

Choosing protein design models with limited experimental measurements

Michael Doane

mdoane@alumni.stanford.edu

Independent researcher

Research Article

Keywords: protein engineering, model evaluation, experimental design, deep mutational scanning, sequential sampling, ProteinGym, value of information

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11098459/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Choosing protein design models with limited experimental measurements

Michael Doane

Independent researcher, Cary, North Carolina, USA

ORCID: <https://orcid.org/0009-0003-0521-8981>

Corresponding author: Michael Doane; [email address to be supplied]

Abstract

Background: Protein engineers often need to choose which amino acid variants to make and test. Computational models rank these variants by their expected performance, such as enzyme activity, but different models recommend different sets. Before committing to a larger experiment, a laboratory needs to know whether a recommended set contains enough variants with the desired experimental performance. This study compared ways to choose measurements that answer that question.

Results: Methods were evaluated retrospectively using GB1 and adeno-associated virus data, followed by 184 ProteinGym datasets covering 159 proteins. Each ProteinGym model recommended its highest-ranked 10% of variants. A variant counted as a success if it belonged to the experimentally measured top 10% for the protein property being tested. A recommended set was acceptable if more than 20% of its variants were successes. Known experimental outcomes were hidden and revealed sequentially to simulate measuring selected variants.

Ten measurement policies were compared. Exact sequential value of information (SeqVOI) had the lowest mean measurement cost: 776.0 measurements, compared with 870.5 for a greedy rule prioritizing variants shared among unresolved model recommendations, and 829.6 for a learned switch between these policies. Costs included initial measurements; runs without a correct decision were charged their full budget. Correct-certification rates were 82.8%, 83.8%, and

24 82.6%, respectively. The learned switch reduced cost relative to greedy by 5.60%, averaged
25 across proteins, but its one-sided 97.5% lower confidence bound of 4.31% missed the required
26 5% improvement. Active Testing/LURE recommended sets with success rates closer to the best
27 available model's than the learned switch at every measured budget. Missing original input-
28 preparation records limit reproducibility.

29 **Conclusions:** Existing model recommendations can guide which protein variants to measure
30 when checking whether a proposed library meets a minimum experimental success rate. Exact
31 SeqVOI had the lowest average cost for this decision in the recorded datasets. Choosing the
32 model with the highest success rate favored different policies. The learned switch did not reduce
33 average cost below that of exact SeqVOI.

34 **Keywords:** protein engineering; model evaluation; experimental design; deep mutational
35 scanning; sequential sampling; ProteinGym; value of information

36 **Background**

37 Protein engineering can produce far more candidate sequences than a laboratory can measure.
38 Machine-learning models help prioritize these candidates, but different models can recommend
39 different libraries [1, 2]. The resulting choice has practical consequences. A model that ranks
40 variants well across a broad benchmark may still propose too few successful designs for a
41 particular campaign. Measurements are therefore needed to assess the proposed library before
42 committing further experimental capacity.
43 The decision considered here is whether a model's proposed variants contain a sufficient fraction
44 of experimental hits. A hit is a variant meeting a specified assay target. A model's design yield is
45 the expected fraction of hits obtained by sampling from its proposed variants. Establishing that
46 yield exceeds a minimum can support proceeding with a library. Establishing that every

47 candidate model falls below the minimum can support stopping or changing the design strategy.
48 A separate decision is which model has the highest yield at a fixed measurement budget. These
49 objectives can favor different measurement rules because a satisfactory model may be easier to
50 identify than the best model.

51 Fannjiang and Park formalized reliable algorithm selection for machine-learning-guided design
52 [3]. Their framework combines predictions and held-out measurements to assess candidate
53 design distributions when the relevant sampling-probability ratios are known. The present study
54 uses that decision setting to examine sequential measurements from a known finite collection.
55 Every candidate assigns known probabilities to its proposed variants. An observed hit contributes
56 a known amount to the candidate's yield, while unobserved outcomes contribute an uncertain
57 remainder. Overlapping proposals allow one measurement to inform several candidate models.
58 The analysis compares ways of choosing those measurements and calculating the evidence
59 needed for a decision. Greedy acquisition selects variants that account for the most unresolved
60 design probability. Exact SeqVOI uses predicted outcomes to estimate how much a measurement
61 could advance certification. A learned controller, Guarded-SeqVOI, chooses between these two
62 rules using information available before the next batch is revealed. Comparators include random
63 sampling and adaptations of active model-evaluation methods [4–7]. Both greedy and SeqVOI
64 update their choices as measurements arrive.

65 GB1 and adeno-associated virus (AAV) provide controlled comparisons of query access,
66 evidence calculation, and ordering. ProteinGym provides a broader retrospective evaluation
67 using published mutation-effect scores and experimental measurements [8]. Deep mutational
68 scanning makes these comparisons possible by recording outcomes for many variants of the
69 same protein [9]. The contribution is a comparison of measurement strategies for this biological

70 decision, including their certificate frequency, label cost, and recommendation quality. The
71 learned switch’s specified efficacy test is retained alongside comparisons with its component
72 policies. No new protein scoring model or laboratory assay was developed.

73 **Methods**

74 **Experimental data and study design**

75 GB1 uses the measured landscape of Wu et al. through the official Fannjiang–Park
76 implementation [3, 10]. AAV uses the validation-chip data associated with Bryant et al. [11].
77 The initial GB1 experiment assessed the additional benefit of SeqVOI over greedy acquisition.
78 Each original landscape study used 200 development and 1000 confirmatory trials in each of
79 three conditions: several acceptable candidate models, one detectable acceptable model, or no
80 acceptable model. These conditions refer to model yields relative to a threshold, rather than the
81 number of successful individual variants.

82 The subsequent GB1/AAV comparison suite contains 7900 paired trial bundles in 55
83 experimental settings, with eight methods per bundle. Six main settings cover the three
84 conditions in each landscape, with 500 trials per setting. Another 49 settings contain 100 trials
85 each and vary the decision threshold, initial sample size, batch size, or AAV design-set size.
86 Canonical settings use 1000 initial random draws and subsequent budgets of 4000 for GB1 and
87 8000 for AAV, in batches of 50. The candidate menus contain 21 GB1 and nine AAV
88 procedures. Each canonical AAV procedure selects 2000 variants. Thresholds are constructed
89 from the recorded truth to create controlled decision conditions.

90 The broader analysis uses retained results associated with ProteinGym release v1.3, including its
91 substitution-assay labels and published zero-shot model scores [8, 12]. Models and their scores
92 remain fixed; no protein language model is trained or executed for this analysis. Such scores

93 provide rankings without fitting a new model to each assay [13]. The saved selection criteria
94 exclude GB1, AAV, and the specified development assays. They require at least 1000 usable
95 variants, at least six candidate models with finite scores on 95% of variants, at least 50 hits and
96 50 nonhits, and a hit fraction between 5% and 20%.

97 A usable variant has a finite experimental score. Duplicate mutant identifiers retain the first
98 official row. Experimental percentiles are ascending ranks divided by the usable variant count,
99 with average ranks for ties. A hit has percentile strictly above 0.90. The target therefore
100 represents the assay's experimental top decile, determined from the complete recorded outcomes
101 before the simulation. It is a retrospective target, rather than an unknown quantile estimated
102 during acquisition.

103 Each ProteinGym candidate selects its highest-scoring 10% of variants, rounding the set size
104 upward, and assigns equal probability within that set. Missing model scores are excluded and
105 score ties retain official row order. Score directions follow the official registry; the merged
106 archive already applies these adjustments. A yield threshold of 0.20 asks whether more than 20%
107 of the proposed set belongs to the experimental top decile. The outcome defining an individual
108 hit and the yield threshold defining an acceptable model are distinct.

109 The retained cohort contains 184 assays, 159 UniProt protein groups, and 95 model-score
110 configurations per assay. Related model sizes and ensembles are included, so the menu contains
111 fewer independent model families. Sixteen trials per assay produce 2944 paired bundles and
112 29440 method records. Assays sharing a UniProt identifier form one training and evaluation
113 group. A later application of the saved criteria reproduced the same assay list from the 217-entry
114 reference. The original preparation files remain unavailable, and HIS7 has a specific input-source

115 discrepancy described below. The reported results consequently concern the recorded cohort.

116 The reconstructed selection ledger accompanies Additional file 2.

117 **Establishing whether a design set meets the target**

118 Let G_i equal one when variant i is a hit and zero otherwise. Candidate k assigns probability $q_k(i)$

119 to variant i , with nonnegative probabilities summing to one over the N variants. Its true design

120 yield is

$$121 \quad \theta_k = \sum_{i=1}^N q_k(i) G_i. \quad (1)$$

122 For observed variant identities O_t , the smallest possible yield is the probability mass already

123 assigned to observed hits:

$$124 \quad L_{k,t} = \sum_{i \in O_t} q_k(i) G_i. \quad (2)$$

125 The largest possible yield allows every unobserved variant to be a hit:

$$126 \quad U_{k,t} = L_{k,t} + \sum_{i \notin O_t} q_k(i). \quad (3)$$

127 For example, a uniform design set of 20 variants with five observed hits and seven observed

128 failures has a yield between 0.25 and 0.65. The eight unobserved variants determine the

129 remaining uncertainty. These bounds contain the recorded yield for every possible set of

130 observations, including sets chosen adaptively. Exact refers to this finite-population accounting.

131 It requires no calibrated outcome predictor. A first measurement narrows every relevant

132 candidate's interval by the probability assigned to that variant; measuring the same identity again

133 adds no information.

134 A positive certificate establishes that a candidate exceeds the required yield. A null certificate

135 establishes that every candidate falls below it. ProteinGym uses a numerical margin of 10^{-10}

136 around 0.20; equality establishes neither decision. The deterministic containment argument
137 applies in exact arithmetic. Its implementation requires correct identities, normalized
138 probabilities, and correct updates; the margin is not a proved bound on floating-point error.
139 Additional file 1 gives the proof, boundary cases, and the weaker numerical checks used in the
140 exported error flags.

141 Anytime SeqVOI instead samples randomly without replacement and accumulates evidence that
142 a candidate's yield exceeds the threshold. Predictions for unmeasured outcomes provide a
143 baseline; probability-weighted observed residuals correct its conditional mean [14–16]. Sampling
144 probabilities and predictions use only information available before each draw. Every remaining
145 variant proposed by any candidate has positive sampling probability. A mixture of nonnegative
146 betting processes controls false positive certification across repeated checks and the fixed
147 candidate menu, following established anytime-inference principles [17–20]. The error level is
148 0.025 in ProteinGym and 0.10 in the GB1/AAV comparison suite. The implemented anytime
149 procedure issues positive certificates only. It has no null decision or exact-bound fallback.

150 Additional file 1 specifies the construction. Both types of guarantee concern recorded outcomes,
151 conditional on the stated assumptions about sampling and bookkeeping.

152 **Measurement budgets and acquisition rules**

153 ProteinGym begins with a shared uniform sample of approximately 2% of each assay, bounded
154 between 32 and 128 variants and below the assay size. The post-initial budget is the smallest of
155 3000 labels, half the assay rounded upward, and the number of unobserved variants. Batches
156 contain at most 16 variants. Within each trial, methods share the initial observations. Master seed
157 260904 and explicit assay, trial, and method identifiers define their random streams.

158 The locked greedy policy averages design probabilities across unresolved candidates and
159 measures the largest remaining values. Stable index order resolves ties. Its rule and settings were
160 fixed before evaluation; the candidate set changes as certificates or eliminations occur. Exact
161 SeqVOI estimates the immediate decrease in a weighted uncertainty score after a potential hit or
162 failure. A candidate receives zero unresolved width in that score when the proposed outcome
163 reaches its nominal decision threshold. The score combines these gains using a predicted hit
164 probability and current candidate weights. It is a local expected-gain approximation, without
165 optimization over future sequences of measurements.

166 The hit predictor places variants in ten bins using their mean percentile across public model
167 scores. Observed hit rates within those bins are smoothed toward the overall observed rate.

168 Candidate working means are smoothed hit rates within their proposed sets. These means
169 determine relative weights for acquisition and most recommendations. The weights are heuristic
170 scores; they are not calibrated probabilities that a candidate is best. Additional file 1 supplies the
171 smoothing, weighting, and proposal formulas.

172 Anytime SeqVOI uses a randomized proposal based on predicted outcome variance and
173 candidate design probabilities, with 10% of its probability allocated to a shared design-
174 probability component. This preserves support over the union of candidate sets. The reference-
175 replay control samples uniformly with replacement from the full ProteinGym assay and charges
176 repeated identities. It represents a fixed reference proposal, rather than an observed laboratory
177 sampling process. Passive random sampling uses the full assay without replacement. Prediction-
178 disagreement sampling prioritizes disagreement in whether models include a variant in their
179 proposed sets. These controls use the common exact certification rules.

180 Recommendations are recorded at 5%, 10%, 20%, 40%, and 100% of each post-initial budget.
181 For most policies, the recommendation is the candidate with the largest working mean among
182 those certified by the exact bounds, if any, and otherwise among all candidates. Comparator-
183 specific estimators are described below. The simulation checks certification after each revealed
184 item, saves the first decision and its label count, and continues acquisition to obtain later
185 recommendations. Acquisition choices update at batch boundaries. These continued curves
186 assess the value of additional measurements after an early certificate; a campaign that actually
187 stops at that certificate would incur no later acquisition cost.

188 **Learning when to use SeqVOI**

189 Guarded-SeqVOI chooses either the greedy batch or the exact SeqVOI batch and retains the
190 same exact bounds. The classifier has six inputs: how evenly candidate weights are distributed,
191 the gap between the two leading working means, disagreement among proposed sets, the
192 observed hit rate, the average fraction of each proposed set already measured, and the remaining
193 budget fraction. All are computed before the next outcomes are revealed. The observed hit rate
194 describes the queried sample; it is not an unbiased estimate of assay-wide prevalence. The
195 coverage feature is stored under the legacy name effective sample size, but does not estimate an
196 inferential effective sample size.

197 Training uses one greedy and one exact-SeqVOI trajectory per assay, ending at a certificate or
198 after 32 batches. Recorded training outcomes score both proposed next batches. The training
199 utility rewards lower summed unresolved widths and adds a bonus of two for an additional
200 terminal decision. A positive utility labels the state as favorable to SeqVOI. Thus training targets
201 the next batch, whereas the primary evaluation concerns total labels to certification.

202 For each evaluation protein group, all its training states are excluded. Inner leave-one-protein-
203 group-out validation tunes standardized, class-balanced logistic regression over inverse
204 regularization strengths 0.1, 1, and 10, and choice thresholds 0.45, 0.55, and 0.65. Scaling uses
205 the current training fold only. Tuning maximizes mean local utility with equal group weighting;
206 ties favor fewer SeqVOI choices, stronger regularization, and the larger threshold. Refitting on
207 the outer training groups produces the evaluation controller. This nested procedure separates
208 tuning and evaluation by protein identity [21, 22]. It does not exclude homologous proteins from
209 different groups.

210 Training trajectories cover at most 512 post-initial labels, while evaluation can reach 3000.
211 Switching can also create histories absent from either fixed-policy training trajectory.
212 Retrospective diagnostics reveal outcomes for the unchosen proposal after the action has been
213 selected; these outcomes do not determine the held-out action. A laboratory implementation
214 would omit these diagnostic reveals. The term guarded denotes the switch to greedy. Exact
215 bounds preserve certificate validity under either choice, but no result bounds the switch’s label
216 cost relative to greedy.

217 **Published comparator adaptations**

218 Active Testing/LURE combines informative acquisition with bias-corrected estimation at fixed
219 observation counts [4, 5]. Its proposal uses average surrogate binary classification loss and 10%
220 uniform sampling support. The yield estimator applies LURE to the population remaining after
221 the initial sample and adds the initial sample’s known contribution. Estimates remain unclipped
222 and determine recommendations; common exact bounds determine certificates. A correction to
223 the original implementation restored that initial contribution and used the correct remaining-
224 population size. All reported LURE results use the corrected tables. The estimator’s fixed-count

225 unbiasedness does not automatically apply to a selected maximizer or a data-dependent stopping
226 time.

227 The ASE-XWED adaptation uses an eight-member beta-bin ensemble and expected weighted
228 disagreement to choose measurements, combining observed outcomes with predictions for
229 unobserved variants [7]. The online active model-selection adaptation uses pool-based vote
230 uncertainty and exponential weight updates from squared design-membership loss, with learning
231 rate 0.25 [23]. Both use common exact certification. Their surrogate models and acquisition
232 settings differ from the original published tasks. In particular, the implemented LURE proposal
233 differs from Hara et al.’s pairwise-loss-difference proposal for model selection [6]. Conclusions
234 are limited to the stated adaptations.

235 **Outcomes and statistical analysis**

236 The primary cost counts the initial sample and subsequent labels to the first recorded certificate
237 when that certificate is correct. An incorrect first certificate or no certificate within budget incurs
238 the full budget. This capped cost is observed for every run and does not estimate how many
239 further measurements an unresolved run would require. Correct-certification frequency is
240 reported alongside cost to show the fraction of trials producing a useful decision.

241 For each assay, costs are first averaged across its 16 trials. The guard’s relative reduction for
242 assay a is

$$243 \quad r_a = 1 - \frac{\bar{c}_{a,\text{guard}}}{\bar{c}_{a,\text{greedy}}}. \quad (4)$$

244 Assay reductions are averaged within proteins, then equally across the 159 proteins. This gives
245 each protein the same weight and differs from the ratio of overall mean costs. Selection accuracy
246 is the proportion of recommendations attaining the maximum recorded yield within 10^{-12} .

247 Regret is the best yield minus the recommended yield; 0.01 represents a one-percentage-point
248 shortfall in expected hit fraction. Primary accuracy and regret differences average trials and the
249 five budgets before averaging within and across proteins. Reporting by protein limits the
250 influence of multiple assays from the same source [24].

251 The archived analysis uses 10000 bootstrap resamples of protein summaries [25]. Its 2.5th and
252 97.5th percentiles supply nominal one-sided 97.5% bounds. Fitted controllers remain fixed, and
253 outer folds share training data. These bounds omit retraining variation and have no established
254 coverage for a newly fitted controller evaluated on a new protein population.

255 The specified guard success rule requires a lower bound of at least 5% for relative label saving, a
256 nonnegative lower bound for selection-accuracy difference, and a nonpositive upper bound for
257 regret difference. At least 70% of assays must also satisfy an observed joint rule: mean cost
258 reduction of at least -5% , nonnegative accuracy difference, and nonpositive regret difference.

259 This last rule uses observed assay means, without assay-specific confidence tests. Validity and
260 minimum-cohort requirements complete the rule. The saved configuration establishes these
261 criteria; the retained materials do not establish a public preregistration timestamp.

262 GB1/AAV results use a separate endpoint: the relative reduction in pooled mean post-screen
263 costs over the two positive conditions, charging failed positive certification the budget. Bootstrap
264 clusters pair conditions and matched paths by landscape and trial index, giving 1000 pooled
265 clusters. They therefore should not be directly compared with ProteinGym's average of assay-
266 relative total-cost reductions.

267 All-method recommendation curves use 155 protein groups with complete checkpoints. Across
268 four other groups, 64 final-budget anytime checkpoint records are missing after exhaustion of the
269 remaining query support; every method and budget excludes these groups for that comparison.

270 The primary guard-greedy analysis retains all 159 groups. Secondary analyses include
271 component-policy comparisons, an evaluation-only HIS7 exclusion, and a descriptive summary
272 of local acquisition states. A later cost-only sensitivity charges every member of a committed
273 batch by rounding the recorded stopping count upward to the next batch of 16, within the
274 original budget. This sensitivity leaves acquisition paths fixed and supplies no new confidence
275 bounds or success test. Full definitions and tables are in Additional files 1 and 2.

276 **Computational assistance**

277 OpenAI tools assisted with code generation, manuscript drafting and language editing under the
278 author's discretion.

279 **Results**

280 **Targeted measurements reduced certification cost in GB1 and AAV**

281 In the original GB1 experiment, SeqVOI reduced post-screen labels by 2.97% relative to greedy,
282 with a one-sided 97.5% lower bound of 1.83%, below the specified 20% target. The later
283 comparison suite examined which experimental choices could produce larger savings.
284 Targeted exact SeqVOI used 67.02% fewer post-screen labels than exact reference-distribution
285 replay across GB1 and AAV, with a lower bound of 65.57% (Table 1). The comparison changed
286 access to individual unmeasured variants and their ordering together. With the same randomized
287 query paths, exact evidence reduced pooled cost by 2.00% relative to anytime evidence. That
288 effect was substantial on GB1 and approximately zero on AAV. The pooled percentage uses
289 absolute costs and consequently differs from the arithmetic mean of the two landscape
290 percentages.

291 **Table 1. Label reductions in the controlled GB1 and AAV comparisons**

Comparison	GB1 (%)	AAV (%)	Pooled (%)	Pooled lower bound (%)
Targeted exact SeqVOI versus exact reference replay	96.47	52.29	67.02	65.57
Exact versus anytime evidence on matched query paths	45.49	−0.05	2.00	1.76
Exact SeqVOI versus fixed-order exact acquisition	3.91	23.73	23.16	21.88

292 Positive reductions favor the first method. Costs exclude the initial screen and charge failure to
293 certify a success the full post-screen budget. Lower bounds are nominal one-sided 97.5%
294 bootstrap bounds. The first row changes access and ordering; the second holds query paths fixed.
295 The fixed-order control in the third row differs from the stronger mass-based greedy control.
296 Compared with fixed-order exact acquisition, SeqVOI reduced pooled cost by 23.16%. Across
297 the 35 positive-condition sensitivity settings, targeted access and ordering gave nonnegative
298 mean savings in all 35; exact evidence on matched paths did so in 26. These settings vary
299 experimental choices within the same two biological landscapes and do not provide 35
300 independent biological replications.

301 **Exact SeqVOI had the lowest mean capped cost in ProteinGym**

302 The lowest mean capped total cost among the ten policies was 776.0 labels for exact SeqVOI
303 (Table 2). Greedy used 870.5 labels and Guarded-SeqVOI used 829.6. Their correct-certification
304 frequencies were 83.8%, 82.8%, and 82.6% for greedy, exact SeqVOI, and Guarded-SeqVOI,
305 respectively. The cost advantage therefore coexisted with a slightly lower decision frequency
306 than greedy (Figure 1).

307 **Table 2. Certification cost and decision frequency across the recorded ProteinGym cohort**

Policy	Mean total labels	Correct certificate (%)
Greedy	870.5	83.8
Exact SeqVOI	776.0	82.8
Guarded-SeqVOI	829.6	82.6
Anytime SeqVOI	1136.0	74.9
Reference replay	1719.1	15.4
Passive random	1670.4	33.5
Prediction disagreement	956.8	78.3
Active Testing/LURE (corrected)	1305.3	70.2
ASE-XWED adaptation	1675.1	18.3
Online model-selection adaptation	959.6	79.9

308 Means give equal weight to the 159 proteins after averaging trials within assays and assays
309 within proteins. All rows use 184 assays and include initial labels. Unresolved or incorrect first
310 decisions incur the full budget. Exact procedures can issue positive or null certificates; anytime
311 SeqVOI provides positive certification only. The comparator names denote the adaptations
312 described in Methods. These are descriptive comparisons without separately reported paired
313 confidence bounds.

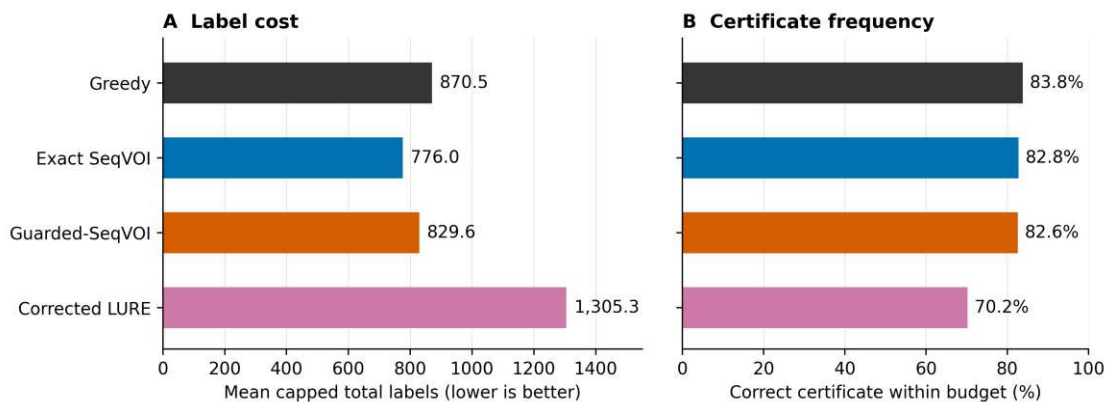


Figure 1. Label cost and correct certification in the retained ProteinGym benchmark. (A)

Capped total labels, including the warm screen, for greedy, exact SeqVOI, Guarded-SeqVOI,

and corrected Active Testing/LURE. A trial without a correct certificate is charged the full

budget. (B) Fraction of trials with a correct certificate by the budget limit. Both panels average

16 trials per assay, assays within UniProt groups, and then 159 groups equally (184 assays).

Values are descriptive means; smaller cost should be interpreted alongside certificate frequency.

The guard saved an average of 40.9 labels per trial relative to greedy, including an average initial

sample of 71.3 labels. Its specified average relative reduction was 5.60%. The standalone exact

policy's relative reduction, calculated descriptively using the same assay-relative endpoint, was

12.04%. This additional comparison retains the original controller test and supplies no new

confirmation claim. Relative reductions and differences in overall mean label counts summarize

the data differently.

The learned switch missed its minimum saving requirement

The guard's one-sided 97.5% lower bootstrap bound for label reduction was 4.31%, below the

required 5% (Table 3). Its budget-averaged selection-accuracy difference relative to greedy was

+6.67 percentage points, with a lower bound of +5.17 points. Its regret difference was -0.00394 ,

with an upper bound of -0.00237 . The observed joint assay rule was met by 133 of 184 assays,

332 or 72.28%. These favorable secondary requirements did not offset the missed minimum-saving
 333 criterion.

334 **Table 3. The specified Guarded SeqVOI requirements and observed outcomes**

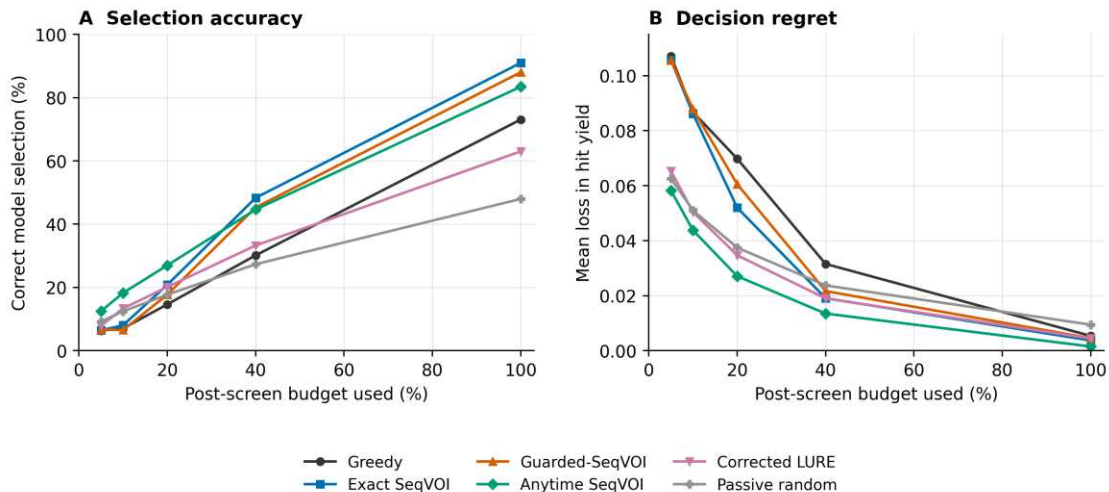
Requirement	Estimate	Decision quantity	Required value	Outcome
Relative total-label saving	5.60%	Lower bound 4.31%	At least 5%	Missed
Budget-averaged accuracy difference	+6.67 pp	Lower bound +5.17 pp	At least 0	Met
Budget-averaged regret difference	−0.00394	Upper bound −0.00237	At most 0	Met
Joint observed assay rule	133/184	72.28% of assays	At least 70%	Met
Anytime false-certificate flags	0/2944	Zero-event upper bound 0.1252%	At most 2.5%	Diagnostic met

335 pp denotes percentage points. Cost includes initial labels; recommendation differences average
 336 five budgets. Bootstrap bounds hold fitted controllers fixed. The joint assay rule uses observed
 337 means without individual assay confidence tests. The zero-event bound is a separate diagnostic
 338 of average flag probability under independent trial randomization. Minimum assay and protein
 339 counts and the zero structural-failure requirement were also met.

340 No structural-failure flags or false-certificate flags were recorded for the reported exact
 341 procedures. Anytime SeqVOI recorded zero false positive flags in 2944 trials. These checks use
 342 implementation-specific tolerances and do not establish a complete numerical audit of every
 343 intermediate bound. The within-trial error argument is separate from the zero-event frequency
 344 diagnostic.

345 **Recommendation quality depended on the budget and decision metric**

346 The common-cohort curves cover 180 assays in 155 protein groups (Figure 2). Exact SeqVOI
347 had 91.0% full-budget selection accuracy, compared with 88.0% for the guard, 83.4% for
348 anytime SeqVOI, 73.1% for greedy, and 63.0% for corrected LURE. Final mean regret was
349 0.00372, 0.00453, 0.00153, 0.00538, and 0.00446, respectively. Thus anytime SeqVOI had the
350 smallest final mean regret among all ten methods despite its higher certification cost and lower
351 exact-maximizer accuracy than exact SeqVOI.



352
353 **Figure 2. Model recommendations improve differently as labels accumulate.** (A) Fraction of
354 recommendations identifying a best model. (B) Loss in true hit yield relative to a best model.
355 Points correspond to 5%, 10%, 20%, 40%, and 100% of each assay's post-screen budget; lines
356 connect recorded checkpoints. Six policies are shown on the same complete cohort of 155
357 UniProt groups (180 assays). All 10 policies appear in the accompanying table. Averages first
358 combine trials within assays and assays within groups. Four groups with missing final anytime-
359 SeqVOI checkpoints are excluded from every curve. LURE uses the corrected estimator. These
360 are descriptive recommendation outcomes; they are distinct from certification.

LURE had higher selection accuracy than the guard at the three earliest budgets and lower mean regret at every budget. In the full 159-group comparison, final mean regret was 0.004420 for the guard and 0.004399 for LURE. The paired guard-minus-LURE difference was 0.000021, with bootstrap endpoints -0.003040 and 0.004570 . These estimates are similar, but the interval does not establish equivalence. A large difference in best-model identification can have little mean yield consequence when alternative models propose nearly as many hits. The LURE estimator repair changed 1003 of 14720 checkpoint recommendations. It left certification costs, certificate decisions, and the guard–greedy primary outcome unchanged. The corrected curves are the basis for all LURE comparisons here. Additional file 2 supplies all-method curves and the separate full-cohort comparison.

Reconstruction and practical cost checks

A later reconstruction reproduced the recorded 184-assay list. Of the 217 reference entries, 29 failed the usable-variant requirement, three were excluded as development-related, and one failed the target requirement. The original prepared arrays and preparation ledger were not recovered. For HIS7, the standalone label archive contains 407055 usable variants, while historical records contain 496137. The experimental outcome column in the official merged-score archive reconstructs the larger population and the retained LURE outputs; predicted scores were not substituted for experimental labels. The original preparation path remains unresolved. Excluding HIS7 from evaluation left 183 assays and 158 groups. The guard’s relative saving was 5.63%, with a lower bound of 4.36%, preserving the missed efficacy criterion. The controllers remained fixed, so HIS7 training states can still contribute to other groups’ fits. This sensitivity isolates the assay’s evaluation contribution.

383 Charging full committed batches reduced the guard's descriptive saving from 5.60% to 5.12%
384 and exact SeqVOI's from 12.04% to 11.34% (Additional file 1, Figure S1). This accounting
385 change preserves recorded query paths and does not include assay prices or repeat
386 measurements. A separate feasibility calculation shows that PHOT and the historical HIS7
387 population cannot produce an exact certificate within their budgets, even with ideal ordering.
388 They require at least 3351 and 9923 observed hits for positive certification, respectively, against
389 at most 3128 total observations.

390 **Discussion**

391 **Implications for protein design campaigns**

392 The practical finding is that proposed design sets contain useful information about where
393 verification measurements should be directed. A variant included by several unresolved models
394 can inform all of them. An unmeasured variant with substantial assigned probability can narrow
395 the possible yield more than a variant that appears in few proposed sets. Exact accounting makes
396 that contribution explicit and allows arbitrary query ordering while retaining deterministic
397 bounds on the recorded population.

398 The GB1/AAV comparisons show large savings for the tested combination of targeted access
399 and ordering. The matched-path comparison shows that changing the evidence calculation alone
400 can have a smaller and landscape-dependent effect. These distinctions matter when translating a
401 computational result to the laboratory. Individually requesting selected variants may require
402 different synthesis or sampling operations from measuring an existing pooled library. The
403 reported access comparison describes those different measurement opportunities together with
404 their ordering rules.

405 The ProteinGym results give a reason to retain simple component policies in the evaluation of
406 learned experimental controllers. The guard used fewer labels than greedy, but always using
407 exact SeqVOI had lower mean capped cost and better final recommendation point estimates. The
408 guard’s training objective rewards immediate progress, whereas the evaluation counts labels until
409 a later decision. Its short training trajectories and the histories created by switching can further
410 weaken that relationship. These point estimates provide no evidence that fitting the controller
411 improved aggregate performance over always using exact SeqVOI. Any proposal to use it only
412 in selected assay classes would require a new evaluation after defining those classes.

413 Certification and recommendation should remain separate outcomes. A certificate can establish
414 that one proposed library is acceptable while leaving the identity of the best library uncertain.
415 Conversely, a model can be recommended accurately without enough evidence to cross a
416 conservative certification threshold. Regret describes the biological consequence of the
417 recommendation as a shortfall in expected hit fraction. Best-model accuracy records whether the
418 exact maximizer was selected. The smallest regret and the lowest certification cost favored
419 different policies in this study, making a single ranking of methods inadequate for every
420 campaign objective.

421 **Relation to existing statistical and active evaluation methods**

422 The decision problem follows reliable design-algorithm selection [3]. The experiments extend its
423 use to sequential queries from a known finite population and compare the resulting measurement
424 costs. Superiority over the original framework would require matching its access assumptions
425 and implementation. Predictive coverage for adaptively designed molecules is another reliability
426 objective, addressed by conformal methods under feedback covariate shift [26]. The present
427 certificates concern a fixed model’s design yield.

428 Local value-of-information acquisition has antecedents in knowledge-gradient and entropy-
429 search methods [27, 28]. Active evaluation methods similarly use predictions to allocate labels
430 and improve performance estimation. Best-model selection is related to best-arm identification,
431 with shared observations here arising from overlap among design sets [29]. This study tests
432 specified adaptations for that setting; it does not establish a new general principle of adaptive
433 sampling or model selection.

434 The learned fallback also has close precedents. BayesAME combines Bayesian active evaluation
435 with a random fallback in specified conditions [30]. Robust sampling and balanced active
436 inference address protection against unhelpful acquisition under their own estimation settings
437 [31, 32]. Their guarantees do not apply to the SeqVOI controller. Hara et al.’s selection-focused
438 LURE proposal is a particularly relevant untested comparator [6]. Holding acquisition fixed
439 while varying the estimator, or holding the estimator fixed while varying acquisition, would help
440 distinguish their contributions. The current comparisons support conclusions about the
441 implemented combinations.

442 **Limits of the retrospective evidence**

443 The recorded labels define both the target and the truth used to score decisions. A future
444 campaign would need to specify a measurable success threshold before acquisition, or allocate
445 separate data to estimating its target. Measurements also contain noise. The exact bounds certify
446 the recorded outcomes; they do not establish the underlying biological response without
447 additional assumptions or replicate measurements. No new laboratory experiment, prospective
448 library, or model training was performed.

449 Grouping by UniProt prevents the same protein identity from entering both controller training
450 and evaluation within a fold. Homologs may remain in different groups, and published models

451 have heterogeneous pretraining histories. Related model configurations also influence consensus
452 and disagreement scores. These choices limit claims about transfer to evolutionarily distant
453 proteins or a differently weighted model menu. ProteinGym and FLIP illustrate how task
454 construction changes the meaning of protein-model generalization [8, 33].

455 Uncertainty is conditional on the retained fits. Protein resampling omits refitting, and outer
456 training sets overlap. The nominal bootstrap bounds therefore describe the configured analysis
457 without established coverage for the full training-and-evaluation procedure on new proteins. The
458 assay-level joint rule is an observed proportion. Neither quantity supplies a guarantee of benefit
459 on a future assay. The descriptive regime map also uses repeated states visited by the guard and
460 bins defined from evaluation data. Only the highest observed-hit-rate bin had positive mean local
461 width reduction for SeqVOI; this is a hypothesis about individual batches, not a validated
462 deployment rule.

463 Reconstruction reproduced the assay identities but left the original preparation history
464 unresolved. HIS7 is the identified population discrepancy, and its evaluation-only exclusion
465 cannot remove any influence on other controller fits. The release permits inspection of recorded
466 outcomes, fitted controllers, corrected comparator tables, and the reconstruction record. It does
467 not establish complete reproduction of the original input preparation. This distinction should
468 remain explicit in subsequent use of the benchmark.

469 **From label counts to experimental cost**

470 Before selecting an acquisition rule, the proposed set size and threshold should be checked
471 against the available measurement budget. The PHOT and HIS7 examples cannot reach the exact
472 certificate even if every available measurement supplies the needed outcome. Their capped costs

473 consequently cannot distinguish how close different exact policies came to certification. A
474 feasible threshold and budget are prerequisites for a useful policy comparison.
475 Initial measurements and committed batches also affect the practical saving. The ProteinGym
476 endpoint includes initial labels, and the added batch sensitivity shows a modest reduction in the
477 reported relative savings when all batch members are charged. Synthesis failures, assay
478 replication, processing time, and the price of targeted versus pooled measurements remain
479 unmeasured. Controller training time and per-batch computational latency were not separately
480 recorded. These quantities would be needed to determine whether a reduction in label count
481 repays the cost of implementing a policy in a specific laboratory.

482 **Conclusions**

483 Sequential acquisition can use existing protein-model rankings and overlap among proposed
484 variants to direct measurements toward a design-quality decision. Targeted access and ordering
485 produced substantial label savings in the GB1/AAV comparisons. In the recorded ProteinGym
486 cohort, exact SeqVOI had the lowest mean capped label cost, while other policies provided
487 different recommendation trade-offs. The learned switch missed its specified minimum-saving
488 criterion and had higher mean cost than standalone exact SeqVOI. These results support
489 evaluating measurement rules against the campaign's actual decision, with certificate frequency,
490 recommendation regret, and total experimental cost reported together. Prospective application
491 requires reproducible inputs, a predefined assay target, and a feasible budget.

492 **Abbreviations**

493 AAV, adeno-associated virus; ASE, active surrogate estimator; DMS, deep mutational scanning;
494 GB1, protein G domain B1; LURE, levelled unbiased risk estimator; SeqVOI, sequential value
495 of information; XWED, expected weighted disagreement.

496 **Declarations**

497 **Ethics approval and consent to participate**

498 Not applicable. This study analysed previously published protein-assay measurements and model
499 scores and involved no new human participant recruitment or animal experiments.

500 **Consent for publication**

501 Not applicable.

502 **Availability of data and materials**

503 Study code and retained analysis artifacts are available in the SeqVOI repository at
504 <https://github.com/MRDOANE/sequential-voi-biomolecular-certification> and the version-
505 specific archive at <https://doi.org/10.5281/zenodo.22420027> [34]. ProteinGym v1.3 is identified
506 by <https://doi.org/10.5281/zenodo.15293562> [12]; GB1 and AAV originate from the cited
507 experimental studies and associated repositories [3, 10, 11]. Upstream data retain their source
508 licenses. Additional file 2 contains the reconstructed assay ledger, revised descriptive tables, and
509 analysis code. The missing original ProteinGym prepared arrays and the HIS7 input-history
510 limitation are documented in Methods and Results. Reconstructed records are identified
511 separately from original exports.

512 **Competing interests**

513 [Author to confirm the competing-interest statement before submission.]

514 **Funding**

515 [Author to confirm the funding statement before submission.]

516 **Authors' contributions**

517 Michael Doane conceived the study, specified the analyses, reviewed the implementation and
518 results, interpreted the findings, and wrote the manuscript.

519 **Acknowledgements**

520 The author thanks the investigators who made the experimental landscapes, benchmark
521 implementations, and model scores publicly available.

522 **Additional files**

523 **Additional file 1: SeqVOI_BMC_Additional_File_1.docx.** Supplementary methods and
524 analyses. Contains exact-bound arguments, the anytime construction, acquisition and corrected
525 LURE details, the descriptive regime analysis, and supplementary figures.

526 **Additional file 2: SeqVOI_BMC_Additional_File_2.zip.** Analysis tables and reconstruction
527 audit. Contains full-precision policy summaries, matched-budget results, the reconstructed 217-
528 entry assay ledger, descriptive cost sensitivity, and scripts documenting the additional
529 calculations. Original large result archives remain in the cited repository and research release.

530 **References**

- 531 1. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering.
532 Nature Methods. 2019;16(8):687–694. <https://doi.org/10.1038/s41592-019-0496-6>.
- 533 2. Brookes D, Park H, Listgarten J. Conditioning by adaptive sampling for robust design. Proceedings of
534 the 36th international conference on machine learning. 2019;97:773–782.
535 <https://proceedings.mlr.press/v97/brookes19a.html>.
- 536 3. Fannjiang C, Park JW. Reliable algorithm selection for machine learning-guided design. Proceedings
537 of the 42nd international conference on machine learning. 2025;267:16078–16101.
538 <https://proceedings.mlr.press/v267/fannjiang25a.html>.

539 4. Kossen J, Farquhar S, Gal Y, Rainforth T. Active testing: Sample-efficient model evaluation.
540 Proceedings of the 38th international conference on machine learning. 2021;139:5753–5763.
541 <https://proceedings.mlr.press/v139/kossen21a.html>.

542 5. Farquhar S, Gal Y, Rainforth T. On statistical bias in active learning: How and when to fix it.
543 International conference on learning representations. 2021. <https://arxiv.org/abs/2101.11665>.

544 6. Hara S, Matsuura M, Honda J, Ito S. Active model selection: A variance minimization approach.
545 Machine Learning. 2024;113:8327–8345. <https://doi.org/10.1007/s10994-024-06603-1>.

546 7. Kossen J, Farquhar S, Gal Y, Rainforth T. Active surrogate estimators: An active learning approach to
547 label-efficient model evaluation. Advances in neural information processing systems. 2022;35.
548 https://proceedings.neurips.cc/paper_files/paper/2022/hash/9b9cfd5428153ccfbd4ba34b7e00730
549 5-Abstract-Conference.html.

550 8. Notin P, Kollasch A, Ritter D, Niekerk L van, Paul S, Spinner H, et al. ProteinGym: Large-scale
551 benchmarks for protein fitness prediction and design. Advances in neural information processing systems.
552 2023;36:64331–64379. Datasets and Benchmarks Track.
553 https://proceedings.neurips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-
554 Abstract-Datasets_and_Benchmarks.html.

555 9. Fowler DM, Fields S. Deep mutational scanning: A new style of protein science. Nature Methods.
556 2014;11:801–807. <https://doi.org/10.1038/nmeth.3027>.

557 10. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is
558 facilitated by indirect paths. eLife. 2016;5:e16965. <https://doi.org/10.7554/eLife.16965>.

559 11. Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, et al. Deep diversification of an AAV
560 capsid protein by machine learning. Nature Biotechnology. 2021;39:691–696.
561 <https://doi.org/10.1038/s41587-020-00793-4>.

562 12. ProteinGym contributors. ProteinGym release v1.3. Zenodo. 2025. Dataset and published model
563 scores. Accessed 10 September 2026. <https://doi.org/10.5281/zenodo.15293562>.

564 13. Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of
 565 the effects of mutations on protein function. *Advances in neural information processing systems*.
 566 2021;34:29287–29303.
 567 [https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-](https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html)
 568 [Abstract.html](https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html).
 569 14. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe.
 570 *Journal of the American Statistical Association*. 1952;47(260):663–685.
 571 <https://doi.org/10.1080/01621459.1952.10483446>.
 572 15. Angelopoulos AN, Bates S, Fannjiang C, Jordan MI, Zrnic T. Prediction-powered inference. *Science*.
 573 2023;382(6671):669–674. <https://doi.org/10.1126/science.adf6000>.
 574 16. Zrnic T, Candès E. Active statistical inference. *Proceedings of the 41st international conference on*
 575 *machine learning*. 2024;235:62993–63010. <https://proceedings.mlr.press/v235/zrnic24a.html>.
 576 17. Howard SR, Ramdas A, McAuliffe J, Sekhon J. Time-uniform, nonparametric, nonasymptotic
 577 confidence sequences. *The Annals of Statistics*. 2021;49(2):1055–1080. [https://doi.org/10.1214/20-](https://doi.org/10.1214/20-AOS1991)
 578 [AOS1991](https://doi.org/10.1214/20-AOS1991).
 579 18. Waudby-Smith I, Ramdas A. Confidence sequences for sampling without replacement. *Advances in*
 580 *neural information processing systems*. 2020;33:20204–20214. <https://arxiv.org/abs/2006.04347>.
 581 19. Ramdas A, Grünwald P, Vovk V, Shafer G. Game-theoretic statistics and safe anytime-valid
 582 inference. *Statistical Science*. 2023;38(4):576–601. <https://doi.org/10.1214/23-STS894>.
 583 20. Waudby-Smith I, Ramdas A. Estimating means of bounded random variables by betting. *Journal of*
 584 *the Royal Statistical Society Series B: Statistical Methodology*. 2024;86(1):1–27.
 585 <https://doi.org/10.1093/jrsssb/qkad009>.
 586 21. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in
 587 performance evaluation. *Journal of Machine Learning Research*. 2010;11:2079–2107.
 588 <https://jmlr.org/papers/v11/cawley10a.html>.

589 22. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC
590 Bioinformatics. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>.

591 23. Reza Karimi M, Gürel NM, Karlaš B, Rausch J, Zhang C, Krause A. Online active model selection
592 for pre-trained classifiers. Proceedings of the 24th international conference on artificial intelligence and
593 statistics. 2021;130:307–315. <https://proceedings.mlr.press/v130/reza-karimi21a.html>.

594 24. Demšar J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning
595 Research. 2006;7:1–30. <https://www.jmlr.org/papers/v7/demsar06a.html>.

596 25. Efron B. Bootstrap methods: Another look at the jackknife. The Annals of Statistics. 1979;7(1):1–26.
597 <https://doi.org/10.1214/aos/1176344552>.

598 26. Fannjiang C, Bates S, Angelopoulos AN, Listgarten J, Jordan MI. Conformal prediction under
599 feedback covariate shift for biomolecular design. Proceedings of the National Academy of Sciences.
600 2022;119(43):e2204569119. <https://doi.org/10.1073/pnas.2204569119>.

601 27. Frazier PI, Powell WB, Dayanik S. A knowledge-gradient policy for sequential information
602 collection. SIAM Journal on Control and Optimization. 2008;47(5):2410–2439.
603 <https://doi.org/10.1137/070693424>.

604 28. Hennig P, Schuler CJ. Entropy search for information-efficient global optimization. Journal of
605 Machine Learning Research. 2012;13:1809–1837. <https://jmlr.org/papers/v13/hennig12a.html>.

606 29. Kaufmann E, Cappé O, Garivier A. On the complexity of best-arm identification in multi-armed
607 bandit models. Journal of Machine Learning Research. 2016;17(1):1–42.
608 <https://jmlr.org/papers/v17/kaufman16a.html>.

609 30. Cordero Encinar P, Cemgil T, Doucet A, Aglietti V, Chiappa S. BayesAME: Bayesian active model
610 evaluation. 2026. Preprint, version 1, 29 July 2026. <https://doi.org/10.48550/arXiv.2607.27023>.

611 31. Li P, Zrnic T, Candès E. Robust sampling for active statistical inference. Advances in neural
612 information processing systems. 2025;38. <https://doi.org/10.52202/085713-2312>.

613 32. Chen B, Zhou Z, Peng L, Wang Z. Balanced active inference. Advances in neural information
614 processing systems. 2025;38. <https://doi.org/10.52202/085713-1724>.

615 33. Dallago C, Mou J, Johnston KE, Wittmann BJ, Bhattacharya N, Goldman S, et al. FLIP: Benchmark
616 tasks in fitness landscape inference for proteins. Proceedings of the neural information processing systems
617 track on datasets and benchmarks. 2021;1. [https://datasets-benchmarks-](https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/2b44928ae11fb9384c4cf38708677c48-Abstract-round2.html)
618 [proceedings.neurips.cc/paper/2021/hash/2b44928ae11fb9384c4cf38708677c48-Abstract-](https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/2b44928ae11fb9384c4cf38708677c48-Abstract-round2.html)
619 [round2.html](https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/2b44928ae11fb9384c4cf38708677c48-Abstract-round2.html).

620 34. Doane M. SeqVOI: Sequential finite-population certification for biomolecular design. Zenodo. 2026.
621 Software and retained results, version 0.4.3. Accessed 10 September 2026.
622 <https://doi.org/10.5281/zenodo.22420027>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SeqVOIBMCAdditionalFile1.docx](#)
- [SeqVOIBMCAdditionalFile2.zip](#)