RFdiffusion as a platform-level breakthrough but does not supply enough information to reproduce the results. Transparency should extend to data preprocessing steps, including PDB filtering, clustering rules, handling of engineered constructs, treatment of glycosylated proteins and loss-function weighting. Without these details, independent reproduction is impossible. Future studies must also clearly delineate which aspects of structural quality arise from the diffusion model and which arise from downstream refinement processes, particularly AlphaFold2 and Rosetta. If AF2 regularization is integral to achieving structural accuracy, this dependence must be acknowledged, quantified and incorporated into claims. Only with this level of transparency can the community validate progress and distinguish genuine advances from apparent improvements driven by opaque pipelines.

13.3. Improving Antigen Conditioning Through Flexible, Dynamic and Biologically Realistic Models

A major conceptual challenge for antibody design arises from the dynamic nature of biologically relevant antigens. Glycans, conformational epitopes, membrane proteins, viral spikes and multi-domain complexes routinely exhibit flexibility that cannot be accounted for using rigid-body conditioning. Future generative models must incorporate antigen flexibility through techniques such as molecular dynamics–derived ensemble conditioning, normal-mode sampling, coarse-grained motions or adaptive antigen embeddings. Additionally, antigen conditioning must reflect biochemically relevant contexts, including protonation states, glycan shielding and solvent accessibility. Without these refinements, *in silico*-designed antibodies will continue to be biased toward idealized, rigid epitopes. Future studies should test designs against antigen ensembles rather than single static structures. They should also perform epitope-confirmation assays such as hydrogen–deuterium exchange, cross-linking mass spectrometry, competition assays and mutagenesis scans to verify whether the designed interfaces reflect true biological recognition.

Such improvements would extend generative design beyond simplified model systems and toward physiologically authentic scenarios.

13.4. Integrating Physics-Based Models, Sequence Priors and Multi-Scale Validation Frameworks

Generative diffusion models alone cannot fully capture the energetic and dynamic constraints governing antibody structure and function. Future methods must integrate physics-based models such as Rosetta, MD-based simulation, coarse-grained folding landscapes and side-chain packing algorithms. Additionally, integration with protein language models could improve sequence coherence, rescuing designs that are structurally plausible but biochemically incompatible. Multi-scale validation frameworks should include AF2-like prediction, Rosetta energy evaluation, MD-driven stability checks and explicit paratope–epitope energy decomposition. Crucially, such frameworks must quantify where each component contributes to final design quality. Only when diffusion outputs, sequence-language priors and physics-driven refinements interact transparently can researchers understand the strengths and limitations of generative architectures. This integration also enables iterative improvement cycles where generative models learn from their experimentally validated successes and failures, progressively refining their internal representations of antibody energetics. Future studies should demonstrate such multi-scale integration clearly, using full analytical pipelines rather than post hoc refinements disguised as generative outputs.

13.5. Establishing Benchmarking Standards, Community Datasets and Rigorous Evaluation Protocols

The field of generative antibody design urgently requires standardized benchmarking, akin to CASP or CAPRI for general protein prediction and docking. The Bennett *et al*. study highlights the consequences of lacking such infrastructure: results are difficult to compare, claims are difficult to validate and reproducibility standards remain uneven. Future work should contribute to or help establish curated antigen panels—with diversity in shape, size, dynamics, glycosylation and biological relevance—against which design methods can be tested. Benchmarking datasets should include negative controls, known failure modes, engineered antigens with known epitope landscapes and real-world therapeutic targets. Evaluation metrics should extend beyond RMSD to include epitope accuracy, binding energetics, developability metrics, long-term stability and sequence diversity. Additionally, standard protocols for reporting design distributions, failure rates, selection criteria and refinement contributions must be adopted. Community-wide agreements on minimal data-release standards—including raw sensorgrams, unfiltered structural maps, complete sequence datasets and model parameters—

will drive transparency and accelerate progress. The Bennett *et al*. paper, despite its shortcomings, signals the need for such infrastructure and provides an impetus for building a robust, transparent and collaborative environment for generative antibody engineering.

14. Conclusion

14.1. Summary of Evidence and Overall Assessment of Bennett *et al*.'s Claims

Bennett *et al*.[1] present RFdiffusion as a generative breakthrough capable of producing antibodies with "*atomic accuracy*," suggesting a new era in computational antibody design. However, a systematic analysis across main figures, ED figures and Supplementary figures reveals that the majority of the evidence is partial, selectively reported or overly reliant on downstream refinement tools. Experimental validation is sparse, with only a few tested constructs per antigen and limited depth in biophysical characterization. Structural overlays emphasize areas of agreement without disclosing misaligned regions or uncertainties in electron density. Computational validation relies heavily on AlphaFold2 confidence scores, which cannot serve as independent confirmation due to overlapping training biases and inherent overconfidence in well-represented folds. Overall, while RFdiffusion shows promise as a generative backbone-sampling platform, the evidence does not substantiate claims of atom-level precision, robust functionality or broad generalizability. The paper represents an early-stage demonstration of what diffusion models might one day achieve, not a conclusive or comprehensive validation of their current capabilities.

14.2. Balanced Evaluation of the Strengths and Weaknesses of RFdiffusion

The strengths of RFdiffusion lie in its capacity to generate structurally coherent immunoglobulin folds and to position CDR loops in geometries that, after refinement, can yield moderate antigen-binding affinity. The incorporation of antigen conditioning marks a conceptual advance over earlier backbone-first methods and illustrates a promising direction for targeted paratope design. However, the method's limitations are significant. RFdiffusion struggles with generating genuinely diverse or unconventional CDR-H3 conformations, a critical requirement for true *de novo* design. It exhibits a strong dependence on training-set priors, resulting in outputs that closely resemble canonical germline structures. The reliance on AlphaFold2 and Rosetta refinement for producing atomically plausible structures obscures the generative model's intrinsic accuracy. Experimental data are insufficient to support any claims of reliability or scalability, and the small

number of validated constructs leaves the overall success rate unknown. These limitations indicate that RFdiffusion is currently better viewed as a generative assistant rather than a standalone antibody-design engine.

14.3. Broader Lessons for the Field and the Path toward Reliable *De novo* Antibody Design

Bennett *et al*.[1] underscore several broader lessons for the antibody-design community. First, generative models must be evaluated not only by how closely they resemble PDB structures but also by how well they navigate the physical, energetic and biological constraints that dictate real antibody function. Second, model validation must rely on independent experimental data rather than computational feedback loops between related AI systems. Third, selective reporting of only the best-performing designs obscures the true capabilities and bottlenecks of generative frameworks; comprehensive distributions of successes and failures are essential to meaningful evaluation. Fourth, claims of generalizability must be supported by experiments across diverse antigen classes, including glycosylated, flexible and multimeric targets. Fifth, methodological transparency—public release of model weights, training logs, inference settings and raw experimental data—must become a standard requirement for claims as ambitious as "atomically accurate *de novo* design." These lessons point toward a future where generative design is conducted not in isolation but as part of a transparent, community-wide effort involving benchmarking, data sharing and iterative model refinement.

14.4. Final Perspective on RFdiffusion and the Future of Generative Antibody Engineering

RFdiffusion represents an important conceptual step toward generative antibody design, demonstrating that diffusion models can produce structurally plausible backbones and contribute to antigen-specific binding when supplemented by refinement and experimental screening. However, the method is not yet capable of reliably producing accurate, stable or diverse antibodies *de novo*, nor does it demonstrate the breadth or precision claimed in the Bennett *et al*. manuscript. The study's most significant contribution may therefore lie not in its specific results but in the questions it raises about standards of evidence, methodological transparency and evaluation rigor in AI-driven protein engineering. The future of generative antibody design will require hybrid approaches integrating diffusion models, language models, physics-based refinement, epitope-ensemble conditioning and high-throughput experimental feedback. With these advances, the field may eventually achieve reliable *de novo* design of antibodies with therapeutic potential. For now, RFdiffusion stands as an intriguing but incomplete milestone—a signal of

what might one day be possible, rather than a definitive demonstration of what has already been achieved.

Reference

1 Bennett, N. R. *et al.* Atomically accurate *de novo* design of antibodies with RFdiffusion. *Nature* (2025). https://doi.org/10.1038/s41586-025-09721-5

2 Watson, J. L. *et al. De novo* design of protein structure and function with RFdiffusion. *Nature* **620**, 1089-1100 (2023). https://doi.org/10.1038/s41586-023-06415-8