TxBench–Antibody Development

Benchmarking AI agents on the decisions that take a therapeutic antibody from lead nomination to the first human dose

Joshua Badshah[1], Arjun Banerjee[1], Yuyi Liu[1], Anders Eliasen[1], Ashika Jain[1], Tim Proctor[1], Dillon Flood[1], Kenny Workman[1]

[1]LatchBio, San Francisco, CA, USA

Correspondence: kenny@latch.bio

Abstract

Once a therapeutic antibody is nominated, its development depends on a series of decisions: whether the molecule is developable, whether a batch meets specification, how fast it clears, whether it will provoke anti-drug antibodies (ADA), what the toxicology allows and what dose is safe to give first. AI agents are increasingly used to support these decisions, but no benchmark tests whether they can make them from the underlying records. We introduce TxBench–Antibody Development, a verifiable benchmark of 60 evaluations spanning the seven stages from lead nomination to the first-in-human (FIH) dose. Each evaluation is built from patents, regulatory reviews, trial protocols or published studies, asks for a decision with the quantities that support it, and is graded deterministically. Across 34 model–harness configurations (21 models, three attempts each), the highest pass rate was 39.4% (Grok 4.7 / Grok Build), followed by Opus 5 / Claude Code (37.2%) and Sonnet 5.5 / PI (36.7%). Six models led at least one of the 20 categories, and no failure class had a mean pass rate above 35%.

Introduction

Between lead nomination and an investigational new drug (IND) filing, decisions about an antibody program are made by specialists in separate functions. A developability scientist reads size-exclusion chromatography (SEC) monomer, differential scanning fluorimetry (DSF) melting temperatures, hydrophobic interaction chromatography (HIC) retention and polyspecificity for each candidate, and compares them with values for clinical-stage antibodies [1, 2]. A pharmacokineticist reads mouse and monkey serum profiles, where neonatal Fc receptor (FcRn) recycling and target-mediated drug disposition (TMDD) set clearance [3]. A toxicologist sets the no-observed-adverse-effect level (NOAEL), and a clinical pharmacologist integrates these data into a first dose.

Our group has released benchmarks that grade whether an agent recovers the decision the evidence supports: TxBench–PP for small-molecule preclinical pharmacology [4], TxBench–Antibody Discovery for the discovery phase of biologics [5] and OncologyBench for multiagent oncology decisions [6]. Others have benchmarked prediction and design, from the Therapeutics Data Commons [7] and the antibody fitness benchmark FLAb [8] to the prospective AIntibody design challenge [9]. None of these asks an agent to make the development decisions themselves. With TxBench–Antibody Development we asked how well current agents make these decisions (Figure 1), whether one model ranking holds across them, and where the failures come from.

Benchmark Construction

The evaluations follow the seven stages of antibody development (Figure 2A): nomination and developability triage (12 evaluations; SEC, DSF and nanoDSF, HIC, polyspecificity), CMC and expression liabilities (5; titer, scale-up by-products, formulation factorials, clipping by LC-MS), glycosylation and analytical QC (5; glycan profiles, FcγRIIIa binding, multi-laboratory potency, release trending against specification [10]), PK, FcRn and TMDD (8), immunogenicity and deimmunization (8; humanized variant panels and clinical ADA incidence [11]), toxicology and safety (10; cytokine-release margins, effector-function hazards, adverse-event tables) and the FIH dose (12; minimal anticipated biological effect level (MABEL)-, NOAEL- and highest non-severely toxic dose (HNSTD)-based limits [12, 13, 14], including MABEL for T-cell engagers [15]). The molecules are IgG antibodies, bispecific T-cell engagers, antibody–drug conjugates, Fc fusions and single-domain antibodies. Each evaluation also carries one decision type, named as in TxBench–Antibody Discovery so the two can be compared (Figure 2C), one main evidence source (Figure 2D) and one primary failure class from its main trap [5].

The records come from 41 patents, 11 regulatory review documents (PMDA and FDA reviews and European public assessment reports), four published studies, three trial protocols or registry entries and one public dataset (Figure 2B). We removed identifiers that could reveal the answer and kept the information that determines it: comparators, replicates, batches, donors, dose and route. Every plausible incorrect approach is contradicted by data available to the agent. A fixed three-fold affinity bar on the parent's single binding reading, for example, would drop a humanized variant whose function in the same run is intact, and the potency tables beside it show this. Graders match categorical answers exactly and numeric answers against admissible ranges, and a run passes only if every check passes. Each configuration made three offline attempts per evaluation under the PI, Claude Code, Codex or Grok Build harness.

Results

Leaderboard and efficiency

Grok 4.7 / Grok Build passed 39.4% of attempts (95% CI 28.6–50.3), Opus 5 / Claude Code 37.2% and Sonnet 5.5 / PI 36.7%; the weakest configuration, GPT-5.6 Luna / Codex, passed 2.8% (Figure 1A). The low ceiling does not reflect unsolvable tasks: 57 of the 60 evaluations were solved by at least one configuration, 48 by models from two or more labs, and 45 were passed in all three attempts by some configuration.

The harness mattered for few models. Opus 5 gained 7.2 points with Claude Code over PI (95% CI 1.3–13.2) and GPT-5.6 Sol lost 6.7 points with Codex; for the other eleven models run under two harnesses the interval included zero. Higher cost did not predict higher accuracy (Figure 1B–E). GPT-6 Astra / PI passed 32.8% of attempts at $0.46 and 52 thousand tokens per run, the same rate as Opus 5.5 / Claude Code at $3.47 and 764 thousand tokens.

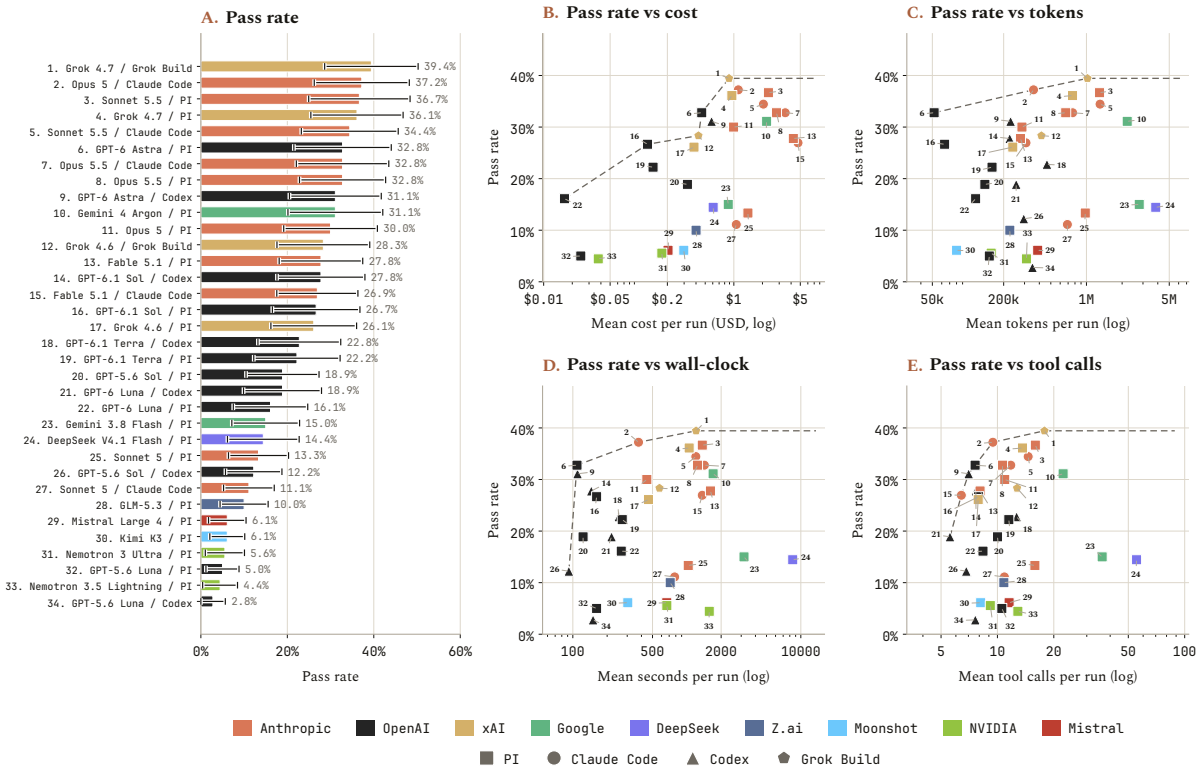


Figure 1: Pass rate by configuration. (A) Mean pass rate over the 60 evaluations, three attempts each, with 95% Student's t intervals taken over per-evaluation pass rates; colour marks the model's lab. (B–E) Pass rate against mean cost, tokens, wall-clock time and tool calls per run, on log scales. The dashed line marks the efficiency frontier.

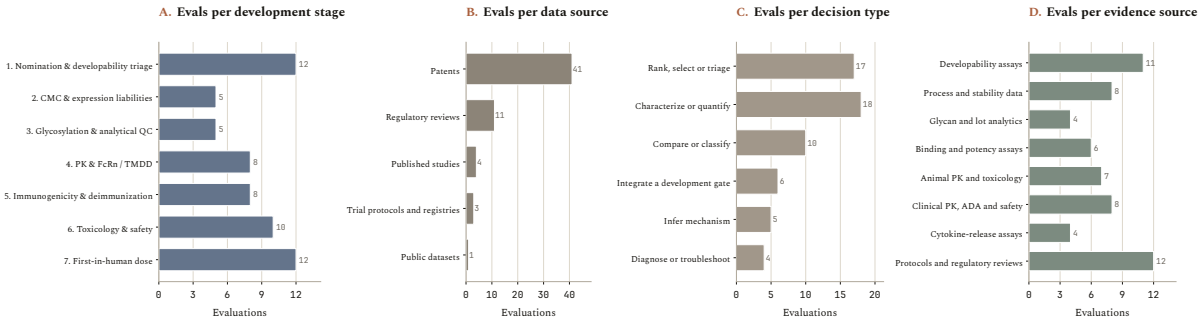


Figure 2: Benchmark anatomy. (A) Evaluations per development stage, from lead nomination to the first-in-human dose. (B) Evaluations per data source. (C) Evaluations per decision type, using the class names of TxBench–Antibody Discovery. (D) Evaluations per main evidence source. Each evaluation has one label on each axis.

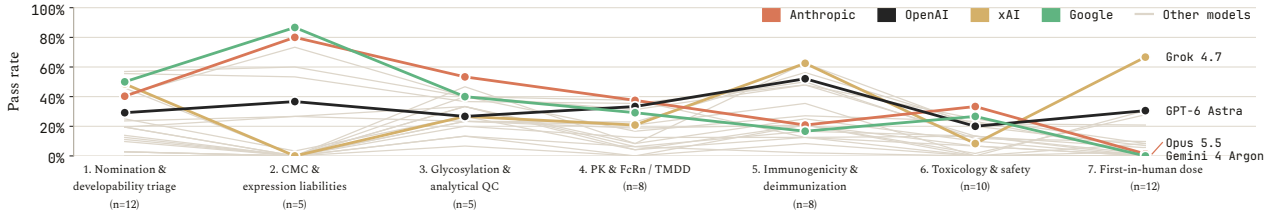


Figure 3: Pass rate across development stages. Mean pass rate per stage for one flagship model from each lab, in colour, and every other model, in grey; both harnesses pooled and averaged at the evaluation level.

Performance by category

Pass rates fell from the molecule stages to the clinical ones. Pooled over configurations, glycosylation and analytical QC reached 30.6% and immunogenicity 33.3%, while toxicology and the FIH dose reached 13% (Figure 3). The flagship models cross each other between almost every pair of stages.

Figure 4 gives the full breakdown by decision type, evidence source and failure class, with the leader of each column outlined. Six models lead at least one of the 20 columns. Grok 4.7, the overall leader, leads six; Opus 5.5, ranked fourth overall, leads seven and is also the weakest of the top ten models in two others, binding and potency assays (3%) and evidence integration (2%). Only Opus 5.5 passed any cytokine-release evaluation (17%); the other models used a donor average or the first-listed cytokine instead of the most sensitive donor, or read the EC50 as the minimal effect level.

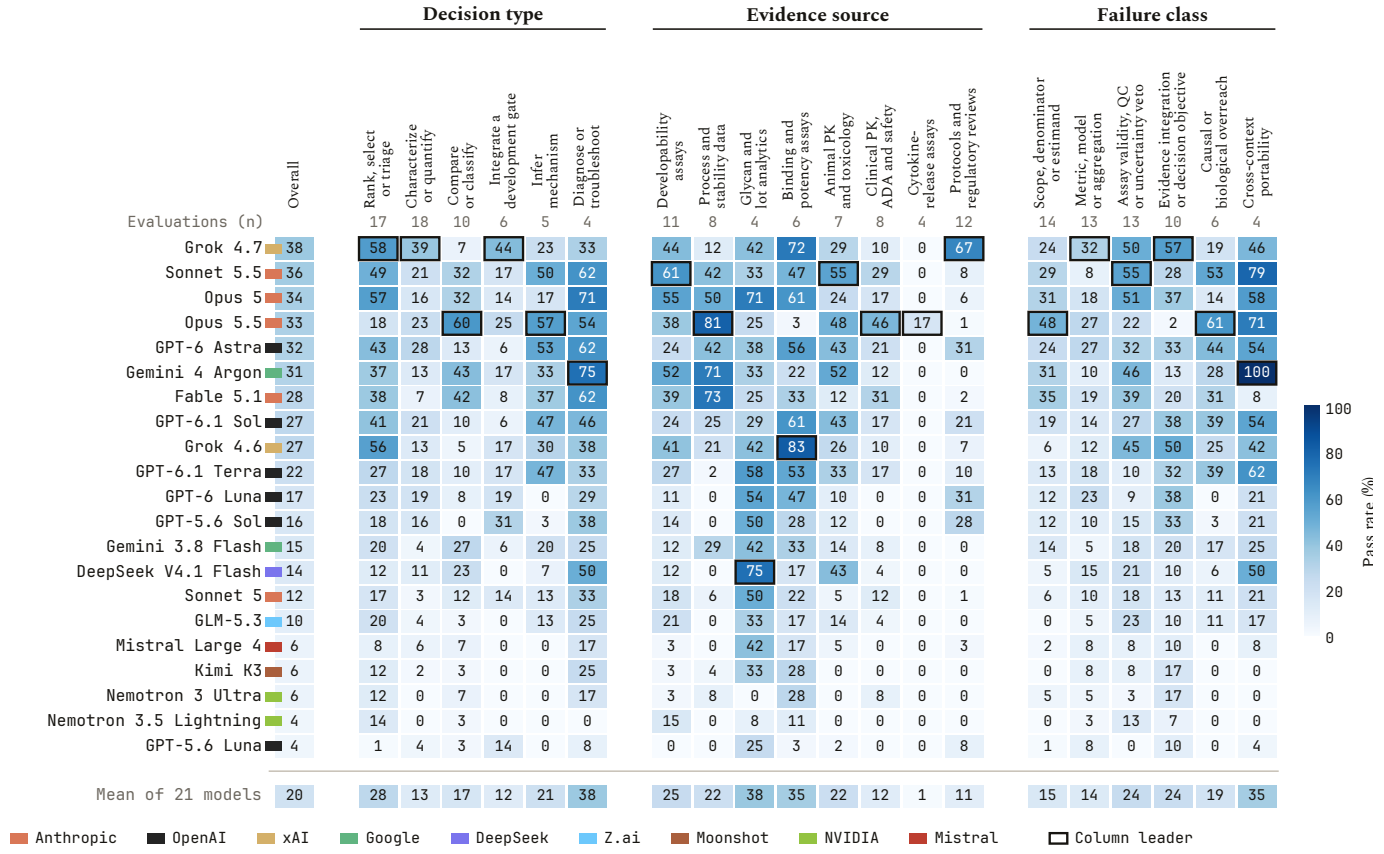


Figure 4: Pass rate by decision type, evidence source and failure class. All 21 models, both harnesses pooled, ordered by overall pass rate. Each evaluation carries one label on each axis, so each block sums to the 60 evaluations. The outlined cell is the leader of each column and the bottom row is the mean over models.

Failure decomposition

Five of the six failure classes concern interpretation: the estimand or denominator (14 evaluations), the validity of an assay, control or single comparator reading (13), the integration of evidence into the decision objective (10), a cause assigned from a confounded comparison (6) and a reference bar transferred from another context (4). One class is computational, a wrong metric, model or aggregation (13): a dose-scaling or unit slip, a one-time offset fitted as a trend, or a best-fitting surrogate extrapolated to a withheld outcome.

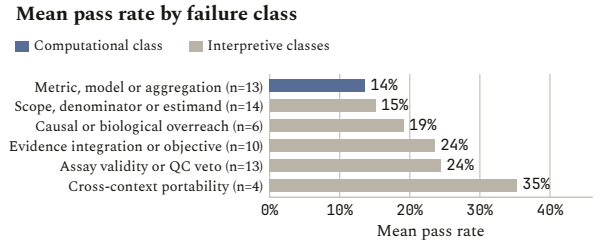


Figure 5: Mean pass rate by failure class. Mean over the 21 models of each model's pass rate on the evaluations in the class, both harnesses pooled; n is the number of evaluations in the class.

Because classes are assigned per evaluation, the share of failed attempts in each class tracks the number of evaluations in it, so we report the pass rate per class (Figure 5). No class was easy. The computational class was the hardest, with a mean pass rate of 14%, followed by the estimand class at 15%; cross-context portability, at 35%, was the easiest. Which classes a model failed differed as much as its overall rate (Figure 4).

A five-review development chain

Because development decisions build on each other, we also built one program as a chain, a separate evaluation outside the 60. Four candidates from two humanized binder families of a bispecific immune agonist pass through five reviews in one conversation, from developability to the FIH dose (Figure 6A). Each review carries its own trap, from a HIC limit set from this format's own values to a starting dose that must be set before the missing IND studies. Each review was also run on its own from the correct earlier answers.

Only Opus 5.5 completed the chain, in one of five runs (Figure 6B). Grok 4.7 and GPT-6 Astra failed the first review in every run, and Gemini 4 Argon, which always passed it on its own, passed it once inside the chain. Only Opus 5.5 read one candidate's low first PK sample as a dosing artifact rather than fast clearance (13 of 13 runs, against 0 of 36 for the other models), and it used that finding to pass the safety and nomination review in two chain runs, a review that no model passed on its own (Figure 6C).

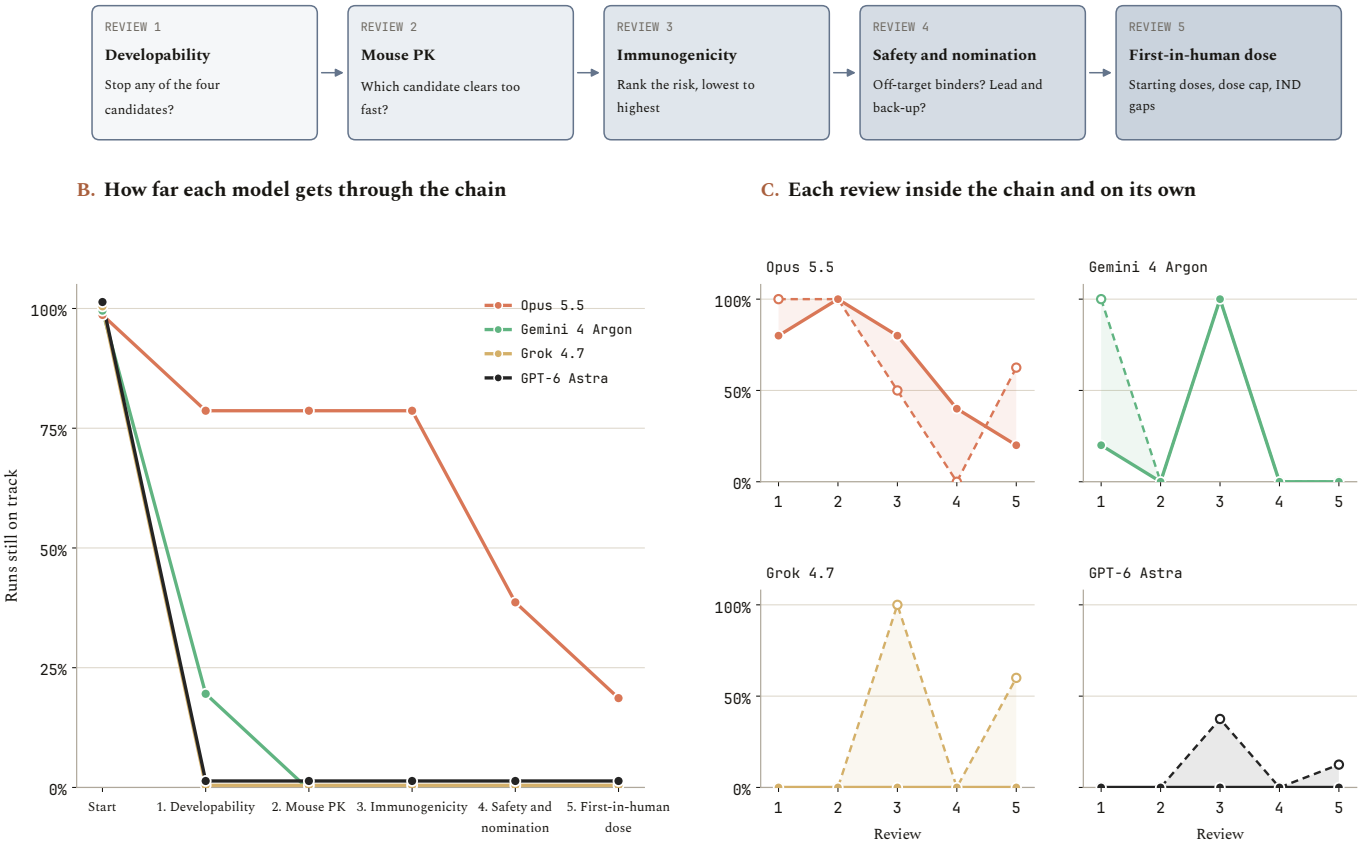


Figure 6: A five-review development chain. (A) The five reviews, each asked in turn in one conversation about the same four candidates. (B) Share of the five runs per model still on track after each review. (C) Pass rate of each review inside the chain (solid), where the model builds on its own earlier answers, and of the same review run on its own from the correct earlier answers (dashed; five to eight runs per model).

Discussion

Current agents are not yet reliable development decision-makers. Almost every evaluation was solved by some configuration, so the shortfall reflects judgment on particular kinds of evidence rather than task difficulty. Six models led at least one category and the leader changed between stages, so a single aggregate score obscures where each model is reliable. The two hardest failure classes, a wrong estimand or denominator and a wrong metric or aggregation, are also two of the three largest, which suggests that harnesses should check the comparator, the estimand and the decision rule before committing to an answer, rather than extending the run. In the chain, building on earlier answers propagated early errors and, for one model, an early correct finding.

With the known limitations of our work, we believe TxBench–Antibody Development is a useful measure of how far current agents are from making these decisions reliably. The records are public and retrospective, three attempts per evaluation make single cells indicative, the failure classes are assigned per evaluation rather than per attempt, and the chain is one program with five runs per model. Next steps are expert baselines and evaluations built from unpublished programs.

References

1. Jain T, Sun T, Durand S, et al. Biophysical properties of the clinical-stage antibody landscape. Proc Natl Acad Sci USA 114:944–949, 2017.
2. Raybould MIJ, Marks C, Krawczyk K, et al. Five computational developability guidelines for therapeutic antibody profiling. Proc Natl Acad Sci USA 116:4025–4030, 2019.
3. Roopenian DC, Akilesh S. FcRn: the neonatal Fc receptor comes of age. Nat Rev Immunol 7:715–725, 2007.
4. Le H, Ramasamy R, Urrutia A, Yazdani M, Proctor T, Workman K. TxBench-PP: Analyzing AI agent performance on small-molecule preclinical pharmacology. arXiv [cs.AI], 2026.
5. Ramasamy R, Le H, Brougher J, Urrutia A, Banerjee A, Poust S, Workman K. TxBench–Antibody Discovery: Benchmarking AI agents on experimentally grounded decisions in therapeutic antibody discovery. benchmarks.bio, 2026.
6. Badshah J, Saye BK, Banerjee A, Shrishrimal S, Ramasamy R, Flood D, Workman K. OncologyBench: Benchmarking multiagent AI systems on grounded systems in oncology research and treatment. latch.bio, 2026.
7. Huang K, Fu T, Gao W, et al. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. NeurIPS Datasets and Benchmarks, 2021.
8. Chungyoun M, Ruffolo J, Gray JJ. FLAb: Benchmarking deep learning methods for antibody fitness prediction. bioRxiv 2024.01.13.575504, 2024.
9. Erasmus MF, Spector L, Ferrara F, et al. AIntibody: an experimentally validated in silico antibody discovery design challenge. Nat Biotechnol 42:1637–1642, 2024.
10. ICH. Q6B: Specifications: Test procedures and acceptance criteria for biotechnological/biological products, 1999.
11. Shankar G, Arkin S, Cocea L, et al. Assessment and reporting of the clinical immunogenicity of therapeutic proteins and peptides: harmonized terminology and tactical recommendations. AAPS J 16:658–673, 2014.
12. European Medicines Agency. Guideline on strategies to identify and mitigate risks for first-in-human and early clinical trials with investigational medicinal products. EMEA/CHMP/SWP/28367/07 Rev. 1, 2017.
13. US Food and Drug Administration. Guidance for industry: Estimating the maximum safe starting dose in initial clinical trials for therapeutics in adult healthy volunteers, 2005.
14. ICH. S9: Nonclinical evaluation for anticancer pharmaceuticals, 2009.
15. Saber H, Gudi R, Manning M, Wearne E, Leighton JK. An FDA oncology analysis of immune activating products and first-in-human dose selection. Regul Toxicol Pharmacol 81:448–456, 2016.