Arc Science
Technical Report III

Judging, Specificity, Clinical Liabilities, Portfolio Selection, and Evaluation Protocol

Part III of the Omniverso Specification (Decision Stack and Supplementary Section S3)

Danil Kotelnikov

Abstract. Pre-clinical candidate selection can fail through polyreactivity, immunogenic peptides, sequence liabilities, or post-selection bias. This volume specifies complex judging, self-proteome specificity panels, clinical liability bands, and cautious scoring inside a Kullback–Leibler ball. It defines fixed-size portfolio selection committed before scores are read, conformal risk control, and power calculations. Supplementary Section S3 provides the 53-gate falsification registry. The report covers the design and its verification gates.

Contents

1 Judging and specificity **3**

1.1 The signed panel and pair-specific losses 5

1.2 Thresholds and conformal risk control 6

2 Liability, immunogenicity and glycan channels **7**

3 Decisions under ambiguity and portfolio selection **10**

3.1 Fixed-size portfolio 12

4 Evaluation design **14**

4.1 Reporting for the base model 14

4.2 Comparison with published numbers 15

4.3 Splits and statistical rules for the full system 16

4.4 Hypotheses, ablations and validation levels 18

5 Evaluation protocol **19**

6 Scope, Ethics, Licence, and Declarations **24**

6.1 Scope, Dual Use, and Licensing 24

6.2 Declarations 26

1. Judging and specificity

The judging layer of the full system has two parts. The complex judge proposes and ranks antibody-antigen poses, which is a question of geometry. The specificity instrument measures cross-reactivity against a named panel of other proteins. Table 3 gives the contract of the complex judge. The co-folding models and physics programs named below are published systems that the design calls as external tools. Their outputs sit in separate fields, and no field is converted into another.

Input and output. The judge reads a candidate, a protein antigen record, a declared epitope and an antigen-state profile. It returns an ensemble of proposed complexes, one selected pose and the registration of that pose against the epitope. The profiles cover a VHH or a paired Fv against one protein state, several registered antigen conformers with one report per state, and a peptide antigen. An antigen residue is identified by six items: structure checksum, model, chain, author residue number, insertion code and a normalised index. A chain letter and an integer are insufficient, because a revised structure or an insertion code can move the atoms that a number points to. The epitope uses the same identifier. A judge version binds the weights, the template cut-off, the alignment databases, the sampling settings, the seed algorithm and the numerical precision. The native complex of a test case is excluded from the antibody templates.

Seed ensembles and clusters. For candidate i and a roster of S registered seeds, the co-folder returns poses x_{ij}, $j = 1, \dots, S$. The number S is fixed for a profile, and a different S needs a new calibration. Seed replicates measure the variability of the algorithm for one pair. They carry no information on binding, which needs independent candidates with measured endpoints. In the benchmark reviewed in the Introduction, more samples per target improved structure and left the separation of cognate from shuffled pairs unchanged [1] (Technical Report I (Section "Evaluation, specificity and developability")), so the design spends compute on seeds and negative controls. The comparator is $S = 15$, and qualification selects the smallest fixed S that meets the registered pose margin at matched compute.

Poses are clustered by interface RMSD after antigen alignment, with the algorithm, cut-off, linkage, atom mask and handling of missing residues fixed in advance. Write $C_{i1}, \dots, C_{im}$ for the clusters of candidate i. The largest cluster $C_{ic_i^*}$ is selected, with ties broken by the best reranker score and then the smallest seed index. The pose is a seed majority when $|C_{ic_i^*}| > S/2$. Otherwise the pose is called multimodal, and the largest cluster supplies a descriptive representative that no downstream gate treats as a majority.

Reranking. A reranker r_η picks a pose inside the selected cluster. It trains on reference complexes, where each sampled pose has a geometric label Y such as DockQ, a continuous score from 0 to 1 that combines the fraction of native contacts with two RMSD terms [2]. Training pairs (u, v) come from one target and satisfy $Y_u > Y_v$. The loss rewards a higher score for the better pose of every pair:

$$\mathcal{L}_{\text{rank}} = -\sum_{t} \sum_{(u,v)\in\mathcal{O}_t} w_{uv} \log \sigma\big(r_\eta(x_u) - r_\eta(x_v)\big).$$

Here t runs over targets, $\mathcal{O}_t$ is the set of ordered pose pairs of target t, w_{uv} is a pair weight and σ is the logistic function. The reranker learns the order of seeds for one antibody-antigen pair. Ordering different candidates by binding is a separate task that needs assayed candidates

under one protocol. The emitted pose score is a frozen monotone readout of the reranker and the ensemble features, with a model-specific scale. Search quality (does the ensemble hold a good pose) and ranking quality (does the rule pick it) are reported separately, because benchmarks show a gap between them. The internal confidence of a co-folder often fails to identify the best sampled model, and the reported success on recent antibody and VHH complexes is low [3, 4] (Technical Report I (Section "Evaluation, specificity and developability")).

Epitope registration and state margin. Registration reuses P_E, R_E and EpiF1 of Technical Report II (Section "Antigen conditioning and Interface-JEPA"), computed on the selected pose. They describe placement against a declared patch. With a native complex the report adds interface and ligand RMSD, the fraction of native contacts and DockQ. For antigen conformers the judge runs the same seeds on each and reports a state margin, the pose score on the intended state minus the best score on any other state. A margin describes geometry. Selectivity between states needs assays on both.

Admission error and yield. A pose gate admits some candidates, and an external endpoint later succeeds or fails for each. Let $A_i = 1$ when candidate i passes the gate, $Y_i = 1$ when the registered endpoint succeeds and $B_i = 1 - Y_i$. Over n candidates the report gives the bad-admission incidence, the admission yield, the bad share among admitted candidates and the good yield:

$$\widehat{\mathrm{BAI}} = \frac{1}{n}\sum_i A_i B_i, \quad \widehat{y}_{\mathrm{admit}} = \frac{1}{n}\sum_i A_i, \quad \widehat{\pi}_{\mathrm{bad|admit}} = \frac{\sum_i A_i B_i}{\sum_i A_i}, \quad \widehat{y}_{\mathrm{good}} = \frac{1}{n}\sum_i A_i Y_i.$$

The third quantity is undefined when nothing is admitted. The four are reported together, because the incidence alone reaches zero for a gate that admits nothing. The cohort is an untouched set from a frozen generator, steering policy, triage rule and judge profile, and the endpoint Y_i is an assay outcome. Acceptance limits for each judge component are listed in Supplementary Section S3.

Co-folders and physics tools. The judge draws on three published co-folders. AlphaFold 3 predicts the joint structure of proteins, nucleic acids, small molecules and ions with a diffusion-based network [5]. Boltz-2 is a structure and affinity model with open weights, and its affinity head is described for small-molecule binders, so the judge uses it for geometry only [6]. Chai-1 predicts complexes with or without sequence alignments [7]. Boltz-2 is the primary co-folder and Chai-1 adds a second model family. Antibody chains enter with empty alignments and the antigen reads a precomputed alignment. Each invocation draws one diffusion sample with 3 recycling steps and 200 sampling steps. Co-folders from different families are not independent judges, because their training sets overlap, so the design measures their dependence.

Physics tools run as a cascade on a shortlist, and each tier has an admitting rule and its own endpoint type (Table 1). Short molecular dynamics with OpenMM [8] tests whether the pose is retained. MM-GBSA gives a protocol-bound end-state energy rank [9], and Flex ddG ranks mutations from a backrub ensemble in Rosetta [10]. Well-tempered metadynamics [11] gives a one-dimensional free-energy profile along the interface distance, which is read as a rank, and a rank is not a binding free energy. None of the four outputs enters a field for pose, pair binding or affinity. With S seeds over M survivors the campaign needs $S \cdot M$ co-fold inferences, and a pre-triage by the complex head of Interface-JEPA (Technical Report II (Section "Antigen conditioning and Interface-JEPA")) could reduce this number.

Table 1: Tiers of the physics cascade. Distances are in angstroms and times in nanoseconds. The slope bound is a one-sided 95% moving-block bootstrap bound.

Tier	Settings	Admitting rule
Co-fold ensemble	15 seeds	Every survivor
Reranked shortlist	Ensemble poses ranked by the reranker	Top decile of reranker score and cross-tool ipTM agreement within 0.15
Short MD	5 ns at 300 K, 4 fs step, first 1 ns discarded	At least 20 interface heavy atoms
MM-GBSA	Protocol-bound end-state energy rank	Mean interface RMSD below 3.0 Å and upper bound of the Theil-Sen slope at most 0.25 Å per ns
Flex ddG	Backrub ensemble	MM-GBSA rank in the top third and a mutation question from the maturation loop
Metadynamics	50 ns, well-tempered	Top 8 by rank

1.1. The signed panel and pair-specific losses

Panel. Specificity is relative to a panel, and the instrument reports risk for a named panel only. For a candidate X with intended antigen A^+, the signed panel is $\Pi = \Pi_{\text{req}} \dot{\cup} \Pi_{\text{forbid}}$, the disjoint union of required members, which the binder should bind (for example orthologs of the pharmacology species), and forbidden members, which it should not bind (for example human paralogs). The panel is built from the campaign and the target and never from the candidate. Its inputs are the indication, route, species, target family, reference proteome, tissue-exposure rules, isoform policy and available structures. The campaign profile sets the tiers: for systemic use the forbidden tiers hold accessible human paralogs and membrane, extracellular and secreted self proteins, and for a reagent antibody they hold proteins of the sample matrix and common homologs. Each member records its sign and tier with the rationale, its identifier, release, isoform and sequence checksum, the source of its structure, its exposure class, the rule that builds its surface patch, its reference mass ν_k^{panel} and, when it is sampled, its sampling probability q_k. Linear peptides, folded proteins and computed surface patches test different interaction modes and keep different tier labels. A peptide-negative result holds for the linear tier only.

Readouts. Three readouts are kept apart. The pair-binding probability $o_{\text{pair}}(X, \pi)$ belongs to a candidate and one named panel member. The pose score belongs to a proposed pair and measures geometric plausibility. The generic polyspecificity $o_{\text{PSR}}(X)$ belongs to the candidate alone: it is the response to a polyspecificity reagent, a detergent-solubilised insect-cell membrane preparation in the nanobody assay [12, 13]. It flags broadly sticky candidates and cannot say whether a given paralog is bound. Co-folder confidence does not estimate a pair-binding probability. In the nanobody benchmark reviewed in the Introduction, ipTM separated cognate from shuffled pairs only weakly [1] (Technical Report I (Section "Evaluation, specificity and developability")). An admissible pair oracle therefore consumes both X and π, predicts a named binding endpoint and is calibrated on held-out pair labels.

Losses. Let $p_k = o_{\text{pair}}(X, \pi_k) \in [0, 1]$ be a calibrated pair-binding probability under one assay family. The forbidden-hit mass is the weight of forbidden members predicted to bind above a threshold λ_f. The required-miss mass is the weight of required members predicted to bind at or

below a threshold λ_r:

$$\ell_f(X) = \sum_{k \in \Pi_{\text{forbid}}} w_k^f \mathbb{1}[p_k > \lambda_f], \qquad \ell_r(X) = \sum_{k \in \Pi_{\text{req}}} w_k^r \mathbb{1}[p_k \le \lambda_r].$$

The weights are non-negative and sum to one within each sign, and $\mathbb{1}[\cdot]$ is 1 when its condition holds and 0 otherwise. The hit mass falls as λ_f rises and the miss mass rises with λ_r, so the thresholds are separate. The two numbers keep off-target binding apart from missing intended breadth. When members are sampled with probability q_k to stand for a reference measure with masses ν_k^{panel}, each carries the importance ratio ν_k^{panel}/q_k, and the estimate is the ratio-weighted mean of the indicators. Support ($q_k > 0$ wherever $\nu_k^{\text{panel}} > 0$) is checked first. The report lists the normalised weights $\tilde{w}_k$ and the effective panel size $K_{\text{eff}} = 1/\sum_k \tilde{w}_k^2$, which falls from the number of members towards 1 as one member takes all the weight. Uncertainty comes from repeated, independently seeded panel draws and from resampling candidates, stored apart.

Coverage. Every panel risk is reported with the number of members per tier, the reference mass covered (defined only for a finite, normalised member frame) and the fraction of sequence clusters represented. It adds the fraction of members with compatible structural inputs, the fraction with a pair-specific assay or calibrated oracle, the reference mass whose exposure class is unresolved, and K_{eff}.

Oracles and delivered sets. Pose, pair-binding and generic polyspecificity oracles stay in separate fields. Two oracles are combined only when they measure the same endpoint under compatible conditions and a declared measurement model defines the combination. Agreement of two oracles is a diagnostic, since both can agree while ranking the measured outcomes in reverse. A combined risk summary counts as panel risk only after its frozen form has been checked against held-out, construct-matched measured outcomes.

Candidates chosen for a portfolio form a delivered set, which differs from the off-target panel. Epitope competition between binders is stored as a directed graph, and each edge records whether one binder blocks, displaces or sandwiches another under the protocol of the assay. Competition is not transitive. Antibodies at adjacent or slightly overlapping epitopes can displace one another [14], so A may compete with B and B with C while A and C do not compete. Epitope bins are therefore a derived view under a named rule, either identical blocking fingerprints or complete-linkage clustering at a fixed cut height. Section 3 describes the selection.

1.2. Thresholds and conformal risk control

Populations. The three populations P_{hist}, P_{cal} and P_{gen} are those of Technical Report II (Section "Interface-JEPA"). Oracles train on P_{hist}. The calibration population P_{cal} is divided into a map-fitting subset and a risk-calibration subset. The score-calibration map is fitted on the map-fitting subset and then frozen, and thresholds are chosen on the risk-calibration subset, which is exchangeable with the candidates it serves. A map fitted on the same candidates would make the losses depend on one another and void the exchangeability that the guarantee below needs. The cohort P_{gen} comes from one frozen generator, guide, selector, panel profile and assay protocol, and it measures risk after the shift to generated designs. It keeps every selected candidate, including failures of synthesis, expression or assay, because removal by outcome would change the quantity measured. The deployment gap is the empirical risk on P_{gen} minus that on the risk-calibration subset at the chosen threshold, reported with a cluster-bootstrap interval.

Two-threshold grid. Thresholds are first chosen on the risk-calibration subset from finite grids. The grids, the ceilings α_f and α_r, a cluster bootstrap and a tie rule are fixed before calibration. A pair of thresholds is acceptable when the bootstrap upper endpoints of both mean losses lie at or below their ceilings. If none is, panel export is blocked. This rule carries no coverage guarantee, because the two losses move in opposite directions. The reference configuration therefore uses one alarm threshold on the whole panel, since a false-alarm rate of 10 percent per pair would raise an alarm on 88 percent of designs against 20 independent negatives if the alarms are independent.

Conformal risk control. A guarantee needs a different construction. Conformal risk control picks a threshold $\widehat{\lambda}$ from n risk-calibration examples so that the expected loss on a new example is at most α [15]. The loss is $L_i(\lambda) = \mathbb{1}[X_i \in \widehat{\mathcal{D}}_\lambda]\, B_i$, with the failure indicator $B_i = 1 - Y_i$: candidate i is called designable at threshold λ and later found bad. The theorem needs four conditions. The loss is right-continuous in λ, which an admission rule $s(X) > \lambda$ with a strict inequality secures. It is non-increasing in λ. It is at most B_L, with some $\lambda_{\max}$ at which it is at most α. And the $n+1$ losses are exchangeable. The threshold is

$$\widehat{\lambda} = \inf\Big\{\lambda :\ \widehat{R}_n(\lambda) \le \alpha - \frac{B_L - \alpha}{n}\Big\}, \qquad \widehat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} L_i(\lambda),$$

If the feasible set is empty, set $\widehat{\lambda} = \lambda_{\max}$, where $\lambda_{\max}$ admits no candidate; otherwise use the displayed infimum. Then $\mathbb{E}[L_{n+1}(\widehat{\lambda})] \le \alpha$. For $B_L = 1$ and $\alpha = 0.10$ the empirical inequality can be feasible only when $n \ge 9$. Independently tuned thresholds do not define the required nested scalar family. A fixed scalar score, such as the minimum of two scores oriented so that larger values favour admission, defines nested sets through $s(X) > \lambda$. Supplementary Section S1 gives the proof and the fallback rule (Theorem 5.6).

Limits of the guarantee. Two limits remain for a valid path. First, the guarantee controls the marginal rate of candidates that are called and bad. The share of bad candidates among those selected can differ from it by an arbitrary factor. A guarantee on that share needs Learn then Test [16] or conformal selection [17].

Second, a prospective campaign is a shifted population. Let Q be the campaign distribution and W a bound on the density ratio $\mathrm{d}Q/\mathrm{d}P$. The threshold chosen on the risk-calibration subset then has risk at most $\min(W\alpha, B_L)$ under Q. The bound is uninformative once $W \ge B_L/\alpha$, which is 10 at $B_L = 1$ and $\alpha = 0.10$. Weights for the design procedure can restore control in a weighted version of the rule, if they can be computed and the rule that labels a candidate bad is unchanged [15, 18]. The second condition can fail when the antigen changes, because a candidate is bad by its binding to the deployment antigen.

2. Liability, immunogenicity and glycan channels

The full system evaluates three typed channels on the final assembled construct: chemical-liability flags, a predicted presentation load on major histocompatibility complex class II (MHC-II), and glycan-occlusion scenarios. A sequence screen nominates sites, and a structural or measured endpoint decides what a campaign does with them. A channel output is a computed score. A claim about stability, immunogenicity or binding rests on a matched assay. Guidance (Technical Re-

port II (Section "Guidance and sequential Monte Carlo")) and maturation (Technical Report II (Section "Maturation and editing")) read the scores.

Scanned sequence. The scanned sequence is the product as it would be made: VH and VL domains or the VHH, linkers, tags, junctions, cleavage remnants, any Fc sequence, and every arm of a multispecific assembly. An isolated variable domain would miss peptides and motifs that form across a junction, so every window that crosses a component boundary is labelled.

Chemical liabilities. A sequence motif only nominates a site. Let $s_1 \dots s_L$ be the final sequence and r a liability family. A deterministic scanner returns $m_r(i; s) \in \{0, 1\}$ for the window anchored at position i, with value 1 when a susceptible context is present. A structural hotspot model returns $h_{\psi,r}(i; X, \kappa) \in [0, 1]$, where X is the hard-decoded structure and κ holds the conditions: pH, temperature, buffer, stress duration, redox state and construct format. The structural liability score sums the product over positions:

$$L_r(X; \kappa) = \sum_{i=1}^{L} m_r(i; s)\, h_{\psi,r}(i; X, \kappa).$$

The hotspot term estimates whether the structure around a motif resembles that of known hotspots. A score is interpretable only under the conditions and calibration cohort of $h_{\psi,r}$. An exposed Asn in an NG window with $h_{\psi,r} = 0.8$ is therefore ranked for confirmation ahead of a buried one with $h_{\psi,r} = 0.1$ (illustrative values). Structure-aware evidence exists for Asn deamidation and Asp isomerisation. In a large uniform set of antibody variable-domain residues, degradation hotspots measured by mass spectrometry were characterised by conformational flexibility, the size of the following residue and secondary-structure parameters. A CDR predictor was derived from them [19]. Other families (Met and Trp oxidation, N-terminal cyclisation, fragmentation) give sequence flags only. A site keeps the label of a sequence motif or a structural estimate until a site-resolved forced-degradation assay on the product, with binding retention after stress, replaces it.

The scanner also reports every N-X-S/T sequon (X is any residue except Pro) in the binder. Its occupancy is unknown until a glycopeptide mass-spectrometry measurement on the same construct and host exists, and no occupancy probability is computed. Surface charge patches at a declared pH are developability descriptors, and a gate may use one only after calibration against a named assay. Steering uses a smooth expectation of each motif indicator, whereas admission uses the hard-decoded sequence.

Predicted MHC-II presentation load. The load summarises predicted peptide presentation by a declared panel of human leukocyte antigen (HLA) class-II molecules. It serves as a sequence-level design penalty and as an input to immunogenicity assessment. It does not measure an anti-drug antibody response, and an industry commentary names the missing correlation between predicted T-cell epitope content and clinical anti-drug antibody outcome as a limit of such tools [20]. Humanness is a separate field.

The final sequence s is cut into overlapping windows of $L_p = 15$ residues, so the window set is $\mathcal{W}_{15}(s) = \{s_{i:i+14} : i = 1, \dots, |s| - 14\}$. NetMHCIIpan-4.0 supports variable peptide lengths and defaults to 15-mers for FASTA input, and it reports the binding core it identifies [21]. The design pins one release (4.3k), and every threshold names that release. The 4.3 update adds inverted peptide binders and covers HLA-DR, DP and DQ [22]. Strong-binder and weak-binder labels are release defaults with no clinical meaning: version 4.0 uses ranks below 2 and 10 percent [21], and the 4.3k service lists 1 and 5 percent.

The panel H lists complete molecule identifiers, with alpha and beta chains for HLA-DP and HLA-DQ. The weight π_h of molecule h comes with its source population, frequency data and normalisation. Let r_{wh} be the tool rank of window w for molecule h, φ a monotone transform of it oriented so that larger values mean more burden, and $\omega_{\text{region}}(w)$ the weight of the region that contains w. The load averages the burdens of each window with the panel weights, scales them by region weight and sums over windows:

$$L_{\text{MHCII}}(s; H) = \sum_{w \in \mathcal{W}_{L_p}(s)} \omega_{\text{region}}(w) \sum_{h \in H} \pi_h \, \varphi(r_{wh}).$$

A weight vector is a population weight when it has a population source and a design weight otherwise. Region weights separate CDR, framework, constant-region, linker, tag and junction windows, and a window that crosses two components takes the junction weight. Burdens 0.7 and 0.2 on two molecules weighted 0.6 and 0.4 give $0.6(0.7) + 0.4(0.2) = 0.50$ for one 15-mer before its region weight. The adapter also counts the deduplicated physical core and HLA events that the pinned release labels strong. A change in this load supports a lower predicted core load and no other conclusion.

Glycan occlusion. Antigen glycans are mobile and can shield large parts of a protein surface. GlycoSHIELD grafts sampled glycan conformers onto a static structure in minutes, and its authors apply it to N-cadherin, coronavirus spike proteins and gamma-aminobutyric acid receptors [23]. Let $\mathcal{H} = \{h_1, \ldots, h_M\}$ be a finite set of glycosylation scenarios for antigen A. Each fixes the protein conformer, occupied sites, site-specific glycoforms, conformer-library version and sampling seed. For epitope residue j, the accessibility in scenario h is the fraction of N_h sampled conformers that leave the residue reachable to a probe:

$$\sigma_j^{(h)} = \frac{1}{N_h} \sum_{n=1}^{N_h} \mathbb{1}[\text{probe reaches residue } j \text{ in conformer } (h, n)].$$

Because N_h is finite, every value carries Monte Carlo uncertainty. Scenarios stay unweighted until independent occupancy and glycoform evidence supplies weights. The envelope of residue j runs from the minimum $\underline{\sigma}_j$ to the maximum $\overline{\sigma}_j$ over scenarios. For an epitope E with residue weights a_j, the aggregate is $S_E^{(h)} = \sum_{j \in E} a_j \sigma_j^{(h)} / \sum_{j \in E} a_j$, with lower envelope $\underline{S}_E = \min_h S_E^{(h)}$. Required residues E_{req} also receive $\underline{\sigma}_{\text{req}} = \min_{h,\, j \in E_{\text{req}}} \sigma_j^{(h)}$, which is not applicable for an empty set. The channel sends central accessibility, envelope width, sampling uncertainty and a support mask to antigen conditioning (Technical Report II (Section "Antigen conditioning and Interface-JEPA")).

Floors for $\underline{S}_E$ and $\underline{\sigma}_{\text{req}}$ are set before evaluation. An epitope is accessible when its lower envelope clears both floors and occluded when its upper envelope lies below a floor. It is unresolved when the envelope or the Monte Carlo interval crosses a floor, and out of scope when the site, glycoform, chemistry or antigen class falls outside the library's support. These states describe accessibility in the registered ensemble. Measured binding is a separate readout.

The scenario set covers glycosylation of the antigen only. A candidate that carries a binder sequon is therefore unresolved for glycan effects, because its contact scenarios omit binder glycoforms. A variable-domain glycan can raise or lower affinity depending on its position [24] and adds chemical heterogeneity [25]. Admission needs a sequon-free sequence, a qualified aglycosylated format, or measured occupancy with matched binding data. The admission predicate of Technical Report I (Section "Immunoglobulin contract") already rejects sequons inside the CDRs.

Order of evaluation. The hard-decoded candidate passes six steps in a fixed order. The first is validity of sequence, atoms and disulfides (Technical Report I (Section "Immunoglobulin contract")). The next three are the chemical motif and structure-context label with binder sequons, the MHC-II presentation load of the final construct, and glycan occlusion of the antigen epitope. The fifth is the pose, pair-binding, affinity and panel-risk endpoints (Section 1), and the last is experimental confirmation where the campaign profile specifies one. A missing required input leaves the result unavailable, and an interval that crosses a decision floor leaves it unresolved. Each channel acts on steering or admission only after a pre-specified comparison of channel-active against channel-inactive designs under the same conditioning, seeds and step budget (Supplementary Section S3). The liability comparison tests the structural proxy, and a forced-degradation assay measures degradation.

Developability predictors. Developability is treated as a vector of axes, and a missing axis stays missing: it is filled with neither zero nor a mean. The axes come from published predictors: AbNatiV for VHH nativeness [26], NanoMelt [27] and TEMPRO [28] for the apparent melting temperature T_m, BioPhi with OASis for humanness [29], the sequence-liability scanner above, and the MHC-II load. Disagreement between the two T_m predictors is descriptive. Each predictor returns, for every registered metric, a value, units, an interval and a flag that states whether it is calibrated. A registry fixes the orientation of each metric (higher is better, lower is better or two-sided), its units and its valid range, and a value outside the range is an error. A predictor whose output is unavailable gives no admission decision.

Intervals follow an empirical rule on a named held-out set of natural antibodies. For a predictor $\hat{f}$ with residuals $\hat{f}(x_i) - y_i$ on n held-out items, the lower endpoint for a higher-is-better metric is the prediction minus the residual order statistic of rank $\lceil (n+1)(1-\alpha) \rceil$, with $\alpha = 0.10$. A lower-is-better metric uses the mirrored rule. The rank exists only if it does not exceed n, which at this α needs $n \geq 9$. Otherwise the predictor is flagged as uncalibrated, and a percentile filter serves as an engineering screen. These intervals apply to the stated reference population, and application to generated designs requires calibration on that population. A coverage statement needs a procedure whose assumptions hold for the population, such as conformal risk control [15].

Five metrics use clinical-stage reference bands. Raybould and colleagues modelled 242 clinical-stage therapeutics and derived guidelines for five metrics [30]. An amber flag marks the extremes of that distribution and a red flag a value not seen in it (Table 2). Percentiles from one assay, buffer and format are not universal limits, and a VHH has no light chain, so charge symmetry does not apply to it.

A framework residue rule becomes a hard constraint only inside a hashed scaffold profile backed by a primary experiment (Technical Report I (Section "Scaffold profiles, cysteines and pairing")). Without an executable checker no universal residue set is substituted.

3. Decisions under ambiguity and portfolio selection

The selection layer reads scored candidates and returns a set of exactly N designs, or no set when none is feasible or the search budget ends. It first combines the scores of several plausible models into one cautious utility, then chooses the set. Table 3 gives the inputs, the operations, the outputs and the status of the three modules that act on a finished candidate: maturation

Table 2: Reference bands of the Therapeutic Antibody Profiler for five metrics, derived from 242 clinical-stage therapeutics [30]. Total CDR length is in residues, and the patch metrics and charge symmetry are in the units of the profiler.

Metric (orientation)	**Amber**	**Red**
Total CDR length (two-sided)	54 to 60	above 60
Surface hydrophobicity patches, `psh` (two-sided)	83.84 to 100.71 and 156.20 to 173.85	below 83.84 or above 173.85
Positive-charge patches, `ppc` (lower is better)	1.25 to 3.16	above 3.16
Negative-charge patches, `pnc` (lower is better)	1.84 to 3.50	above 3.50
Fv charge symmetry, `sfvcsp` (higher is better)	-20.40 to -6.30	below -20.40

(Technical Report II (Section "Maturation and editing")), pose judging (Section 1) and this layer. Figure 2 draws the upstream steric glycan clash exclusion and zero-sum Plackett–Luce specificity calibration beside the totally unimodular min-cost flow portfolio programme that selects $K = 3$ diverse binders.

Table 3: Contract of the three modules that act on a finished candidate. Each row names what the module consumes, what it does and what it returns. The three modules are specified as designs.

Module	**Inputs**	**Operations**	**Outputs**	**Status**
Maturation	One parent antibody or nanobody, the three nested masks, a mutation budget, and the packed parent state from the frozen encoder	Partial corruption to a re-entry point on the master clock, guided solve to the data end, hard decoding, restoration of the fixed context, acceptance or rejection	A variant of the parent with its realised mutation count, or a rejected draw	Proposed
Pose judging	A candidate, a protein antigen record, a declared epitope and an antigen-state profile	A co-folder run over a fixed roster of registered seeds, pose clustering, reranking inside the selected cluster, registration against the declared epitope	An ensemble of proposed complexes, one selected pose, the pose score, the epitope metrics P_E, R_E and EpiF1, and a state margin where conformers are declared	Proposed
Portfolio selection	A finite candidate pool with its gate results, one utility per plausible world with reference weights, a target specification with weights, the committed size N and the campaign constraints	The cautious utility over the worlds, the coverage objective, and an integer programme under the size, cluster and gate constraints	A set of exactly N designs with its realised coverage, every constraint value and the removal contribution of each member; otherwise a proved absence of a feasible set, or an exhausted budget with no proof	Proposed

Model ambiguity. Calibrated predictors can disagree, and a design that ranks first under one checkpoint can rank tenth under another. The layer therefore scores each design in every one of a declared set of plausible worlds and aggregates the scores with caution. Let $\Omega = \{1, \ldots, M\}$ index the worlds. A world is one surrogate checkpoint, assay mapping, panel perturbation or antigen-family scenario. World j has a reference weight $\pi_j > 0$, with $\sum_j \pi_j = 1$. For candidate X and antigen A, a registered transform puts the raw score of world j on one dimensionless scale, the

utility $U_j(X, A)$, where larger is better. The entropic pessimistic utility lets poor worlds count for more. It multiplies each reference weight by $e^{-\eta U_j}$, which grows as the utility falls, sums, and maps the result back to the utility scale:

$$U_{\text{ent}}(X, A) = -\frac{1}{\eta} \log \Big(\sum_{j=1}^{M} \pi_j \, e^{-\eta U_j(X,A)} \Big), \qquad \eta > 0.$$

The risk aversion η is dimensionless. As $\eta \to 0$ the value tends to the reference-weighted mean $\sum_j \pi_j U_j$, and as η grows it tends to $\min_j U_j$. An identity gives the meaning. Let ν be any weights over the worlds, and let $\mathrm{KL}(\nu \mid \pi) = \sum_j \nu_j \log(\nu_j/\pi_j)$ be the Kullback-Leibler divergence (KL) from the reference weights. Then

$$U_{\text{ent}} = \min_{\nu} \Big\{ \sum_j \nu_j U_j + \tfrac{1}{\eta} \, \mathrm{KL}(\nu \mid \pi) \Big\}, \qquad \nu_j^\star = \frac{\pi_j e^{-\eta U_j}}{\sum_k \pi_k e^{-\eta U_k}}.$$

An adversary shifts weight towards the poor worlds and pays a KL price, scaled by $1/\eta$, for leaving the reference. The minimiser is $\nu^\star$, and the identity follows from $\mathrm{KL}(\nu \mid \nu^\star) \geq 0$ (Supplementary Section S1). To check the identity by hand, take $\pi = (0.5, 0.5)$, $U = (0.9, 0.3)$ and $\eta = 2$. Then $U_{\text{ent}} \approx 0.515$ and $\nu^\star \approx (0.231, 0.769)$, against a plain mean of 0.6.

A hard KL ball caps how far the adversary may move: $U_{\text{KL-ball}}(\delta) = \min\{\sum_j \nu_j U_j : \mathrm{KL}(\nu \mid \pi) \leq \delta\}$. For $\delta > 0$ the point $\nu = \pi$ is strictly feasible, and convex duality gives

$$U_{\text{KL-ball}}(\delta) = \sup_{\lambda_{\text{DRO}} \geq 0} \Big[-\lambda_{\text{DRO}} \log \sum_j \pi_j e^{-U_j/\lambda_{\text{DRO}}} - \lambda_{\text{DRO}} \, \delta \Big].$$

The entropic form fixes $\lambda_{\text{DRO}} = 1/\eta$. The ball chooses λ_{DRO} and adds the radius term, so η and δ describe different families. Both forms reweight only the listed worlds, so a plausible world missing from the list stays out. The Wasserstein set moves probability mass between nearby world descriptions. For a nominal law $\widehat{P}$ on a declared world space, a ground cost c_W and a radius ρ_W, $U_{\text{W}} = \inf\{\mathbb{E}_Q[U] : W_{c_W}(Q, \widehat{P}) \leq \rho_W\}$. The ground cost decides which shifts count as near. The duality that makes the set tractable needs U to be lower semicontinuous, with integrability and growth conditions, and a finite reformulation needs more structure [31]. Guidance (Technical Report II (Section "Guidance and sequential Monte Carlo")) uses the entropic utility, and the Wasserstein value is evaluated offline. Every campaign stores the world membership, the utility transforms, the reference weights and a leave-one-world-out sensitivity.

3.1. Fixed-size portfolio

Coverage objective. A campaign carries a fixed number of designs forward. The layer fixes an integer N before any candidate score is read, then chooses the best set of that size. The N highest scores can be near copies that share one epitope, and one poor epitope would then lose the whole set. Let $\mathcal{V}$ be a finite candidate pool. A target specification $\mathcal{T} = \{(A_j, E_j, w_j)\}_{j=1}^{n_T}$ lists antigens A_j, epitopes E_j and weights $w_j \geq 0$. Let $p_j(X) \in [0, 1]$ be the binding-utility input of candidate X for target j, obtained from the utilities above by a registered map. The coverage of a set $\mathcal{B} \subseteq \mathcal{V}$ is

$$F_{\text{cov}}(\mathcal{B}) = \sum_{j=1}^{n_T} w_j \log \Big(1 + \sum_{X \in \mathcal{B}} p_j(X) \Big).$$

For each target, the inputs of all chosen designs are added, one is added to the sum, and the logarithm is weighted. The logarithm flattens, so a design aimed at a target that is already covered adds little. Write $S_j(\mathcal{B}) = \sum_{X \in \mathcal{B}} p_j(X)$. Adding X raises the coverage by $\sum_j w_j \log\big(1+p_j(X)/(1+S_j(\mathcal{B}))\big)$, which cannot grow when $\mathcal{B}$ grows. Hence F_{cov} is monotone and submodular by direct calculation. As a check of diminishing returns, take two targets of weight 1 and $p(A) = (0.8, 0.1)$, $p(B) = (0.2, 0.7)$. Candidate B alone adds 0.713, and added to $\{A\}$ it adds $1.281 - 0.683 = 0.598$.

Selection problem. The selection solves

$$\max_{\mathcal{B} \subseteq \mathcal{V}} F_{\mathrm{cov}}(\mathcal{B}) \quad \text{subject to} \quad |\mathcal{B}| = N, \quad R_{\mathrm{self}}(\mathcal{B}) \le r_{\max}, \quad \sum_{i<k} \mathrm{collision}(X_i, X_k) \le c_{\max}, \quad \mathcal{B} \in \mathcal{F}_{\mathrm{format}}.$$

R_{self} is a set-level self-risk. The collision sum counts each unordered pair once. The family $\mathcal{F}_{\mathrm{format}}$ holds the chain-pairing, valency, linker and assembly rules of the chosen format (Technical Report I (Section "Immunoglobulin contract")). Every member also passes the per-candidate gates $\ell_f(X) \le \alpha_f$ and $\ell_r(X) \le \alpha_r$ on the forbidden-hit and required-miss masses (Section 1). The reference configuration uses one alarm threshold, because the coverage guarantee of conformal risk control holds for a scalar nested family of thresholds and a grid over two thresholds carries none [15]. A candidate with a missing evaluation is never counted as cleared. The reference configuration sets $N = 5$, $c_{\max} = 2$ and $R_{\mathrm{self}} \equiv 0$, and counts a collision when two designs lie in one CDR-H3 cluster at 80 percent identity. For $N = 5$ this allows at most two designs per cluster.

Integer programme. Each candidate i gets a binary variable z_i, with $z_i = 1$ when it is chosen. The reference configuration has one target, so F_{cov} is an increasing function of $\sum_i z_i p(X_i)$. The programme maximises that sum subject to $\sum_i z_i = N$, at most two chosen designs per CDR-H3 cluster, and $z_i = 0$ for any candidate that fails a gate. HiGHS [32] solves it through the `milp` routine of SciPy [33], with a time limit of 60 s, a relative gap of 10^{-4} and a pool limit of 1,000 candidates. With several targets, the concave logarithm enters through tangent lines: each target gets an auxiliary variable bounded above by tangents of $\log(1 + s)$ at fixed points. Tangents lie above the curve, so the solver bound stays an upper bound on the exact objective, and the incumbent is re-scored exactly. Greedy forward selection reaches at least $1 - (1 - 1/N)^N$ of the optimum under a size constraint alone, which is 0.672 at $N = 5$ and tends to 0.632 as N grows [34]. Once self-risk, collision and format constraints apply, the solver supplies a proved bound in place of this factor.

Outcomes. The solver returns one of four outcomes, and each is a statement about proof. It returns an incumbent set whose upper bound is proved equal to it, or an incumbent set with a proved upper bound and the relative gap. It returns no set when a solver certificate or presolve proves that no feasible set of size N exists, or when the budget ends with neither an incumbent nor a proof. A heuristic that finds no feasible set therefore cannot report infeasibility. The report of a returned set lists the committed N, the realised coverage, every constraint value and, for each member, the removal contribution $\Delta_X^{\mathrm{remove}} = F_{\mathrm{cov}}(\mathcal{B}) - F_{\mathrm{cov}}(\mathcal{B} \setminus \{X\})$, with members in canonical identifier order. An empty acceptable threshold grid upstream blocks panel export before selection (Section 1).

Scope of the guarantees. The entropic identity, the dual of the KL ball and the submodularity of F_{cov} are proved (Supplementary Section S1). A greedy rule has a guarantee under the size constraint alone, and the solver bound covers the added constraints. The Wasserstein duality holds

for a given utility only after its conditions are checked. The marginal control of false-designable calls is a rate over candidates, and the share of failing designs among the N chosen ones can differ from it by an arbitrary factor. These guarantees do not compose into an end-to-end rate.

4. Evaluation design

The evaluation design fixes, before any outcome is read, how the data are split, how the system is compared with published work, how evidence is counted and which claims the evidence can carry. Figure 1 shows three components of the validation design: connected-component leakage isolation across T^* anchored to PDB 4KRL (panel a), conformal risk control under covariate shift (panel b), and a simulated 53-gate Ville e-process monitor (panel c). The base model generates no structure, so its logged quantities describe the course of training and do not measure the quality of a design. Model quality is judged by the evaluation protocol of this volume (Section "Evaluation protocol").

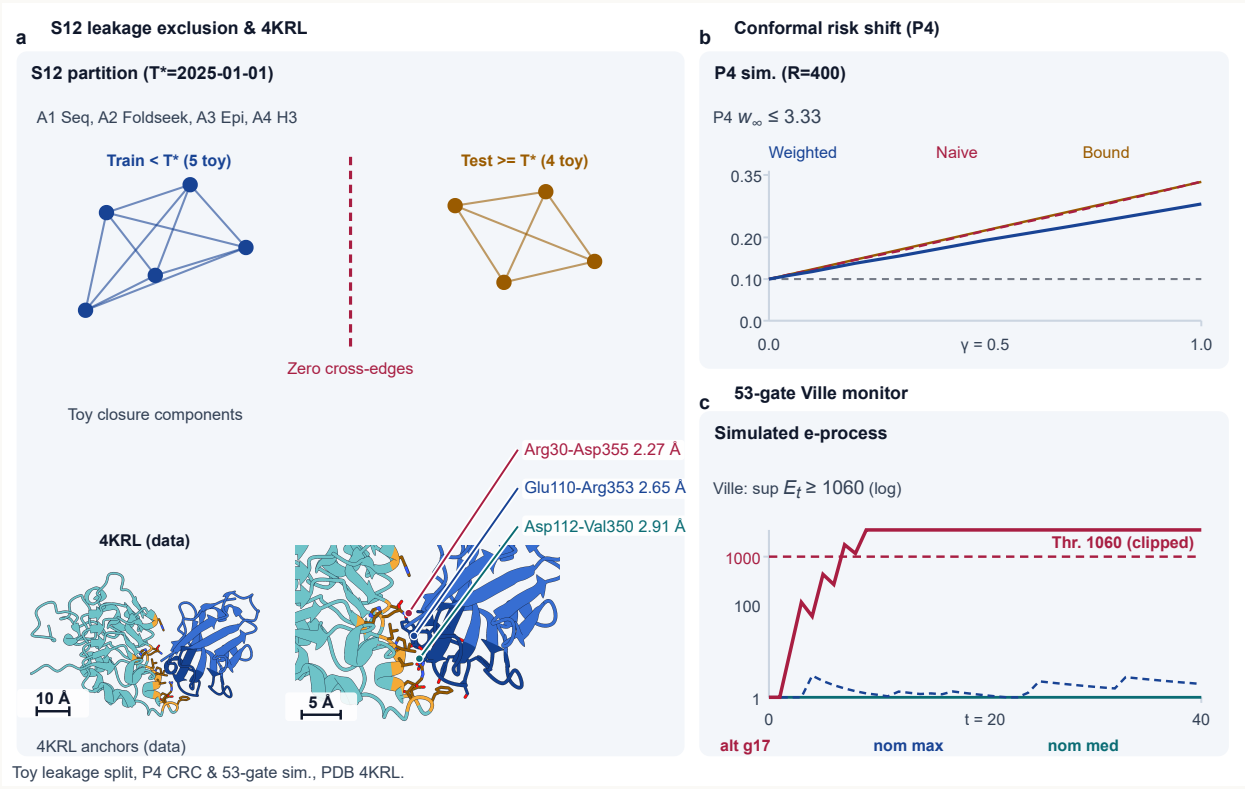

Figure 1: Connected-component closure across T^*, conformal risk shift bounds and a 53-gate Ville monitor structure the evaluation evidence. a, Toy transitive closure across axes A1–A4 separating pre-cutoff training clusters ($< T^*$) from held-out records ($\geq T^*$, $T^* = 2025\text{-}01\text{-}01$) with zero cross-edges, anchored structurally to PDB 4KRL (Arg30–Asp355 at 2.27 Å, Glu110–Arg353 at 2.65 Å, and Asp112–Val350 at 2.91 Å). **b,** Conformal risk shift simulation ($R = 400$, weight bound $w_\infty \leq 3.33$) comparing weighted, naive and upper-bound curves across mixture level $\gamma \in [0, 1]$. **c,** Simulated 53-gate Ville e-process monitor on a logarithmic axis crossing the 1060 exclusion threshold under alternative arm g17 while nominal arms stay near 1. PDB 4KRL distances reflect deposited coordinates; graph components and curves in b and c represent deterministic numerical stress simulations. CRC, conformal risk control; PDB, Protein Data Bank.

4.1. Reporting for the base model

The training script prints one line on rank 0 after every tenth successful update. The line holds the loss summed over the updates since the previous line, the terms of that loss, the global gradient

norm and the mean wall time per update. In the autoencoder stage the terms are the masked coordinate error over present atoms, in nm^2, the masked residue cross-entropy and the latent penalty, which is 0.01 times the masked mean square of the latent. In the denoiser stage the terms are the coordinate flow loss, in nm^2, and the latent flow loss. Each is the masked sum of squared residuals divided by the number of present residues and the block width, 3 for coordinates and 8 for the latent. Component means are taken over the successful updates of each logging window.

Shards store coordinates in ångströms. After centring, the code multiplies them by 0.1, so that the networks, the noise and the coordinate losses work in nanometres. The autoencoder reconstruction is reported as a root-mean-square error per coordinate in ångströms. The script averages the coordinate term over the successful updates of each window and converts the mean to a length:

$$\mathrm{RMSE}_{\text{Å}} = 10\sqrt{\mathrm{MSE}_{\mathrm{nm}^2}}.$$

The square root turns a squared error into a length, and the factor 10 turns nanometres into ångströms. A mean squared error of 0.01 nm^2 gives 1 Å. The result is an error per coordinate, and the root-mean-square distance over atoms is $\sqrt{3}$ times larger. The mean squared error is averaged over the window before the root is taken, and roots of single complexes are not averaged. The mean runs over all present atoms, including the C-alpha atoms, which the decoder receives as input. Because 1 nm^2 equals 100 Å^2, the coordinate term is 100 times smaller than in ångström units for the same structure, and its weight against the cross-entropy depends on the unit.

The minimum population squared-error loss is the expected conditional variance of the target velocity given the noisy state and times [35]; this term can be zero. Logged losses fluctuate because noise and times are sampled, so each value is a Monte Carlo estimate and the curve is read over windows.

4.2. Comparison with published numbers

Other systems enter only as published numbers, cited as reported. The design runs no competitor and loads no competitor weight. Published numbers of BoltzGen, for example, can be cited as reported [36].

A published number belongs to one panel, one cut-off, one judge and one success rule, and these differ between papers. Table 4 lists the panels and cut-off dates of four published works relative to $T^* = 2025\text{-}01\text{-}01$, the temporal split boundary illustrated on toy connected components in Figure 1a (Technical Report I (Section “Leakage axes, temporal holdout and split freezing”)). A number measured on one of these panels cannot be moved to another. A published panel older than T^* is held out of Omniverso training only through the benchmark rule of Technical Report I (Section “Curation, quality control and deduplication”).

Every comparison reports five items beside the published number, once for the source and once for Omniverso. The target set gives the targets, their number and how they were chosen. The cut-off date gives the deposit date after which a test complex counts as unseen, and the training cut-off of each model. The designs per target give the number generated and the number tested or scored. The judge model gives the model that scored the designs, its version and its training cut-off. The success definition gives the metric, its threshold and its unit, and whether it is a computational score or an assay result. A difference in any item is stated next to the comparison. A published number whose protocol cannot be read from its source is not compared.

Table 4: Panels of published works that the design cites: panel size (complexes, pairs or targets) and calendar date rule, as reported by the sources.

Source	Panel	Date rule
Fromm et al. [3]	110 antibody-antigen complexes	Training cut-off of 30 September 2021, shared by all tested methods
FoldBench [37]	172 antibody-antigen pairs	Entries deposited after 13 January 2023 and before 1 November 2024
SNAC-DB benchmark [4]	Public entries	Entries deposited up to 31 March 2025
Origin-1 [38]	Ten targets, fewer than 100 designs per target	Training on entries released before 1 January 2024

The base model generates no design, so only size and budget can sit side by side. The denoiser has 153,102,091 parameters, against about 159 M for the published reference denoiser [39]. That model saw about 374 M training samples, computed from the step counts and batch sizes in its tables (Technical Report I (Table: *Training scale stated by the papers behind the generative models in this section*)). The training budget of the base model follows the rule of Technical Report I (Section "Training of the base model"). Size and budget compare the cost of the models, and accuracy is a separate comparison.

The panels are not matched, so the comparison sets published rates side by side. A paired statement of equal accuracy, superiority or non-inferiority needs a matched panel. Sampling noise adds a second limit. With 30 targets, the normal-approximation 95% margin near a 50% hit rate is about 18 percentage points, and 200 independent units give about 7 points. A gap on targets that are not shared does not identify a cause. A comparison sentence takes one form: "System S reports a rate of r on its own panel of m targets. The panels are not matched to those of this study."

4.3. Splits and statistical rules for the full system

Splits. Leakage control uses the four axes of Technical Report I (Section "Leakage axes, temporal holdout and split freezing") (A1 to A4, thresholds set by the design). Records linked on any axis form one component, and a component goes wholly to training or to calibration, except that a component holding a record deposited on or after T^* leaves training as a whole (Figure 1a, Technical Report I (Section "Leakage axes, temporal holdout and split freezing")). Only records deposited on or after T^* are test records, and their older partners stay out of the test set. Three populations keep roles apart. Historical data fit models. Held-out calibration objects set thresholds. Outputs created after the generator, the selector and the compute policy are frozen are the only population on which a claim about the generator is read. Further routes of leakage are shared interface geometry [40], overlap with the pretraining corpus of a judge [41] and group-level links such as lineages and assay plates [42]. A claim that a model generalises beyond its training records therefore needs a prospective cohort collected after the model is frozen, as blinded challenges with experimental readouts provide for antibodies [43].

Fixed-horizon record. Every gate fixes its evidence record before an outcome is read. The record names the target population and sampling rule, the independent unit, the committed number of units, the statistic, the interval procedure, the error level, the threshold with its failure direction and the multiplicity rule (Supplementary Section S3):

$$\mathcal{E}_k = (\mathcal{P}_k, \mathcal{U}_k, n_k, T_k, I_k, \alpha_k, \tau_k, \mathcal{M}_{\mathrm{mult},k}).$$

The statistic is computed once, at the committed n_k. Early looks and stopping are not allowed, and sequential methods form a separate research interface with other error rules [44]. A floor reads a lower bound and a ceiling reads an upper bound.

A worked example shows why the bound decides and the point estimate does not. A gate commits 200 independent domains and a ceiling of 0.10 on a failure rate. If 18 domains fail, the point estimate is 0.09. The one-sided 95% exact upper bound is the 0.95 quantile of a beta distribution with parameters 19 and 182:

$$\widehat{R}_{200} = \frac{18}{200} = 0.09, \qquad \mathrm{Beta}^{-1}(0.95;\, 19,\, 182) \approx 0.131.$$

The bound lies above the ceiling, so the gate fails. The registry of the full system lists 53 gates. Each carries one of four dispositions. A required gate withholds its capability when it fails. A feature-conditional gate removes its feature. A diagnostic gate is reported and cannot admit anything. A not-applicable gate is a checked absence. Supplementary Section S3 lists the gates.

Sample size. Take a rate with baseline 0.80 and a one-sided exact binomial test at level 0.05. With 200 independent Bernoulli units, the test has 80% power at a true success rate of approximately 0.726, a drop of 7.4 points. Smaller departures from 0.80 have lower power. Failure to reject does not establish equivalence or adequate performance. With 1,000 units, a decrease of approximately 3.3 percentage points gives 80% power, and with 2,400 units a decrease of approximately 2.1 points does. The record marks an arm whose power at the effect of interest is below 80% as underpowered.

Endpoint statistics. Two yields per antigen score the top-ranked designs of an arm. For antigen i, arm m and design rank $k \le 5$, let $a_{ikm} = 1$ if the design passes every frozen on-target criterion and 0 otherwise, and let a missing design score 0. Let $h_{ikm} = 1$ if any member of the off-target panel exceeds the frozen alarm threshold or its evaluation is incomplete. The on-target yield of the top five and the yield that also clears the panel are:

$$Y_{im}^{\mathrm{on}} = \frac{1}{5}\sum_{k=1}^{5} a_{ikm}, \qquad Y_{im}^{\mathrm{clean}} = \frac{1}{5}\sum_{k=1}^{5} a_{ikm}\,(1 - h_{ikm}).$$

The second yield is a computational surrogate and measures no specificity. For the full arm and a comparator, let Δ^{on} and Δ^{clean} be the means over antigens of the paired differences, and let $L(\cdot)$ be the lower one-sided bound. The acceptance bar is $L(\Delta^{\mathrm{on}}) > -0.10$ and $L(\Delta^{\mathrm{clean}}) > 0$, both required. The margin lets the full arm lose half a passing design per portfolio of five on average, in return for a panel-clean selection.

The bound has level 97.5% on a final panel and 90% on a go or no-go panel, with Holm adjustment across comparisons. A yield takes only six values, so the bound inverts a conservative test, with a bootstrap-t interval as a check. Coverage is simulated on at least 100,000 data sets per scenario before a protocol is frozen.

Panel size sets the power. With a standard deviation of 0.25 for the paired difference and 25 antigens, the standard error is 0.05. Under a normal approximation, a true gain of 0.10 in the clean yield passes its bound 52% of the time and a gain of 0.15 passes 85% of the time. Any new panel needs its own calculation. The comparators in these tests are the own arms of the ablations below, and other groups enter through published numbers only.

4.4. Hypotheses, ablations and validation levels

Hypotheses. Six hypotheses give the programme its questions (Table 5). Each names an estimand, a population and split, a comparator, metrics, a refuting condition, a compute tier and a promotion gate. H5 requires assay-family data.

Table 5: Hypotheses H1 to H6 with their comparators and refuting conditions.

Id	Question	Comparator	Refuting condition
H1	Does the candidate depend on the true antigen?	Framework-only baseline; ungated antigen embedding. Controls: antigen ablation, shuffle, same-antigen epitope swap	A control has a non-positive lower bound, or the framework-only baseline matches the full conditioner on contact quality
H2	Do predicted contact probabilities match observed contacts?	Direct antigen embedding without a contact plan	Calibration slope or error outside the registered limits on a primary split
H3	Does sequence diversity rise without loss of specificity?	Reference decoder, contact-first arm, discrete-sequence arm	Diversity rises while antigen dependence or pose margins fail
H4	Does frame transport add value over C-alpha plus latent transport?	Partially latent flow at matched candidate count and GPU-hours	No lower-bound gain on pose metrics, or a memory or runtime penalty beyond the hardware profile
H5	Do proxy gates track assay outcomes?	Proxy-only triage against an assay-confirmed endpoint	Proxy success with assay failure above the registered ceiling, or a denominator changed after outcomes
H6	Does quality transfer from VHH to paired Fv?	VHH profile, paired-Fv extension	Paired Fv gains on one proxy by breaking pairing, humanness or polyreactivity gates

Ablations. Six axes isolate one cause each, and a missing separation on an axis has a stated meaning, so that no arm is read as a general result.

- *Antigen channel:* true antigen, ablation, shuffle and same-antigen epitope swap. A missing gain means that target conditioning is unavailable.
- *Framework shortcut:* full framework, framework dropout and framework only. A missing gain means that antigen evidence is too weak for admission.
- *Contact object:* none, predicted plan, transport-refined plan and contact first. A missing gain means that a miscalibrated plan cannot drive chemistry.
- *Decoder:* continuous latent, discrete sequence and paired Fv. A decoder-specific gain does not transfer across formats.
- *Geometry:* partially latent flow and frame flow. A geometry gain must beat its cost and numerical risk.
- *Evidence mode:* fixed horizon and sequential preview. A preview cannot serve as release evidence.

A failed hypothesis is reported with the split on which it failed and is not averaged into a larger result.

Validation levels. A capability enters the manifest of the full system only through evidence at the

level that matches its claim (Supplementary Section S3). Six levels are used. At the algebra level, shapes, units, masks, signs and limiting cases are checked by derivations and numerical fixtures. At the component level, one module is checked against its declared contract on cluster-disjoint examples with named gates. At the interface level, cross-module fixtures check that producer and consumer agree on type, scale, mask and provenance. At the system level, a fixed campaign is replayed end to end under one configuration. At the external level, the frozen system is tested on registered benchmark populations of independent complexes, panels or assay units. At the experimental level, selected candidates are tested against measured biological endpoints under predeclared wet-lab protocols.

For the base-model code the specified checks lie at the algebra level (shapes, masks, units and parameter counts) and at parts of the interface and system levels (shard reader and resume). Two outcome classes are defined. A computational nomination has passed every active computational gate and carries no experimental statement. An experimental release needs a wet-lab evidence package.

Claim boundary. The boundary holds at every level. Results are computational. A pose score, contact calibration, humanness score, predicted MHC-II rank and generic polyreactivity prediction may rank or quarantine candidates inside their endpoint scope. Binding, affinity, specificity, immunogenicity, aggregation and developability are assay-family claims. A prediction ranks a design. Only a wet-lab test can say that it binds, and a claim of any of these kinds about the model rests on such tests.

5. Evaluation protocol

The evaluation protocol of the full system is a set of gates. A gate is a pass-or-fail test with a fixed population, statistic, threshold and sample size. The registry holds 53 gates. All limits below are design values that are fixed before any evaluation. A change made after an outcome is read counts as exploratory analysis and cannot clear a confirmatory gate. The base model is evaluated only through its training diagnostics (reconstruction error, residue cross-entropy and flow loss), which describe the course of training and do not measure model quality. Quality is judged by the evaluation protocol: the gates of this section and the evaluation design of this volume (Section "Evaluation design").

Figure 1b and c draw the conformal risk shift curve under covariate perturbation alongside the simulated 53-gate sequential Ville e-process monitor. The text follows the order of the protocol: the record, the registered gates, the sample-size rules and the validation levels.

The fixed-horizon record. A gate is a promise written before the data exist. It says what is counted, what is measured, how uncertainty is handled and which result counts as failure. Writing it first stops a favourable result from choosing its own test. A consumer is one use of evidence. A chirality rate, an antigen-shuffle contrast, a panel-level false-discovery proportion and a co-folding score are four consumers with four records. The record of consumer k has eight fields:

$$\mathcal{E}_k = (\mathcal{P}_k, \mathcal{U}_k, n_k, T_k, I_k, \alpha_k, \tau_k, \mathcal{M}_{\mathrm{mult},k}).$$

- $\mathcal{P}_k$, the population: the objects tested and the rule that samples them.

- $\mathcal{U}_k$, the independent unit: what counts as one observation.
- n_k, the terminal sample size: the number of units, fixed in advance.
- T_k, the statistic: the number computed once, at n_k.
- I_k, the uncertainty procedure: the fixed-horizon interval or test.
- α_k, the error level: the declared error probability, between 0 and 1.
- τ_k, the threshold: the limit, and the direction in which failure lies.
- $\mathcal{M}_{\mathrm{mult},k}$, the multiplicity rule: the correction across the arms of one gate.

The index k matters because population, unit and loss differ between consumers. The independent unit is what is resampled. Residues of one domain, or designs against one antigen cluster, are not independent, so they enter through a domain-level or cluster-level unit. The statistic is computed once, at the committed n_k, with no early look and no optional stopping. The configuration accepts a positive integer for the evaluation horizon and the panel size and rejects the value `auto`.

A worked example shows why the bound decides and the point estimate does not. Suppose a gate commits 200 independent domains and a ceiling of 0.10 on the rate of domains with at least one registered chirality failure. Let K be the number of failing domains. If every domain fails independently with the same probability p, then K is binomial with $n = 200$. The one-sided upper limit at error level α is a quantile of a beta distribution:

$$I_\beta(k, n; \alpha) = \mathrm{Beta}^{-1}(1 - \alpha;\, k + 1,\, n - k) \;\; (k < n), \qquad I_\beta(n, n; \alpha) = 1.$$

This is the exact Clopper–Pearson limit [45]. For every p it lies above p with probability at least $1 - \alpha$. With 18 failures the point estimate is $18/200 = 0.09$, and the limit at $\alpha = 0.05$ is $\mathrm{Beta}^{-1}(0.95;\, 19,\, 182) \approx 0.131$. The limit is above the ceiling, so the gate fails although the point estimate is below it. Twelve failures give a limit of 0.0954 and thirteen give 0.1014. At most twelve failures in 200 therefore clear the ceiling.

Some gates weight their units. For predetermined weights $w_i \geq 0$ with $\sum_i w_i > 0$ and losses $\ell_i \in [0, 1]$, the weighted risk and the effective sample size are

$$\widehat{R}_w = \frac{\sum_i w_i \ell_i}{\sum_i w_i}, \qquad n_{\mathrm{eff}} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}.$$

Weights 2, 1 and 1 on losses 0, 1 and 1 give $\widehat{R}_w = 0.5$ and $n_{\mathrm{eff}} = 16/6 \approx 2.67$. The effective size measures how concentrated the weights are. An interval may use it in place of the sample size only when the theorem behind that interval allows the substitution.

Three populations keep their roles apart. Historical data fit the model and serve exploration. Held-out calibration objects set thresholds, calibration maps and interval procedures. Generated outputs, made after the generator, the guide, the selector and the compute policy are frozen, are the only population on which a claim about the generator is read.

Sequential evidence is a separate interface (Figure 1c). It needs a declared family of null distributions, a filtration $\mathcal{F}_t$ (everything observed up to step t) and a stake chosen from $\mathcal{F}_{t-1}$ before the next observation. Each factor $e_t \geq 0$ must satisfy $\mathbb{E}_P[e_t \mid \mathcal{F}_{t-1}] \leq 1$ for every P in the null family. The running product $E_t = \prod_{j \leq t} e_j$, with $E_0 = 1$, is then a nonnegative test supermartingale. Under the null, the chance that it ever reaches $1/\alpha$ is at most α [44].

Factors 1.2, 0.8 and 1.1 give the capital path 1.2, 0.96 and 1.056. The arithmetic is a trace, and the validity comes from the conditional-expectation condition. A family of valid terminal e-values may enter the e-BH procedure under that procedure's conditions [46]. The registered gates read fixed-horizon statistics only. F-AV and F-J7 test the sequential code on simulated streams and are diagnostics.

Every claim in a report also sits in an evidence row that binds five fields before any outcome is read. The hypothesis names the claim class and its population. The comparator is the unguided, null, shuffled or baseline arm. The metric is the terminal statistic with its unit. The uncertainty field holds the one-sided bound, the error level and the multiplicity rule. The refuting condition says what would contradict the row.

The registered gates. The 53 gates have identifiers that begin with F-. By family they are: foundations 3, substrate 2, conditioning 4, interface-JEPA 10, steering 6, maturation 6, judge 7, coverage 7, immunoglobulin 4, and liability and glycan 4. Table 6 lists them in registry order. Cohort sizes list the sampled objects; the independent resampling unit is specified separately for each gate. Where the margin of a limit depends on a comparator, it is fixed before evaluation and written as such.

Table 6: The 53 registered gates with their acceptance rules. Rates, correlations and gains are dimensionless; counts identify the stated designs, targets, panels or folds and are not uniformly independent units. Most bounds are one-sided at the 95% level; F-C1 and the random-order arm of F-JUDGE use higher levels. Abbreviations: AUC and AUROC, area under the receiver operating characteristic curve; KS, Kolmogorov–Smirnov; DockQ, docking model quality score; CDR, complementarity-determining region; MHC-II, major histocompatibility complex class II; PDB, Protein Data Bank.

Gate	What is tested	Acceptance rule
F-Gphi	Decoder-metric guidance against a Euclidean control: decoded clashes and bond-geometry violations at matched binding enrichment	Lower bound of the paired relative reduction ≥ 0.20; 500 designs
F-AV	Sequential-evidence code on simulated null and displaced streams (diagnostic)	Upper 95% limit on the exclusion rate ≤ 0.15; 200 streams per arm, horizon 2,000
F-CHI	Chirality and improper-dihedral statistics of decoded all-atom structures against a PDB reference	Rate limits $\leq 0.5\%$; KS distance ≤ 0.10; 1,000 plus 500 decodes, reference ≥ 200 structures
F-KEY	Partially latent interface against an explicit all-atom baseline: CDR-H3 recovery, decoded-versus-refolded RMSD and contacts	Non-inferiority margin, RMSD ceiling and contact floor fixed before evaluation; ≥ 200 complexes
F-DIV	Flow-density arithmetic (traces, step doubling) on synthetic vector fields	Property tests PT-29 to PT-31 pass at registered tolerances; no dataset
F-CB	Dependence on the true antigen: substitution, decoy, removal and renumbering controls	Margin for each control fixed before evaluation
F-MI	Condition-matched contrastive score per residue (a diagnostic, not a mutual-information bound)	Lower limit ≥ 0.20 nats per residue
F-EPI	Epitope precision, coverage and their harmonic mean on predicted complexes	Three floors fixed before evaluation; 1,000 designs
F-CYC	Cycle-consistency accuracy on same-antigen, different-epitope pairs, by antigen-fold stratum	Every stratum lower bound ≥ 0.70; 2,000 pairs, ≥ 100 per stratum
F-J1a	Size of the cognate-versus-decoy score gap, per head	Per-head margin fixed before evaluation; ≥ 200 designs, 10 decoys each
F-J1b	Cognate-versus-decoy AUC, per head	Lower limit ≥ 0.80; as F-J1a
F-J1c	Permutation test of the same gap, per head	$p \leq 0.001$ with 99,999 permutations; as F-J1a
F-J2	Triage recall of co-folding hits at a fixed judge budget	Recall ≥ 0.90 at triage ratio 0.15; gain $\geq +0.10$ over each comparator; 1,500 survivors

Continued on next page

Gate	What is tested	Acceptance rule
F-J3	Latent variance, covariance and task gain against a nominated comparator	Variance floor, covariance ceiling and task margin fixed before evaluation; $\geq$ 200 complexes
F-J4	Effect of adding a head's score on the external judge's target-versus-decoy margin	Lower limit of the margin change ≥ 0; 2,000 target-decoy pairs
F-J5	Violation rate of the head's lower interval on an exchangeable holdout	Upper limit $\leq \beta + \varepsilon$, both fixed before evaluation
F-J6	Residual dependence between head errors and judge errors	Residual correlation ≤ 0.30; score correlation ≤ 0.60; information ≤ 0.10 nats; recall leak ≤ 0.05; $\geq$ 200 complexes, $\geq$ 500 survivors, 5% of rejects
F-J7	Crossing frequency and delay of a diagnostic recurrence on simulated streams (diagnostic)	No threshold; $\geq$ 1,000 streams
F-J8	Polyreactivity transfer from single-domain to paired-Fv format	Paired-Fv AUROC lower limit ≥ 0.70; AUROC ratio lower limit ≥ 0.85; $\geq$ 300 constructs per format
F-SPEC	Negative-design steering: decoy-margin gain without loss on the on-target judge	Decoy-margin floor and on-target tolerance fixed before evaluation; 1,000 paired designs
F-UGAP	Agreement of 8-particle and 256-particle steering ensembles on bounded functionals	Upper endpoint ≤ 0.5 reference standard deviations; 100 contexts
F-TTS	Offline comparison of an evidence-allocated schedule with the fixed schedule (diagnostic)	Enrichment@50 gain ≥ 0.10; diversity ≥ 0.80 of the fixed schedule; 12 antigens $\times$ 200 designs per arm
F-ENT	Entropic reweighting ($\eta > 0$) against $\eta = 0$	Three margins fixed before evaluation
F-MN	Min-norm multi-objective correction against fixed weights	Hypervolume margin and per-objective tolerances fixed before evaluation
F-Dev	Fidelity and effect of the soft developability relaxation	Fidelity, error, disagreement and safeguard limits fixed before evaluation; 1,000 paired designs
F-MAT	Maturation yield, whole-complex rank correlation, humanness and developability minima	Yield margin and two score floors fixed before evaluation
F-M4	Edit-budget fidelity and monotonicity of the budget map	Deviation $\leq \max(1, 0.2k)$ for budget k; 300 parents per length decile
F-M5	Edit targeting by somatic-hypermutation motifs	Descriptive; no threshold
F-M6	Reversion towards germline: humanness gain and affinity retention	Humanness gain ≥ 0.05; affinity fraction ≥ 0.90; no new liability; $\geq$ 100 parents
F-M7	New CDR liabilities among accepted variants	Upper limit ≤ 0.02; 500 variants
F-M8	Offline comparison of an evidence-allocated maturation schedule (diagnostic)	Yield gain ≥ 0.08; edit-count deviation $\leq \max(1, 0.2k)$; $\geq$ 200 parents
F-JUDGE	Multi-seed judge with reranker against native co-folder confidence, random order and a physics ranker	Against native confidence: DockQ gain $\geq +0.05$ and enrichment@50 gain $\geq +0.10$; smaller gains against random order and a physics ranker; $\geq$ 200 targets
F-BAI	Bad-admission incidence of the pose-admission screen	Upper limit ≤ 0.10; 1,000 designs
F-IND	Rank association between judge residuals and steering-surrogate residuals	Absolute Spearman correlation ≤ 0.25, failure side 0.35; 500 designs
F-C1	Ranking-trained reranker against a pointwise reranker	DockQ gain $\geq +0.01$; enrichment gain $\geq +0.02$; 200 targets
F-C2	Epitope-registration test on seeded off-epitope and on-epitope designs	Off-epitope rejection ≥ 0.90; on-epitope rejection ≤ 0.10; $\geq$ 100 per arm
F-C3	State-discrimination margin for targets with two resolved states	AUC lower limit ≥ 0.70; $\geq$ 30 target-binder pairs
F-C4	Offline replay of adaptive pose-search scheduling	Admission incidence ≤ 0.10; cost cut $\geq$ 30%; DockQ within 0.02; enrichment within 0.05; 600 cases
F-COV	Leave-one-cluster-out coverage gap of the specificity panel	Gap ≤ 0.20 on both arms; 1,000 designs in $\geq$ 8 clusters
F-DIL	Sensitivity of the reported risk to a deliberately diluted panel	Upper limit ≤ 0.02 on the risk scale; 300 designs, $\geq$ 3 clusters
F-ORAC	Agreement between the assay-trained polyspecificity oracle and docking	Divergence upper limit ≤ 0.15; positive rank-correlation lower limit; 200 pairs, $\geq$ 60 assay-labelled
F-BIN	Displacement and direction reversal among competition-positive pairs over a concentration series	Displacement fraction ≤ 0.15; reversal fraction ≤ 0.10; $\geq$ 40 pairs

Continued on next page

Gate	What is tested	Acceptance rule
F-SET	Diversified panels against a coverage-only comparator	False-discovery proportion limit ≤ 0.30; diversity gain ≥ 0.05; 400 designs, ≥ 20 panels, ≥ 8 folds
F-SIGN	Whether any grid point meets both tier ceilings	At least one grid point meets both ceilings; 1,000 designs
F-DEP	Conditional dependence between steering and measured oracles	Bands at 0.05 and 0.15 nats; 400 labelled designs
F-IMGT	IMGT numbering round trip against independent curated labels	Agreement lower limit ≥ 0.995; zero undocumented disagreements; zero out-of-frame indices; ≥ 600 domains
F-VHH	Conformance of candidates to a registered scaffold profile	Conformance 1.000 over the whole panel
F-LIAB	Liability channel: adjudicated liability rate, affinity retention and charge shift	Margin and direction fixed before evaluation
F-CAN	Canonical-class assignment of CDR loops	Descriptive; no threshold
F-PAIR	Heavy-light pairing: cognate-versus-shuffled discrimination	Descriptive; no threshold
F-IMM	Predicted MHC-II core-load reduction with safeguards on germline identity, pose and affinity	Relative reduction lower bound $\geq 25\%$; 300 paired designs
F-GLY	Whether glycan accessibility predicts antibody contact (conditioning and epitope admission)	Odds-ratio lower bound ≥ 2.0; 120 complexes
F-GLY-REF	Glycan shielding as part of the reference measure	Gap-difference upper bound ≤ 0.02 and odds-ratio lower bound ≥ 2.0; 1,000 plus 400 designs

A gate may have several arms, and each arm has its own statistic, limit and bound. A gate passes only if every one of its arms passes. An arm below its required sample size fails the gate, and the gate is marked underpowered. A gate with no evaluated arm has neither a pass nor a fail. The registry is the only source of limits. Before evaluation, fix each arm's threshold, sample unit, terminal sample size, uncertainty procedure and multiplicity correction.

Each numerical limit is an engineering operating point or a refutation target. For F-IMM, F-GLY, F-GLY-REF and F-IMGT the cited sources fix the method and the inputs, and the registry fixes the acceptance value. A limit is re-examined on a fresh prospective set. F-J1a, F-J1b and F-J1c form one conjunction on one head: they share at least 200 designs with ten nested decoys each, and a head is admitted only if all three pass. F-GLY certifies the glycan shielding score for conditioning and epitope admission. F-GLY-REF certifies its use in the reference measure, and a record that cites F-GLY for that use is invalid.

Sample size and power. A pass says little if the test could not have failed. Table 7 shows, for the most common design in the registry, how small a fall each cohort size can detect. The design is a Bernoulli endpoint with a baseline success rate of 0.80, a one-sided exact binomial test at level 0.05 and 80% power.

With 200 independent Bernoulli units, the test has 80% power at a success rate of approximately 0.726. Smaller departures from 0.80 have lower power. Failure to reject does not establish equivalence or adequate performance. With 1,000 units, a decrease of approximately 3.3 percentage points gives 80% power. The smallest detectable effect is recorded beside each estimated gate. The table covers rates only, and a correlation, an AUC or a quantile needs its own calculation. The sizes count independent units, and designs for one antigen are not independent. F-TTS lists 2,400 designs, but its test resamples 12 antigens, so its independent units are the 12 antigens.

Validation levels. Six validation levels order the evidence from arithmetic to biology, and a claim needs evidence at the level that matches it. A derivation shows that units agree, and a binding

Table 7: Smallest fall in a success rate, from a baseline of 0.80, that a one-sided exact binomial test at level 0.05 detects with 80% power on independent Bernoulli units. Rates are proportions.

Independent units	Smallest detectable rate	Smallest detectable fall
50	0.632	0.168
100	0.686	0.114
200	0.726	0.074
300	0.737	0.063
500	0.753	0.047
1,000	0.767	0.033
2,400	0.779	0.021

measurement shows that a design binds.

1. **Algebra** asks whether shapes, units, masks, signs and limiting cases agree. The evidence is derivations and numerical fixtures.
2. **Component** asks whether one module meets its declared contract. The evidence is cluster-disjoint examples and named gates.
3. **Interface** asks whether producer and consumer agree on type, scale, mask and provenance. The evidence is cross-module fixtures.
4. **System** asks whether a fixed campaign replays from its run record and artefacts.
5. **External** asks whether the frozen system transfers to registered benchmark populations. The evidence is independent complexes, panels or assay units.
6. **Experimental** asks whether selected candidates meet measured biological endpoints. The evidence is predeclared wet-lab protocols.

An arm whose required model or data are missing is recorded as not evaluated; a missing value is never estimated.

6. Scope, Ethics, Licence, and Declarations

6.1. Scope, Dual Use, and Licensing

Scope. The work is computational and contains no wet-lab step. Predictions of the full system remain predictions until an experiment tests them. A comparison with another group uses its published numbers, names the differences in protocol and states that the panels are not matched.

Dual use and biosafety. A tool that designs binders can be misused, as can any generative model for biochemistry. A drug-discovery model proposed toxic molecules once its objective was inverted [47]. AI-redesigned variants of proteins of concern evaded the screens of nucleic acid synthesis providers until those screens were patched [48]. Synthesis is therefore the control point, and one proposal keeps synthesis records in repositories that are queried only in emergencies [49]. The study takes three positions. The base model contains no component that conditions on a target or emits an antibody sequence. The manuscript prints no designed sequence, and the code package releases none: its only data are synthetic complexes. Any later release of an antigen-conditioned

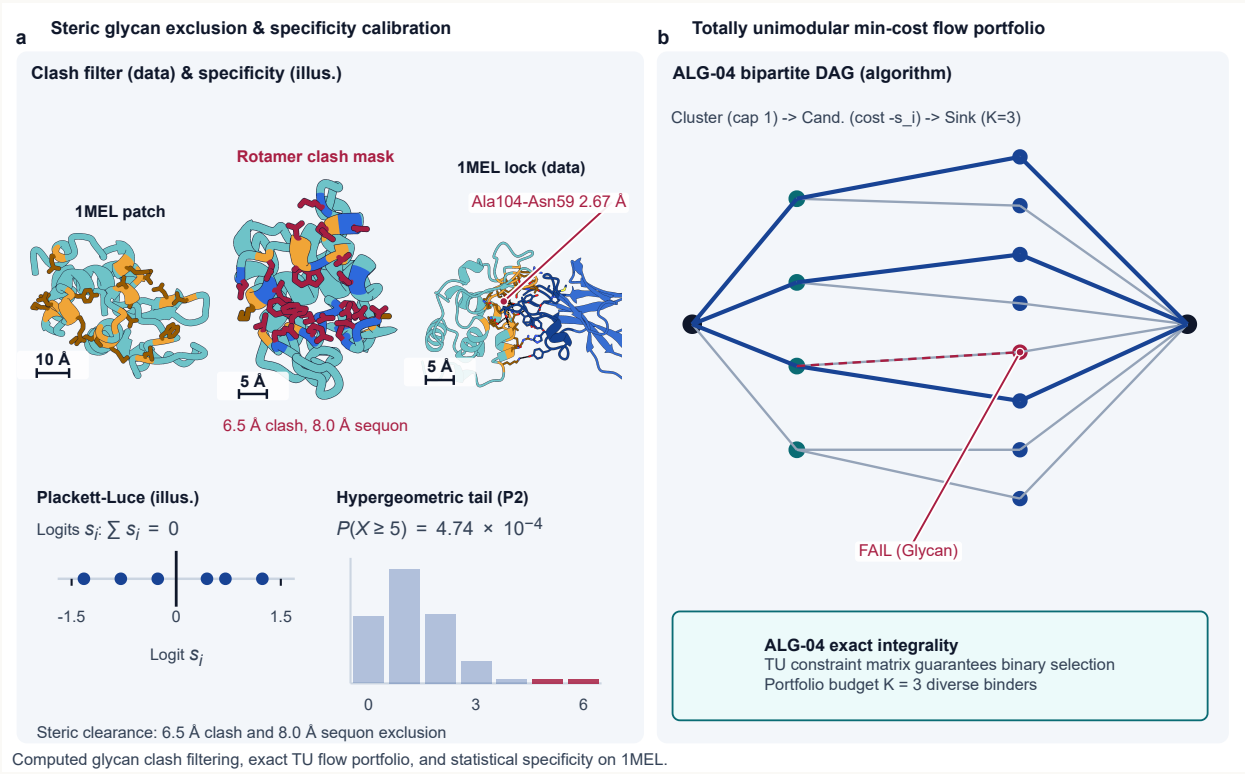

Figure 2: Steric glycan masking, specificity calibration and totally unimodular min-cost flow select a diverse $K = 3$ binder portfolio on PDB 1MEL. a, PDB 1MEL contact patch, rotamer glycan clash mask (6.5 Å clash and 8.0 Å sequon thresholds), and native interface lock (Ala104–Asn59 at 2.67 Å), paired with illustrative zero-sum Plackett–Luce logits ($\sum s_i = 0$) and the exact P2 hypergeometric tail distribution ($P(X \geq 5) = 4.74 \times 10^{-4}$). **b,** ALG-04 bipartite min-cost flow network routing cluster units to $K = 3$ selected candidates while eliminating candidate v_5 after steric glycan failure; total unimodularity guarantees exact binary integrality. Patch geometry derives from PDB 1MEL and the tail probability reflects discrete exact calculation; Plackett–Luce dots and network routing illustrate the algorithmic contract. DAG, directed acyclic graph; PDB, Protein Data Bank; TU, totally unimodular.

generator would first need a review of misuse and a screen of requested targets against curated lists of controlled pathogen accessions. These positions reduce risk and do not amount to a safety claim.

Licences and data. The code package is licensed under the GNU Affero General Public License, version 3 (AGPL-3.0). Its section 13 requires that a modified version offered to users over a network make its source available. The licence covers code and grants nothing on weights or data. Third-party libraries keep their own licences, and the main ones are named in Technical Report I (Section "Computing, software and availability"). Third-party papers appear as citations only. The licence of trained weights follows the licences of the training data (Technical Report I (Section "Sources and licences")). The base model reads public protein structures only and no personal, clinical or patient data, and each downloaded file is stored with its SHA-256 digest (Technical Report I (Section "Shard format and sampling")).

6.2. Declarations

Author contributions. D.K. is the sole author and carried out the conception, design, code and writing of the work.

Data and code availability. The training code, its tests, the configuration files and the container definition are supplied as Supplementary Information and described in Supplementary Section S4. The code is licensed under the GNU Affero General Public License, version 3. Each dataset stays with its provider, and this manuscript redistributes none. The sources and their licences are listed in Technical Report I (Section "Data") and Supplementary Section S2.

Artwork. The protein pictures in Figures 1 and 2 (and the accompanying plates of Technical Report I) were rendered with UCSF ChimeraX from deposited coordinates of the Protein Data Bank (PDB entries 1MEL and 4KRL).

References

[1] Eva Smorodina, Montader Ali, Klara Kropivšek Brumat, Leonardo Salicari, Samo Miklavc, Aibek Kappassov, Chengcheng Fu, Pietro Sormanni, Ario de Marco, and Victor Greiff. Structural Plausibility Without Binding Specificity: Limits of AI-Based Antibody-Antigen Structure Prediction Confidence Scores. *bioRxiv*, 2026. doi: 10.64898/2026.03.02.709004.

[2] Sankar Basu and Björn Wallner. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE*, 11(8):e0161879, 2016. doi: 10.1371/journal.pone.0161879. PMID:27560519.

[3] Samuel Fromm, Marko Ludaic, and Arne Elofsson. Evaluating deep learning based structure prediction methods on antibody–antigen complexes. *Bioinformatics*, 42(4):btag136, 2026. doi: 10.1093/bioinformatics/btag136. PMID:41863324.

[4] Abhinav Gupta, Bryan Munoz Rivero, Ruijiang Li, Jorge Roel-Touris, Yves Fomekong Nanfack, Maria Wendt, Yu Qiu, and Norbert Furtmann. SNAC - DB : An ML -ready database for antibody and NANOBODY ® VHH –antigen complexes with expanded structural diversity and real-world benchmarking. *Protein Science*, 35(7):e70655, 2026. doi: 10.1002/pro.70655. PMID:42252708.

[5] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey

Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w. PMID:38718835.

[6] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

[7] Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955.

[8] Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charlles R. A. Abreu, Stephen E. Farr, Emilio Gallicchio, Anton Gorenko, Michael M. Henry, Frank Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João PGLM Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Philip Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B*, 128:109–116, 2024. doi: 10.1021/acs.jpcb.3c06662.

[9] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, 8:3314–3321, 2012. doi: 10.1021/ct300418h.

[10] Kyle A. Barlow, Shane Ó Conchúir, Samuel Thompson, Pooja Suresh, James E. Lucas, Markus Heinonen, and Tanja Kortemme. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *The Journal of Physical Chemistry B*, 122:5389–5399, 2018. doi: 10.1021/acs.jpcb.7b11367.

[11] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters*, 100:020603, 2008. doi: 10.1103/PhysRevLett.100.020603.

[12] Yingda Xu, William Roach, Tingwan Sun, Tushar Jain, Bianka Prinz, Ta-Yi Yu, Joshua Torrey, Jerry Thomas, Piotr Bobrowicz, Maximiliano Vásquez, K. Dane Wittrup, and Eric Krauland. Addressing polyspecificity of antibodies selected from an in vitro yeast presentation system: a FACS-based, high-throughput selection and analytical tool. *Protein Engineering, Design and Selection*, 26:663–670, 2013. doi: 10.1093/protein/gzt047.

[13] Edward P. Harvey, Jung-Eun Shin, Meredith A. Skiba, Genevieve R. Nemeth, Joseph D. Hurley, Alon Wellner, Ada Y. Shaw, Victor G. Miranda, Joseph K. Min, Chang C. Liu, Debora S. Marks, and Andrew C. Kruse. An in silico method to assess antibody fragment polyreactivity. *Nature Communications*, 13(1):7554, 2022. doi: 10.1038/s41467-022-35276-4. PMID:36477674.

[14] Yasmina Noubia Abdiche, Andy Yik Yeung, Irene Ni, Donna Stone, Adam Miles, Winse Morishige, Andrea Rossi, and Pavel Strop. Antibodies Targeting Closely Adjacent or Minimally Overlapping Epitopes Can Displace One Another. *PLOS ONE*, 12(1):e0169535, 2017. doi: 10.1371/journal.pone.0169535. PMID:28060885.

[15] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. *arXiv*, 2022. URL: https://arxiv.org/abs/2208.02814.

[16] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *arXiv*, 2021. URL: https://arxiv.org/abs/2110.01052.

[17] Ying Jin and Emmanuel J. Candès. Selection by Prediction with Conformal p-values. *arXiv*, 2022. URL: https://arxiv.org/abs/2210.01408.

[18] Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022. doi: 10.1073/pnas.2204569119. PMID:36256807.

[19] Jasmin F. Sydow, Florian Lipsmeier, Vincent Larraillet, Maximiliane Hilger, Bjoern Mautz, Michael Mølhøj, Jan Kuentzer, Stefan Klostermann, Juergen Schoch, Hans R. Voelger, Joerg T. Regula, Patrick Cramer, Apollon Papadimitriou, and Hubert Kettenberger. Structure-Based Prediction of Asparagine and Aspartate Degradation Sites in Antibody Variable Regions. *PLoS ONE*, 9(6):e100736, 2014. doi: 10.1371/journal.pone.0100736. PMID:24959685.

[20] Jochem Gokemeijer, Vibha Jawa, and Shibani Mitra-Kaushik. How Close Are We to Profiling Immunogenicity Risk Using In Silico Algorithms and In Vitro Methods?: an Industry Perspective. *The AAPS Journal*, 19:1587–1592, 2017. doi: 10.1208/s12248-017-0143-z.

[21] Birkir Reynisson, Bruno Alvarez, Sinu Paul, Bjoern Peters, and Morten Nielsen. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, 48(W1):W449–W454, 2020. doi: 10.1093/nar/gkaa379. PMID:32406916.

[22] Jonas B. Nilsson, Saghar Kaabinejadian, Hooman Yari, Michel G. D. Kester, Peter van Balen, William H. Hildebrand, and Morten Nielsen. Accurate prediction of HLA class II antigen presentation across all loci using tailored data acquisition and refined machine learning. *Science Advances*, 9 (47):eadj6367, 2023. doi: 10.1126/sciadv.adj6367. PMID:38000035.

[23] Yu-Xi Tsai, Ning-En Chang, Klaus Reuter, Hao-Ting Chang, Tzu-Jing Yang, Sören von Bülow, Vidhi Sehrawat, Noémie Zerrouki, Matthieu Tuffery, Michael Gecht, Isabell Louise Grothaus, Lucio Colombi Ciacchi, Yong-Sheng Wang, Min-Feng Hsu, Kay-Hooi Khoo, Gerhard Hummer, Shang-Te Danny Hsu, Cyril Hanus, and Mateusz Sikora. Rapid simulation of glycoprotein structures by grafting and steric exclusion of glycan conformer libraries. *Cell*, 187(5):1296–1311.e26, 2024. doi: 10.1016/j.cell.2024.01.034. PMID:38428397.

[24] A. Wright, M. H. Tao, E. A. Kabat, and S. L. Morrison. Antibody variable region glycosylation: position effects on antigen binding and carbohydrate structure. *The EMBO Journal*, 10:2717–2723, 1991. doi: 10.1002/j.1460-2075.1991.tb07819.x.

[25] Gwo-Yu Chuang, Mangaiarkarasi Asokan, Vera B. Ivleva, Amarendra Pegu, Eun Sung Yang, Baoshan Zhang, Rajoshi Chaudhuri, Hui Geng, Bob C. Lin, Mark K. Louder, Krisha McKee, Sijy O'Dell, Hairong Wang, Tongqing Zhou, Nicole A. Doria-Rose, Lisa A. Kueltzo, Q. Paula Lei, John R. Mascola, and Peter D. Kwong. Removal of variable domain-linked glycosylation as a means to improve the homogeneity of HIV-1 broadly neutralizing antibodies. *mAbs*, 12:1836719, 2020. doi: 10.1080/19420862.2020.1836719.

[26] Aubin Ramon, Montader Ali, Misha Atkinson, Alessio Saturnino, Kieran Didi, Cristina Visentin, Stefano Ricagno, Xing Xu, Matthew Greenig, and Pietro Sormanni. Assessing antibody and nanobody nativeness for hit selection and humanization with AbNatiV. *Nature Machine Intelligence*, 6(1):74–91, 2024. doi: 10.1038/s42256-023-00778-3.

[27] Aubin Ramon, Mingyang Ni, Olga Predeina, Rebecca Gaffey, Patrick Kunz, Shimobi Onuoha, and Pietro Sormanni. Prediction of protein biophysical traits from limited data: a case study on nanobody thermostability through NanoMelt. *mAbs*, 17:2442750, 2025. doi: 10.1080/19420862.2024.2442750.

[28] Jerome Anthony E. Alvarez and Scott N. Dean. TEMPRO: nanobody melting temperature estimation model using protein embeddings. *Scientific Reports*, 14:19074, 2024. doi: 10.1038/s41598-024-70101-6.

[29] David Prihoda, Jad Maamary, Andrew Waight, Veronica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A. Bitton. 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. PMID:35133949.

[30] Matthew I. J. Raybould, Claire Marks, Konrad Krawczyk, Bruck Taddese, Jaroslaw Nowak, Alan P. Lewis, Alexander Bujotzek, Jiye Shi, and Charlotte M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116 (10):4025–4030, 2019. doi: 10.1073/pnas.1810576116. PMID:30765520.

[31] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. *arXiv*, 2019. URL: https://arxiv.org/abs/1908.08729.

[32] Q. Huangfu and J. A. J. Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10:119–142, 2018. doi: 10.1007/s12532-017-0130-5.

[33] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020. doi: 10.1038/s41592-019-0686-2. PMID:32015543.

[34] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/bf01588971.

[35] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. *arXiv*, 2022. URL: https://arxiv.org/abs/2210.02747.

[36] Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tally Portnoi, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, Samuel J. Snyder, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Haoyu Fan, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward Universal Binder Design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494.

[37] Sheng Xu, Qiantai Feng, Lifeng Qiao, Hao Wu, Tao Shen, Yu Cheng, Shuangjia Zheng, and Siqi Sun. Benchmarking all-atom biomolecular structure prediction with FoldBench. *Nature Communications*, 17(1):442, 2025. doi: 10.1038/s41467-025-67127-3. PMID:41345395.

[38] Simon Levine, Jonathan Edward King, Jacob Stern, David Grayson, Raymond Wang, Rui Yin, Umberto Lupo, Paulina Kulytė, Ryan Matthew Brand, Tristan Bertin, Robert Pfingsten, Jovan Cejovic, Chelsea Chung, Breanna K. Luton, Andrew Hagemann, Robel Haile, Elliot Medina, Pankaj Panwar,

Oleksii Dubrovskyi, Chase LaCombe, Zahra Anderson, Derrik Mildh, Scott Benjamin, Joe Kaiser, Joseph Ferron, Marta Sarrico, Alexandria Kershner, Apurva Mishra, Kai R. Ejan, Emily K. Marsh, Paul Bringas, Phetsamay Vilaychack, Kyra Chapman, Jacob Ripley, Muttappa Gowda, Kathryn M. Collins, Cailen M. McCloskey, Jeremiah S. Joseph, Rylee Ripley, Shaheed A. Abdulhaqq, David Spencer, Tiffany DeVine, Audree Feltner, Michael Guerin, Jeffrey Goby, Jesse Hendricks, Danielle Castillo, Sean McClain, Douglas Ganini, Derek Shpiel, James Mategko, Eder Cruz Garcia, Masoud Zabet-Moghaddam, John M. Sutton, Zheyuan Guo, Sean M. West, Janani S. Iyer, and Amir Shanehsazzadeh. Origin-1: a generative AI platform for de novo antibody design against novel epitopes. *bioRxiv*, 2026. doi: 10.64898/2026.01.14.699389.

[39] Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling Atomistic Protein Binder Design with Generative Pretraining and Test-Time Compute. *arXiv*, 2026. URL: https://arxiv.org/abs/2603.27950.

[40] Anton Bushuiev, Roman Bushuiev, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Revealing data leakage in protein interaction benchmarks. *arXiv*, 2024. URL: https://arxiv.org/abs/2404.10457.

[41] Leon Hermann, Tobias Fiedler, Hoang An Nguyen, Melania Nowicka, and Jakub M. Bartoszewicz. Beware of Data Leakage from Protein LLM Pretraining. *bioRxiv*, 2024. doi: 10.1101/2024.07.23.604678.

[42] Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16(1):3337, 2025. doi: 10.1038/s41467-025-58606-8. PMID:40199913.

[43] M. Frank Erasmus, Daniel Bedinger, Elizabeth Hopkins, Ginger Ferguson, Justine Strickler, Christilyn P. Graff, Samantha R. Summers, Stacy L. Capehart, Joshua D. Slocum, Crystal Richardson, Sumit Kumar, Zhifei Sun, Yujie Shang, Jixian Zhang, Ming Gu, Lixia Yi, Alon Wellner, Shuangjia Zheng, Wei Lu, Pietro Sormanni, Matthew Greenig, Haiping Zhang, Brendan T. Mann, Mahdi Baghbanzadeh, Ali Rahnavard, Gregory L. Moore, Huaiyu Sun, Ying Ding, Alex Nisthal, Jitendra Kanodia, Matthew J. Bernett, Aurélien Pélissier, Yanjun Shao, Maria Rodriguez Martinez, Karthik Ramesh, Horacio Nastri, Andreas Evers, Anhar Abdelatif, Andrew J. Bordner, Mykola Bordyuh, Lim Heo, Brian A. Kidd, H. Serhat Tetikol, Shuai Wei, Jung-Eun Shin, Ryan Peckner, Leigh Manley, Ajitesh Lunge, Yashas Devasurmutt, Bora Guloglu, Liviu Copoiu, Miles McGibbon, Monica L. Fernandez-Quintero, Nitesh Mishra, Sean M. Callaghan, Olivia M. Swanson, Daniel L. V. Bader, James A. Ferguson, Sai S. R. Raghavan, Benjamin Nemoz, Colleen A. Maillie, Charles Bowman, Bryan Briney, Andrew B. Ward, Paolo Marcatili, Rahmad Akbar, Bing He, Fandi Wu, Jianhua Yao, Bin Hu, Michal Kucer, Kaetlyn Rose Gibson, Rahul Somasundaram, Li-Wei Hung, Tomasz Kaszuba, Daved H. Fremont, Hyeongsun Jeong, Vinodh Babu Kurella, Shipra Malhotra, Satyendra Kumar, Yanyun Liu, Lingling Xu, Joshua Misa, Alexander Nicholas St. John, Jeff Vogt, Fátima A. Dávila-Hernández, Da Xu, Michael Chungyoun, Zyaja D. Huggan, Jeffrey J. Gray, Jonathan Parkinson, Young Su Ko, Wei Wang, Franziska Geiger, Jonathon D. Ziegler, Nikhil Haas, Chance Challacombe, Ahmad Qamar, Akshita Singh, Yi-Ching Tang, Zhiqiang An, Xiaoqian Jiang, Yejin Kim, Xinyan Zhao, Erik Swanson, Jürgen Klattig, Karsten Winkler, Tschimegma Bataa, Volker Sandig, Lilian Denzler, Chunan Liu, Randall J. Brezski, Laura Spector, Katheryn Perea-Schmittle, Sara D'Angelo, Fortunato Ferrara, and Andrew R. M. Bradbury. A blinded, prospective benchmark of in silico antibody discovery anchored to experimental affinity and developability. *Nature Biotechnology*, 2026. doi: 10.1038/s41587-026-03238-6. PMID:42618805.

[44] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-sts894. arXiv:2210.01948.

[45] C. J. Clopper and E. S. Pearson. THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

[46] Ruodu Wang and Aaditya Ramdas. False Discovery Rate Control with E-values. *Journal of the Royal*

Statistical Society Series B: Statistical Methodology, 84(3):822–852, 2022. doi: 10.1111/rssb.12489. arXiv:2009.02824.

[47] Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual Use of Artificial Intelligence-powered Drug Discovery. *Nature Machine Intelligence*, 4:189–191, 2022. doi: 10.1038/s42256-022-00465-9.

[48] Bruce J. Wittmann, Tessa Alexanian, Craig Bartling, Jacob Beal, Adam Clore, James Diggans, Kevin Flyangolts, Bryan T. Gemler, Tom Mitchell, Steven T. Murphy, Nicole E. Wheeler, and Eric Horvitz. Strengthening nucleic acid biosecurity screening against generative protein design tools. *Science*, 390:82–87, 2025. doi: 10.1126/science.adu8578.

[49] David Baker and George Church. Protein design meets biosecurity. *Science*, 383:349, 2024. doi: 10.1126/science.ado1671.