AIR: Inference-Time Refinement for Discrete-Diffusion Antibody Humanization

Anna Karpova [1] **Andrey Shevtsov** [2] **Viacheslav Meshchaninov** [*3 1] **Pavel Strashnov** [*2]

Abstract

Antibody humanization replaces non-human framework residues with human-like ones while preserving the murine CDR loops that determine binding. Discrete-diffusion models such as HuDiff cast this as conditional sequence generation, but their one-pass sampling cannot navigate the trade-off between humanness and structural fidelity. We propose AIR, an inference-time refinement framework that lets a pre-trained discrete-diffusion model audit and revise its own predictions. Each cycle remasks residues flagged as low-quality by the model's own confidence, an external biological scorer, or both, and resamples them within a more complete sequence context. On the Humab25 benchmark, AIR traces a controllable trade-off between humanness and fidelity. Pairing external nativeness scoring with a low-rate self-consistency stage produces sequences that are as human-like as the experimentally validated humanizations and retain the same fraction of the original murine residues. AIR requires no retraining and operates as a drop-in wrapper for existing discrete-diffusion humanization models.

1. Introduction

Murine antibodies must be humanized to reduce immunogenicity while preserving the CDR loops that determine binding. (Foote & Winter, 1992). The central challenge is a trade-off: making a sequence more human-like can disrupt framework residues that physically support the CDR loops, which degrades function.

Humanization methods have evolved from classical CDR grafting to deep-learning sequence models such as Sapiens (Prihoda et al., 2022) and the AbNatiV pipelines (Ramon et al., 2024; 2025). HuDiff (Ma et al., 2024) frames humanization as conditional discrete diffusion, jointly modeling the heavy and light chains. One limitation of one-pass discrete diffusion in this setting is that residues are committed before the full sequence context is available, making it difficult to navigate the preservation–humanness trade-off. Iterative refinement has been explored in adjacent protein-design tasks, including MapDiff (Bai et al., 2025) for inverse folding and RERD (Uehara et al., 2025) for reward-guided refinement of protein and DNA sequences, but neither is designed for humanization or enforces CDR preservation.

We propose AIR, an inference-time refinement framework that wraps a pre-trained discrete-diffusion model. Each refinement cycle identifies residues flagged as low-quality, either by the model's own confidence or by an external biological scorer, remasks them, and resamples within a more complete sequence context. We further introduce a soft proximity prior on framework residues near the CDRs, motivated by the role of Vernier zones in supporting CDR conformation (Foote & Winter, 1992). On the Humab25 benchmark, AIR traces a controllable trade-off between humanness and fidelity. Combining external nativeness scoring with a low-rate self-consistency stage produces sequences that match the experimentally validated humanizations. AIR requires no retraining and operates as a drop-in wrapper for existing discrete-diffusion humanization models.

2. Background

Discrete diffusion with absorbing states. Discrete diffusion for protein sequences operates over a finite amino-acid alphabet. The forward process gradually replaces residues with a special `[MASK]` token. The reverse process, learned during training, predicts the original residue s_0 given a partially masked sequence s_t, approximating $P_\theta(s_0 \mid s_t)$. This corresponds to any-order autoregressive generation in which the unmasking order is not fixed.

HuDiff (Ma et al., 2024) adapts this framework to antibody humanization. The murine CDR loops are kept intact as a fixed condition while framework residues are masked and reconstructed. The model is pre-trained on paired human antibody sequences from the Observed Antibody Space (Olsen et al., 2022) and fine-tuned with biological guidance from AbNatiV. We use HuDiff in its FR mode as the base

[1]HSE University, Moscow, Russia [2]AXXX, Moscow, Russia [3]Constructor University, Bremen, Germany. Correspondence to: Pavel Strashnov <strashnov@axxx.tech>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

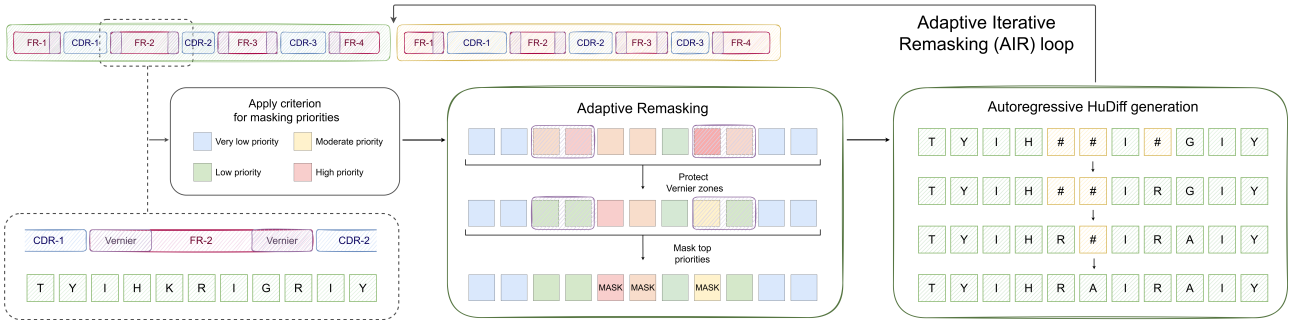

Figure 1. **Adaptive Iterative Refinement (AIR) overview.** Starting from a HuDiff-humanized paired antibody sequence, AIR iteratively scores positions under a chosen criterion, protects framework residues near the CDRs, remasks the highest-priority positions, and regenerates them autoregressively with HuDiff.

model that AIR wraps.

AbNatiV (Ramon et al., 2024) is a VAE-based scorer trained on natural human antibody sequences. It produces a per-residue reconstruction error that quantifies how well each position fits the human distribution, in addition to a whole-sequence nativeness score. We use the per-residue error as a remasking signal inside AIR. AbNatiV-as-humanization-pipeline (and its successor AbNatiV2 (Ramon et al., 2025)) are also baselines we compare against. The per-residue and whole-sequence uses of AbNatiV are distinct.

3. Method

AIR operates as a wrapper around a pre-trained discrete-diffusion humanization model. Given a sequence s produced by a single forward pass through the base model, AIR runs K refinement cycles. Each cycle identifies residues that look problematic under a chosen criterion, returns them to the masked state, and resamples them autoregressively in a randomized order. Because the criterion is evaluated on the most recent sequence, later cycles audit a more complete and humanized context than earlier ones. The procedure is summarized in Algorithm 1.

Linear annealing schedule. The remasking rate p_c decreases linearly from p_{start} to p_{end} across cycles. Early cycles edit large fractions of the sequence to explore globally. Later cycles make small targeted edits to consolidate. When $K = 1$ we set $p_c = p_{\text{start}}$.

Sequential resampling. Within each cycle, the masked positions R are resampled one at a time in a random order, and the sequence s is updated in place after each draw. Each prediction therefore conditions on the most recently resampled tokens, not on the cycle-start state. Parallel resampling, by contrast, would commit all $|R|$ tokens simultaneously and miss the within-cycle context.

3.1. Remasking criteria

The behavior of AIR is determined by the criterion M that scores each residue position for remasking. We consider four criteria, drawing on two information sources. The first is the model's own predictive confidence. The second is an external biological nativeness scorer.

$$\text{Confidence:} \quad w_i = 1 - \max_v P_\theta(s_i = v \mid s) \tag{1}$$

$$\text{External:} \quad w_i = \text{MSE}_i^{\text{AbNatiV}}(s) \tag{2}$$

$$\text{Additive:} \quad w_i = \alpha\, \widetilde{\text{MSE}}_i + (1-\alpha)\, \widetilde{\text{Conf}}_i \tag{3}$$

$$\text{Multiplicative:} \quad w_i = \widetilde{\text{MSE}}_i \cdot \widetilde{\text{Conf}}_i \tag{4}$$

Here $\widetilde{(\cdot)}$ denotes per-sequence min-max normalization to $[0, 1]$, $\text{Conf}_i = 1 - \max_v P_\theta(s_i = v \mid s)$, and $\alpha \in [0, 1]$ controls the weighting in the additive variant. *Confidence* flags positions where the model itself is uncertain. *External* flags positions where AbNatiV's reconstruction error indicates non-native features. The two hybrids combine these signals within each cycle. *Additive* mixes them linearly. *Multiplicative* requires both signals to agree before a position is flagged.

Sequential. A fifth strategy decouples the two information sources across stages. Stage 1 runs AIR with the *External* criterion at moderate remasking rate to push the sequence toward human-like patterns. Stage 2 runs AIR with the *Confidence* criterion at a much lower remasking rate to revise residues that the model itself flags as uncertain in light of the now-humanized context. This is equivalent to two sequential calls to Algorithm 1 with different criteria and schedules.

3.2. Proximity-based protection

Aggressive humanization can disrupt framework residues that physically support the CDR loops, including the *Vernier zones* (Foote & Winter, 1992). External nativeness scorers often assign high error to peri-CDR positions because they may resemble the murine source even in successful humanizations. To protect them, we apply a soft proximity prior to the criterion weights. Any position within a fixed sequence radius of a CDR has its weight scaled by $\gamma \in [0, 1]$, with smaller γ enforcing stronger protection. The mechanism adds no parameters or training cost.

In our implementation the proximity radius is 5 residues on each side of any CDR boundary, chosen from empirical analysis of where AbNatiV-guided refinement concentrates its

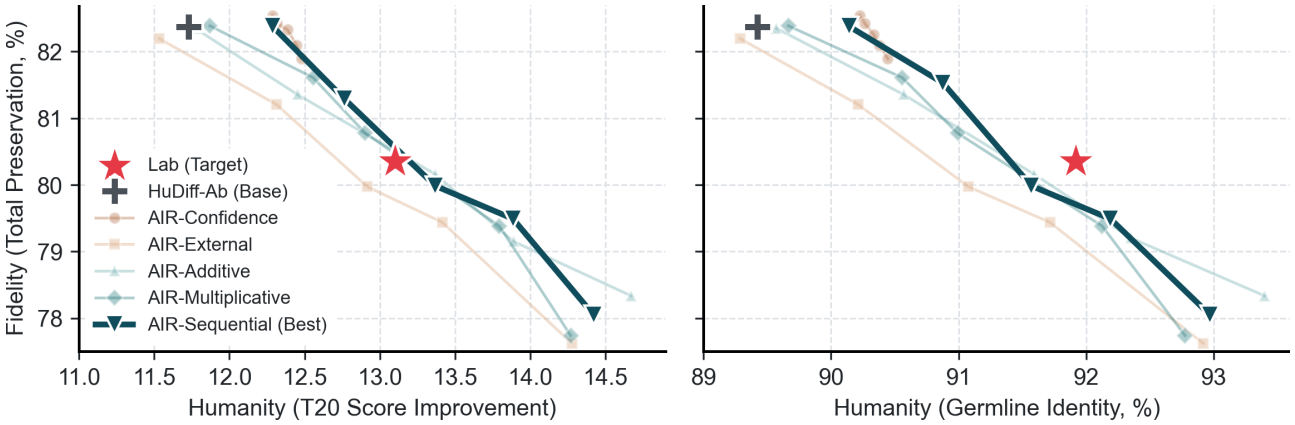

Figure 2. **Humanity–fidelity trade-off on Humab25.** *Lab* is the experimentally validated humanization. *HuDiff-Ab* is the one-pass base model. AIR variants are shown as Pareto-frontier configurations. *Sapiens1*, *AbNatiV*, and *AbNatiV2* fall above the y-axis range and are reported in Tab. 1. *AIR-Sequential* reaches the operating point closest to Lab on T20 and matches it on Total Preservation.

mutations (Appendix A). The 5-residue radius is a sequence-distance proxy for the more anatomically defined Vernier zones. The standard Kabat-defined Vernier preservation metric we report in evaluation overlaps with this proximity set but is not identical to it. The empirical effect of protection is therefore visible primarily in the broader Total Preservation metric (Section 4.3).

4. Experiments

4.1. Setup

Dataset. We evaluate on the Humab25 benchmark (Ma et al., 2024), which consists of 25 therapeutic antibodies that have been clinically validated as humanized variants of murine sources. We leave evaluation on the larger HuAb348 benchmark to future work.

Metrics. Following Ma et al. (2024), we report two humanness metrics and two fidelity metrics. Humanness is measured by the T20 score improvement over the parental sequence (Prihoda et al., 2022) and by germline identity to the closest human V-gene from the IMGT database, reported separately for the heavy (H) and light (L) chains. Fidelity is measured by total preservation, the fraction of residues identical to the parental sequence, and by Vernier preservation, the same quantity computed over Vernier-zone positions.

Algorithm 1 AIR: Adaptive Iterative Refinement

Require: Initial sequence s, cycles $K \geq 2$, rates $(p_{\text{start}}, p_{\text{end}})$, criterion M, protection factor γ
for $c = 1, \ldots, K$ **do**
  $p_c \leftarrow p_{\text{start}} - \frac{c-1}{K-1}(p_{\text{start}} - p_{\text{end}})$
  $w \leftarrow \text{Criterion}_M(s)$ {Eq. (1)–(4)}
  $w_i \leftarrow \gamma \cdot w_i$ for all i within radius r of any CDR {Vernier protection}
  $R \leftarrow$ top-$\lfloor p_c \cdot |s| \rfloor$ positions of w
  Mask positions R in s
  for $i \in \text{shuffle}(R)$ **do**
    $s_i \sim P_\theta(\cdot \mid s)$ {update s in place}
  end for
end for
return s

Baselines. We compare AIR against several baselines on Humab25. *Lab* is the experimentally validated humanization and serves as the gold standard. *HuDiff-Ab* (Ma et al., 2024) is the base model that AIR wraps, reproduced over five seeds. *Sapiens1* (Prihoda et al., 2022) is a fine-tuned protein language model for humanization. We use the values reported in the HuDiff paper. *AbNatiV* and *AbNatiV2* (Ramon et al., 2024; 2025) are deep-learning humanization pipelines that we run end-to-end on Humab25. *HuAbDiffusion* (Liu et al., 2025) is a recent discrete-diffusion humanization model whose code is unavailable and which is not evaluated on Humab25 in the original work. We therefore cite it but cannot include it in the comparison. We use the HuDiff-Ab *FR* sampling mode throughout. The HuDiff-Ab *Inpainting* mode uses germline-template alignment as an additional prior. AIR is orthogonal to this choice and combining the two is left to future work.

AIR configurations. We sweep over K, $(p_{\text{start}}, p_{\text{end}})$, α, and γ, with multiple seeds per configuration. The headline plot reports Pareto-frontier configurations for each AIR variant. The full sweep is described in Appendix B.

4.2. Refinement enables a controllable humanity–fidelity trade-off

Figure 2 shows the humanity–fidelity trade-off across all evaluated methods. The HuDiff-Ab base sits at the low-humanity, high-fidelity corner. Our seeded reproduction (T20 $= 11.73 \pm 0.16$, Total Preservation $= 82.37 \pm 0.10$) matches the reported HuDiff-Ab values (Ma et al., 2024) within seed variance. Sapiens1, AbNatiV, and AbNatiV2 occupy a high-fidelity, low-humanity region above the AIR cloud. They preserve more of the murine residue identity than HuDiff-Ab and Lab, but at humanity levels well below those reached by either, reflecting their conservative editing.

Table 1. **Per-method comparison on Humab25**. For each AIR variant we report the configuration with the highest T20 such that Total Preservation $\geq$ Lab. Per-chain breakdowns are in App. C and recommended hyperparameters are in App. B.

Method	T20	Germline ID (%)	Total Pres. (%)	Vernier Pres. (%)
Lab (Ma et al., 2024)	13.56	91.92	80.35	86.20
Sapiens1 (Prihoda et al., 2022; Ma et al., 2024)	10.20	87.18	88.44	90.00
AbNatiV (Ramon et al., 2024)	9.20	86.06	84.82	85.90
AbNatiV2 (Ramon et al., 2025)	9.29	86.26	86.04	88.68
HuDiff-Ab (Base) (Ma et al., 2024)	11.73±0.16	89.42±0.15	82.37±0.10	85.66±0.77
AIR-Confidence	12.48±0.21	90.45±0.24	81.89±0.12	86.10±0.26
AIR-External	12.51±0.02	90.61±0.08	80.51±0.02	85.45±0.40
AIR-Additive	12.92±0.29	91.08±0.30	80.53±0.08	85.61±0.44
AIR-Multiplicative	12.90±0.14	90.99±0.26	80.78±0.14	85.00±0.59
AIR-Sequential	**13.34±0.13**	**91.52±0.16**	80.43±0.25	85.33±0.74
AIR-Sequential×2	13.11±0.02	91.31±0.06	**81.07±0.13**	85.55±0.27

All five AIR criterion variants extend the base model along a coherent Pareto front. AIR-Confidence is the most conservative, occupying a narrow region near the base. AIR-External, AIR-Additive, AIR-Multiplicative, and AIR-Sequential reach humanity levels approaching the laboratory humanization, with the corresponding decline in preservation. The four non-Confidence variants form overlapping fronts in this region.

Table 1 summarizes the best-balanced configuration of each AIR variant, defined as the highest T20 reached at Total Preservation no lower than the laboratory humanization. AIR-Sequential reaches the operating point closest to Lab on the humanity axis, with T20 $= 13.34 \pm 0.13$ at Total Preservation $= 80.43 \pm 0.25$, against Lab T20 $= 13.56$ and Total $= 80.35$. No other AIR variant reaches T20 within 0.6 of Lab while maintaining Lab Total Preservation. The *Sequential* structure decouples a high-rate external-scoring stage from a low-rate self-consistency stage. The advantage comes from the second stage. It edits only the residues the model itself flags as inconsistent, in the now-humanized context produced by the first stage. Iterating the Sequential procedure twice (External→Confidence→External→Confidence) trades a small amount of T20 reach for additional Total Preservation, reaching T20 $= 13.11$ at Total $= 81.07$. Vernier-zone preservation remains modestly below Lab for all AIR variants we evaluate, with AIR-Confidence reaching the highest Vernier preservation among AIR configurations ($86.10 \pm 0.26\%$ vs Lab 86.20%) at the cost of lower humanity.

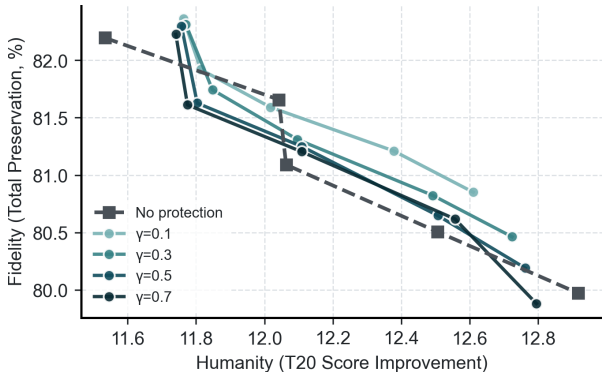

Figure 3. **Proximity-based protection applied to AIR-External at $p = 0.01$.** Each curve corresponds to a protection strength γ, with cycle count $K \in \{1, 3, 5, 10, 15\}$ increasing left to right along the curve.

4.3. Proximity-based protection mitigates the structural cost of external scoring

Figure 3 compares AIR-External with and without proximity-based protection across four protection strengths $\gamma \in \{0.1, 0.3, 0.5, 0.7\}$ at $p = 0.01$. Stronger protection (lower γ) shifts the Pareto curve modestly upward in Total Preservation, with the gap most visible at higher cycle counts where the unprotected baseline drives further into the high-T20 region. At $K = 15$, $\gamma = 0.1$ gains 0.88 percentage points of Total Preservation at the cost of 0.31 T20 relative to the unprotected baseline. The effect is monotonic in γ but modest in absolute terms, with seed variance ($n = 2$ per cell) of comparable magnitude to the inter-γ gap. The same protection mechanism applied to AIR-Multiplicative produces a similar pattern (Appendix D).

The mechanism we implement targets framework residues within five sequence positions of any CDR boundary, while the Vernier preservation metric we report follows the standard Kabat-defined positions. The two sets overlap but are not identical, and the empirical effect of protection is therefore most visible in the broader Total Preservation metric rather than in Vernier Preservation specifically.

5. Conclusion

We presented AIR, an inference-time refinement framework for discrete-diffusion antibody humanization. AIR audits a one-pass sample using either the model's own confidence, an external biological scorer, or both, and resamples low-quality positions within a more complete sequence context. On Humab25, AIR traces a controllable humanity–fidelity trade-off, and a sequential variant reaches the operating point closest to the laboratory humanization, matching Lab on Total Preservation while approaching Lab on T20.

Acknowledgements

This research was supported in part through computational resources of HPC facilities at HSE University. The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4E0002 and the agreement with HSE University №139-15-2025-009.

References

Bai, P., Miljković, F., Liu, X., De Maria, L., Croasdale-Wood, R., Rackham, O., and Lu, H. Mask-prior-guided denoising diffusion improves inverse protein folding. *Nature Machine Intelligence*, 7(6):876–888, June 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01042-6. URL https://www.nature.com/articles/s42256-025-01042-6.

Foote, J. and Winter, G. Antibody framework residues affecting the conformation of the hypervariable loops. *Journal of Molecular Biology*, 224(2):487–499, March 1992. ISSN 00222836. doi: 10.1016/0022-2836(92)91010-M. URL https://linkinghub.elsevier.com/retrieve/pii/002228369291010M.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L. D., Thouvenin-Contet, V., and Lefranc, G. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental and comparative immunology*, 27 1:55–77, 2003.

Liu, D., Hao, X., and Fan, L. Huabdiffusion: a discrete language diffusion model used for antibody humanization. *Briefings in Bioinformatics*, 26(6):bbaf658, 11 2025. ISSN 1477-4054. doi: 10.1093/bib/bbaf658. URL https://doi.org/10.1093/bib/bbaf658.

Ma, J., Wu, F., Xu, T., Xu, S., Liu, W., Yan, D., Bai, Q., and Yao, J. An adaptive autoregressive diffusion approach to design active humanized antibody and nanobody. *bioRxiv*, 2024.

Olsen, T. H., Boyles, F., and Deane, C. M. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. doi: https://doi.org/10.1002/pro.4205. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4205.

Prihoda, D., Maamary, J., Waight, A., Juan, V., Fayadat-Dilman, L., Svozil, D., and Bitton, D. A. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14(1):2020203, 2022. doi: 10.1080/19420862.2021.2020203. URL https://doi.org/10.1080/19420862.2021.2020203.

Ramon, A., Ali, M., Atkinson, M., Saturnino, A., Didi, K., Visentin, C., Ricagno, S., Xu, X., Greenig, M., and Sormanni, P. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nature Machine Intelligence*, 6(1):74–91, January 2024. ISSN 2522-5839. doi: 10.1038/s42256-023-00778-3. URL https://www.nature.com/articles/s42256-023-00778-3.

Ramon, A., Frassetto, N., Zhao, H., Xu, X., Greenig, M., Onuoha, S., and Sormanni, P. Deep learning assessment of nativeness and pairing likelihood for antibody and nanobody design with AbNatiV2, November 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.10.31.685806.

Uehara, M., Su, X., Zhao, Y., Li, X., Regev, A., Ji, S., Levine, S., and Biancalani, T. Reward-guided iterative refinement in diffusion models at test-time with applications to protein and dna design, 2025. URL https://arxiv.org/abs/2502.14944.

A. AbNatiV mutation hotspots

To motivate the proximity radius used in our protection scheme (Section 3.2), we examined where AbNatiV would place edits when applied as a one-shot humanization scorer to the Humab25 sequences. Figure 4 shows the per-position mutation count, aligned to IMGT (Lefranc et al., 2003) numbering and split by chain.

The distribution is sharply non-uniform. Most mutations concentrate in framework positions within a few residues of CDR boundaries. The dominant peak sits immediately C-terminal to H-CDR2, with secondary peaks at the edges of H-CDR1, L-CDR1, and L-CDR2. These positions overlap with the Kabat-defined Vernier zone (Foote & Winter, 1992), known to influence CDR conformation. A radius of 5 residues around each CDR covers these hotspots without extending into framework regions where AbNatiV proposes few edits.

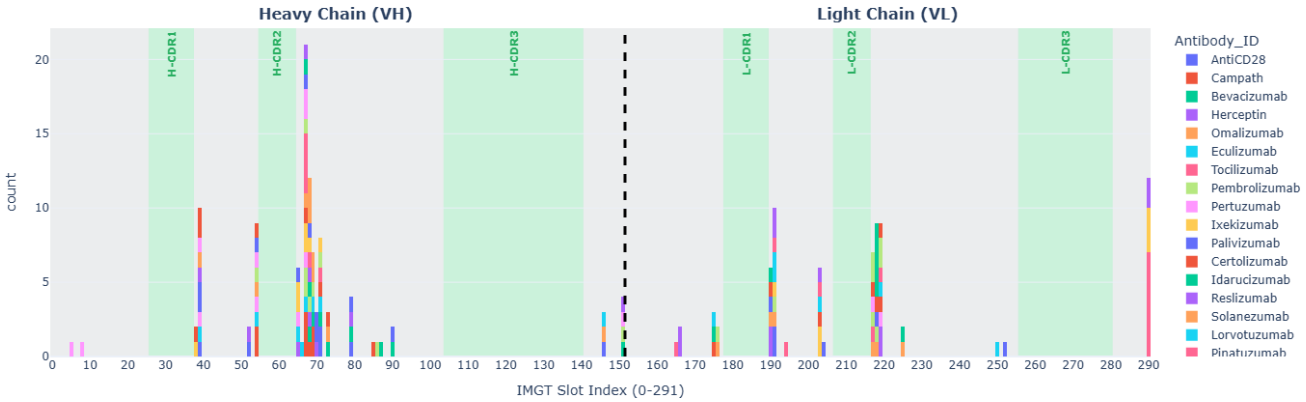

Figure 4. Where AbNatiV proposes mutations on Humab25, by IMGT position. Edits cluster in framework residues adjacent to CDR boundaries, motivating the proximity radius in Section 3.2.

B. Hyperparameter sweep

The configurations reported in Table 1 are selected from a sweep over the cycle count K, the remasking rate p, and the variant-specific weights. For each variant we report the configuration with the highest mean T20 subject to mean Total Preservation no lower than the laboratory humanization, and ties are broken by Total Preservation. Across the full sweep we evaluate 394 unique configurations and 1,695 individual runs. The sweep holds the remasking rate constant across cycles ($p_{\text{start}} = p_{\text{end}} = p$). Protection is disabled ($\gamma = 1$) outside Section 4.3 and Section D.2.

Table 2 lists the values swept for each variant. Table 3 gives the configurations chosen for Table 1.

Figure 5 shows the full distribution of points behind the Pareto fronts in Figure 2. The *Confidence* cloud is concentrated in a narrow region close to the base model, reflecting that the model assigns low confidence to relatively few residues even after several refinement cycles. The four non-*Confidence* variants overlap substantially, and their Pareto fronts run roughly in parallel through the high-T20 region. *Sequential* and *Sequential×2* reach the highest T20 values at Total Preservation no lower than the laboratory humanization, with a small additional reach for *Sequential*. The headline configurations selected in Table 1 sit on these front lines rather than at isolated extreme points.

Table 2. Sweep ranges for each AIR variant on Humab25. Seeds counts the distinct random seeds per configuration. Configs and runs counts are restricted to points with at least two completed seeds.

Variant	Ranges	Seeds	Configs	Runs
Confidence	$K \in \{1, 3, 5, 10, 12, 15, 20, 50, 100, 150, 200, 300, 400\}$, $p \in \{0.01, 0.05, 0.10, 0.15\}$	13	42	444
External	$K \in \{1, 3, 5, 10, 15\}$, $p \in \{0.01, 0.03, 0.05\}$	3	15	43
Additive	$K \in \{1, 3, 5, 10, 20\}$, $p \in \{0.01, 0.03, 0.05\}$, $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$	5	75	368
Multiplicative	$K \in \{1, 3, 5, 10, 20, 50\}$, $p \in \{0.01, 0.03, 0.05, 0.07\}$	5	24	120
Sequential	$K_a \in \{1, 2, 3, 4, 5, 10\}$, $p_a \in \{0.01, 0.02, 0.03, 0.05\}$, $K_w \in \{25, 50, 100\}$, $p_w \in \{0.01, 0.02, 0.03\}$	3	216	646
Sequential×2	As Sequential with two refinement blocks and $p_a \in \{0.01, 0.02\}$	3	22	74

Table 3. Recommended configurations corresponding to the AIR rows of Table 1. All configurations use no protection ($\gamma = 1$).

Variant	Configuration
AIR-Confidence	$K = 100, p = 0.01$
AIR-External	$K = 10, p = 0.01$
AIR-Additive	$K = 20, p = 0.01, \alpha = 0.5$
AIR-Multiplicative	$K = 50, p = 0.01$
AIR-Sequential	Stage 1: $K_a = 4, p_a = 0.02$. Stage 2: $K_w = 50, p_w = 0.02$
AIR-Sequential×2	Two blocks, each with $K_a = 3, p_a = 0.01, K_w = 100, p_w = 0.01$

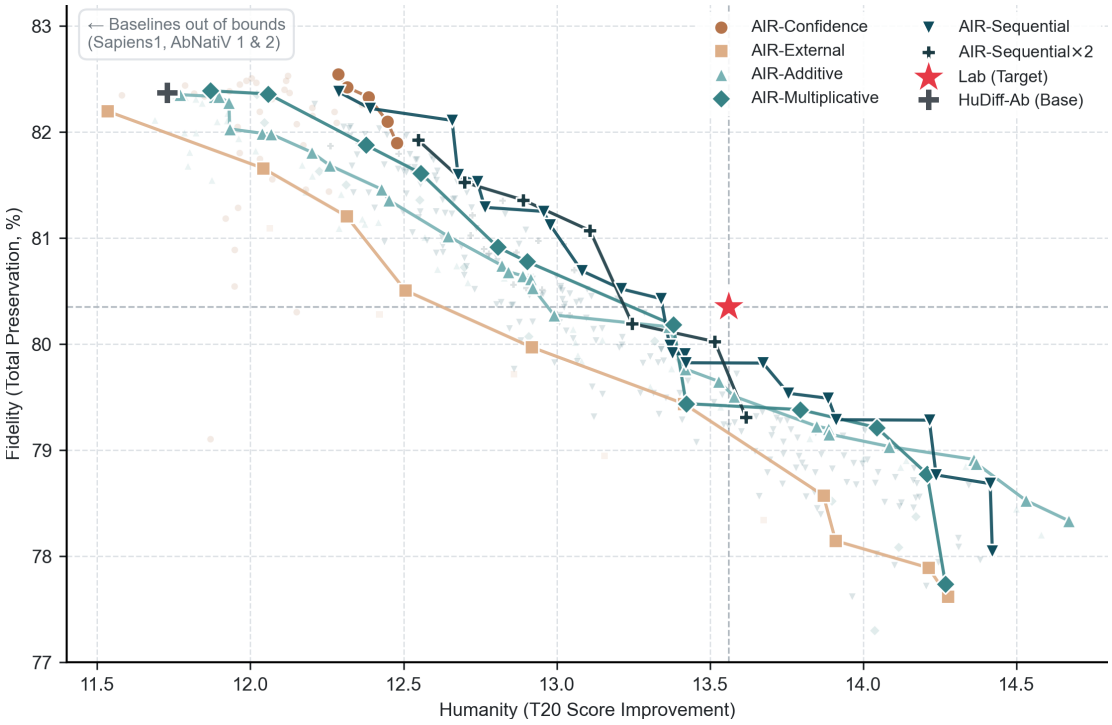

Figure 5. Full hyperparameter sweep on Humab25. Each marker is one configuration, averaged over seeds. Translucent markers are non-Pareto configurations and the connected solid markers are the per-variant Pareto fronts. Anchor points (Lab, HuDiff-Ab, Sapiens1, AbNatiV, AbNatiV2) are shown for orientation. Dotted lines mark Lab T20 and Lab Total Preservation.

C. Per-chain breakdown

Table 4 reports the Humab25 results from Table 1 split into heavy-chain (VH) and light-chain (VL) values. The two chains differ in germline diversity and in CDR-loop length, and humanization difficulty is not symmetric across them.

The HuDiff-Ab base model reaches T20 of 13.98 on VH but only 9.47 on VL, indicating that the light chain is harder to humanize from this starting point under the same model and protocol. AIR yields gains on both chains and the gap between them narrows. For AIR-Sequential, T20 increases by 1.69 on VH (13.98 $\rightarrow$ 15.67) and 1.54 on VL (9.47 $\rightarrow$ 11.01). Total Preservation is roughly balanced across chains for all AIR variants, in the range 80–82%. Vernier Preservation remains markedly higher on VL ($\sim$90%) than on VH ($\sim$80%) across all methods, including Lab, reflecting that the VL Vernier zone is comparatively conserved under any humanization strategy.

T20 is not reported per chain for Lab and Sapiens1 since the HuDiff paper publishes only the aggregate value for these baselines.

Table 4. Per-chain breakdown of the Humab25 results. Mean $\pm$ std across seeds where applicable. T20 is the T20 score improvement over the parental sequence. *Total Pres.* is total preservation. *Vernier Pres.* is Vernier-zone preservation. AIR variants use the configurations selected for Table 1.

Method	Chain	T20 ↑	Germline ID (%)	Total Pres. (%)	Vernier Pres. (%)
Lab (Ma et al., 2024)	VH	–	89.36	78.34	79.25
	VL	–	94.47	82.35	93.14
Sapiens1 (Prihoda et al., 2022; Ma et al., 2024)	VH	–	84.61	88.54	86.00
	VL	–	89.75	88.33	94.00
AbNatiV (Ramon et al., 2024)	VH	–	–	–	–
	VL	–	–	–	–
AbNatiV2 (Ramon et al., 2025)	VH	–	–	–	–
	VL	–	–	–	–
HuDiff-Ab (Base) (Ma et al., 2024)	VH	13.98$\pm$0.19	90.52$\pm$0.25	82.30$\pm$0.15	79.55$\pm$0.51
	VL	9.47$\pm$0.31	88.33$\pm$0.27	82.44$\pm$0.15	91.77$\pm$1.13
AIR-Confidence	VH	14.84$\pm$0.19	91.74$\pm$0.24	81.92$\pm$0.17	80.60$\pm$0.24
	VL	10.12$\pm$0.34	89.15$\pm$0.33	81.87$\pm$0.18	91.60$\pm$0.51
AIR-External	VH	15.17$\pm$0.29	92.18$\pm$0.39	80.26$\pm$0.17	80.42$\pm$0.29
	VL	9.84$\pm$0.29	89.05$\pm$0.38	80.75$\pm$0.13	90.48$\pm$0.72
AIR-Additive	VH	15.41$\pm$0.37	92.60$\pm$0.49	80.21$\pm$0.29	80.42$\pm$0.38
	VL	10.43$\pm$0.45	89.56$\pm$0.23	80.86$\pm$0.31	90.79$\pm$1.15
AIR-Multiplicative	VH	15.52$\pm$0.21	92.56$\pm$0.32	80.51$\pm$0.23	79.43$\pm$0.65
	VL	10.28$\pm$0.23	89.42$\pm$0.45	81.05$\pm$0.19	90.57$\pm$0.57
AIR-Sequential	VH	**15.67$\pm$0.05**	92.85$\pm$0.09	80.04$\pm$0.35	80.00$\pm$0.87
	VL	**11.01$\pm$0.25**	90.19$\pm$0.29	80.82$\pm$0.15	90.67$\pm$0.92
AIR-Sequential$\times$2	VH	15.37$\pm$0.26	92.47$\pm$0.33	80.71$\pm$0.17	80.81$\pm$0.38
	VL	10.85$\pm$0.24	90.16$\pm$0.21	81.43$\pm$0.16	90.29$\pm$0.66

D. Extended ablations

This appendix reports sensitivity analyses on the main hyperparameters of AIR. Each analysis varies one factor while holding the others at the values from Appendix B.

D.1. Cycle count and remasking rate

Figure 6 traces each criterion through (K, p) space. For *Confidence*, T20 rises slowly with K and Total Preservation stays close to the base model across the swept range. The full *Confidence* sweep extends to $K = 400$, with mean T20 climbing from 11.58 at $K = 1$ to 12.48 at $K = 400$ at $p = 0.01$, and Total Preservation in a narrow band of 81.9 to 82.5. The other four criteria trace coherent curves in $(T20, \text{Total Preservation})$ space. At fixed p, increasing K buys T20 at a roughly constant rate of Total Preservation. At fixed K, increasing p moves the operating point along the same direction. The two parameters are therefore largely interchangeable in their effect on the position along the trade-off curve, which is the property that lets a single criterion span the Pareto front of Figure 2.

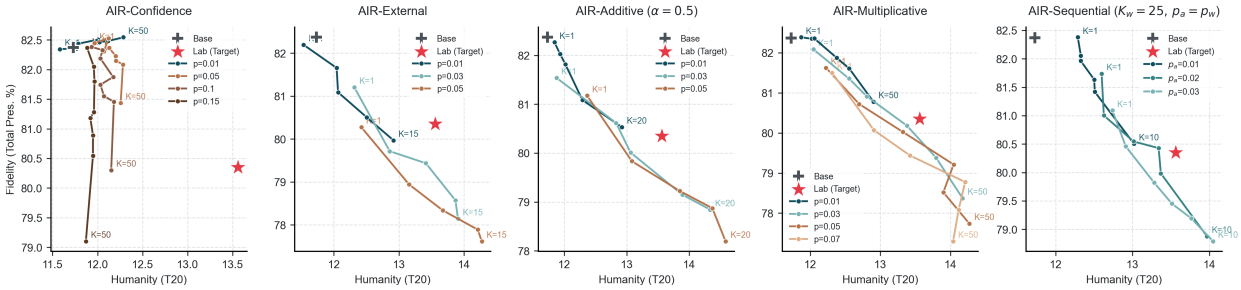

Figure 6. Cycle count and remasking rate sensitivity per criterion. Each curve traces K at fixed p, with the smallest and largest K values along each curve labelled. *Confidence* is shown over $K \leq 50$ for visibility (the full sweep extends to $K = 400$). The *Sequential* panel varies K_a at fixed $K_w = 25$ and $p_a = p_w$. Black markers are Lab and HuDiff-Ab (Base) for reference.

D.2. Proximity protection on AIR-Multiplicative

Section 4.3 reports proximity protection for AIR-External. Figure 7 applies the same protection mechanism to AIR-Multiplicative. The pattern is the same. Protection shifts the curve modestly upward in Total Preservation, with the gap most visible at the higher p values that drive the unprotected baseline further into the high-T20 region. At $p = 0.05$, protection gains 0.4 to 0.8 percentage points of Total Preservation across K, at a cost of 0.0 to 0.6 T20. At $p = 0.01$ the protected and unprotected curves overlap within seed variance. The effect is monotonic in the same direction observed for AIR-External and is of comparable magnitude.

D.3. Sequential second-stage rate

The *Sequential* variant uses a low remasking rate in its second stage. We sweep $p_w \in \{0.01, 0.02, 0.03\}$ at $K_w \in \{25, 50, 100\}$ while holding the first stage at the recommended $K_a = 4$, $p_a = 0.02$. The headline configuration ($K_w = 25$, $p_w = 0.02$) reaches T20 $= 13.34$ at Total Preservation 80.43. At the same K_w, reducing p_w to 0.01 drops T20 to 12.86 at similar Total Preservation, and increasing p_w to 0.03 matches the headline T20 within seed variance at slightly lower Total Preservation. At $K_w \in \{50, 100\}$ the choice of p_w has a smaller effect, with all three values clustered near the headline operating point. The picture is consistent with the second stage saturating at moderate combinations of K_w and p_w. The choice of p_w matters mainly when K_w is small.

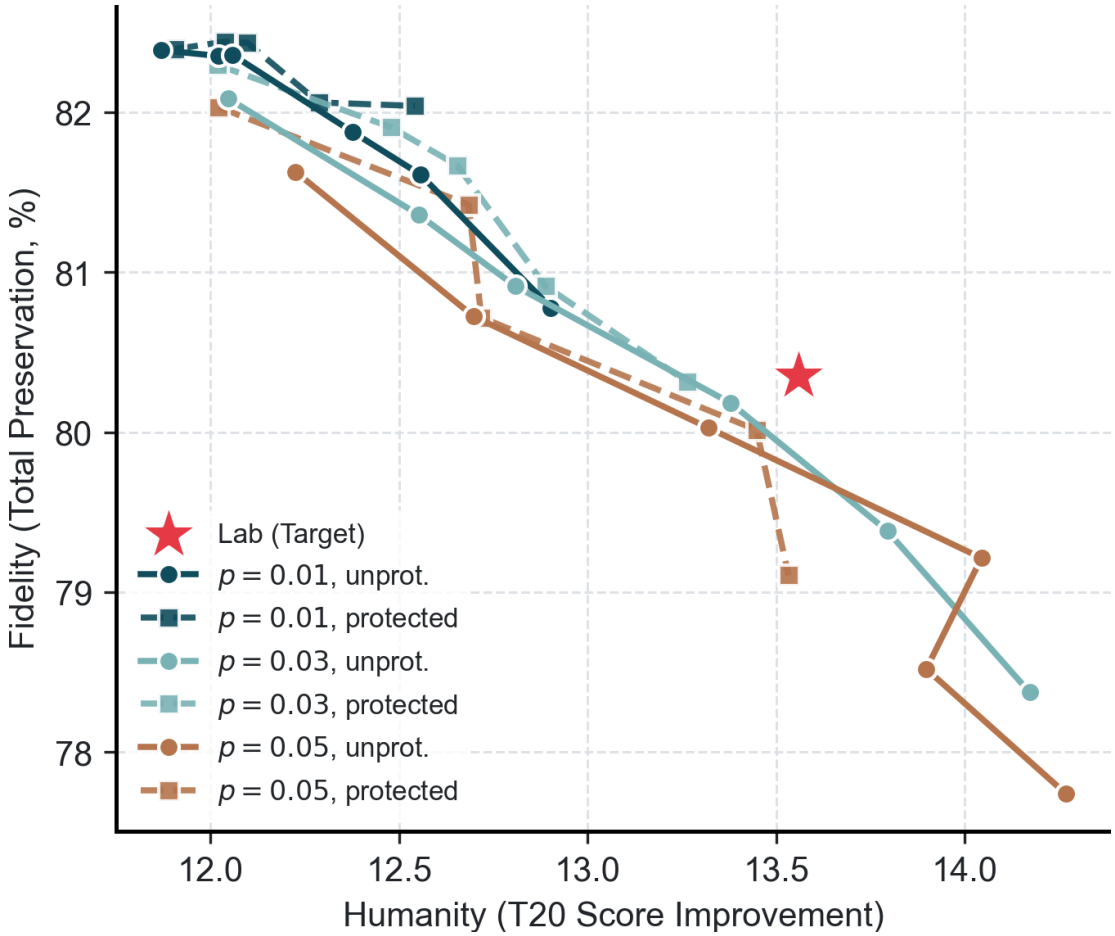

Figure 7. Proximity protection on AIR-Multiplicative. For each remasking rate p, solid curves are runs without protection and dashed curves use protection. K increases left to right along each curve. Five seeds per cell.