Impact of Interval Censoring on Data Accuracy and Machine Learning Performance in Biological High-Throughput Screening

Vanni Doffini[1,2,3,*] **Michael A. Nash**[1,2,3,†]
[1]*University of Basel* [2]*ETH Zurich* [3]*Swiss Nanoscience Institute*

Abstract

High-throughput screening (HTS) combined with deep mutational scanning (DMS) and next-generation DNA sequencing (NGS) have great potential to accelerate discovery and optimization of biological therapeutics. Typical workflows involve generation of a mutagenized variant library, screening/selection of variants based on phenotypic fitness, and comprehensive analysis of binned variant populations by NGS. However, in such cases, the HTS data are subject to interval censoring, where each fitness value is calculated based on the assignment of variants to bins. Such censoring leads to increased uncertainty, which can impact data accuracy and, consequently, the performance of machine learning (ML) algorithms tasked with predicting sequence-fitness pairings. Here, we investigated the impact of interval censoring on data quality and ML performance in biological HTS experiments. We theoretically analyzed the impact of data censoring and propose a dimensionless number, the *Ratio of Discretization* (R_D), to assist in optimizing HTS parameters such as the bin width and the sampling size. This approach can be used to minimize errors in fitness prediction by ML and to improve the reliability of these methods. These findings are not limited to biological HTS techniques and can be applied to other systems where interval censoring is an advantageous measurement strategy.

1 Introduction

High-throughput screening (HTS) [43, 1, 27, 55, 6] has profoundly transformed the molecular engineering field by enabling the rapid assessment of thousands to millions of protein variants for key properties such as enzymatic activity and ligand binding. When coupled with advanced methodologies such as deep mutational scanning (DMS) [18, 42, 24] and next-generation DNA sequencing (NGS) [17, 40, 19, 16], HTS becomes a powerful tool for the identification and evolution of novel therapeutics from vast mutagenized libraries. This integrated approach not only accelerates the discovery process but also reduces associated costs, making it indispensable for modern biotechnology.

A typical workflow for variant discovery is based on fluorescent-activated single-cell sorting (FACS) [13, 35, 46] coupled with NGS, and includes three main steps [47, 48]: (i) generation of a mutagenized library of protein variants containing sequence alterations; (ii) single-cell sorting of the variant pool into different bins based on their fluorescence intensity by FACS, and (iii) analysis of the populations in each bin using NGS. In this case, the screening assay is set up such that the fluorescence intensity of the cell reports on the relevant phenotype, for example, by binding to a fluorescently labeled ligand [22]. This process establishes a genotype-phenotype relationship for each mutant screened and enables the library to be sorted according to the desired fitness trait. Subsequently, the best variants can be propagated to subsequent screening rounds and the procedure repeated to further optimize

*vanni.doffini@unibas.ch
†michael.nash@unibas.ch

the selected sequences in a process known as directed evolution (DE) [36, 4, 51]. This approach has proven successful in improving a range of phenotypic properties, such as enzymatic activity [9] and efficiency [39], ligand binding affinity [7, 25], and thermal stability [5].

Modern variations on this technique [38, 53, 33] have sought to leverage the enormous datasets generated by DMS in order to train machine learning (ML) models to navigate and predict the fitness landscape *in silico* [10, 45, 37]. If such an approach were successful, reliable and broadly generalizable, it would enable efficient multi-objective variant optimization [14] and save significant time and cost, thereby enhancing the efficiency and scope of protein engineering for human therapies. Indeed, ML has been employed to predict beneficial mutations, reduce the experimental search space, and design proteins with desired properties more efficiently [52, 30]. For example, ML models have been applied to enzyme engineering [41] and to predict and design antibodies with high affinity for their targets [28, 44]. Furthermore, ML is not limited to mutagenized data but can also be utilized to generate de-novo proteins [50, 26, 54, 3]. However, the field is still in its infancy. Many challenges exist regarding optimal encoding strategies, algorithm classes, scalability and efficiency, and the literature is plagued by a survivor bias of the successful methods implemented within a limited context.

One challenge that has thus far received only limited attention is that datasets obtained via FACS/NGS assays typically classify variants into specific bins based on fitness. The precise value of each point, therefore, remains unknown and resides only within the defined bin boundaries. This phenomenon, known in statistics as interval censoring, introduces significant uncertainty into the dataset and makes the outcome highly dependent on the gating strategy used to define sorting bins. Further complicating the picture is the fact that each variant is represented by multiple cells and therefore generates multiple data points. Due to intrinsic biological variability and experimental noise, these cellular replicates of the variants frequently fall into different bins, resulting in a censored distribution of fitness values. A prior study focused on inferring protein fitness properties from the classification of variants in two bins [8], but did not address larger numbers of bins or how the number of collected samples influences measured fitness.

Additionally, from a purely practical perspective, it is important to consider not only censoring error but also its associated costs. Enhancing data resolution by increasing the number of gates (i.e., decrease the gate width) and collecting more measurements to achieve a more accurate fitness distribution and improve reproducibility inevitably increases experimental costs. In most cases, the number of gates will be limited by the choice of hardware (i.e., FACS instrument) and is generally limited to low numbers in a single sorting run (e.g., typically 4-channel sorting). Therefore, achieving an optimal trade-off between the various factors is crucial for efficient and cost-effective implementation of HTS in research.

The ultimate goal is to ensure that realistically-sampled censored data are truly representative of each variant-specific theoretically continuous distribution of fitness values. This can be achieved by minimizing the divergence between a specific measure derived from the censored data (i.e., the censored averages $\bar{z}_i$) and the ideal data (i.e., the continuous means $\mu_{y,i}$) for each variant in the library. Moreover, ML presents an additional layer of complexity. In fact, ML models are trained on censored data ($\bar{z}_i$), but their predictions should ideally represent the theoretical fitness values ($\mu_{y,i}$).

A critical aspect of ML in this setting is its ability to learn efficiently from training examples, which can be quantified using learning curves (LCs) [11, 29], a popular concept in ML applied to quantum physical chemistry [20]. LCs reflect how a model's predictive performance improves as more data points are added to the training set, and typically follow a monotonically decreasing trend [49]. Historically, LCs have been observed to follow a power-law decay [2], usually depicted on a log-log scale. However, recent work on ML of protein sequences [15] suggests that when the training set is scaled to include variants with increasing numbers of mutations and normalized using a combinatorial scale, the resulting LCs show discontinuities and different decay behaviors. These are referred to as "mutant-based normalized" LCs.

Although interval censoring has been studied extensively in mathematics [56], with solutions such as Sheppard's correction for statistical moments [23, 12], these solutions often assume a theoretical scenario involving a single probability distribution with infinite samples discretized by narrow gates. The HTS scenario, however, involves the censoring of multiple distributions, finite sampling, and broader gating strategies, which can result in all samples from a variant being enclosed within a single

bin. Furthermore, the impact of interval-censored HTS data on ML performance remains largely unexplored.

To address this knowledge gap, we present a comprehensive study on the effects of interval censoring in HTS experiments on data accuracy and ML performance. After introducing the HTS parameters and the parallels between our approach and a hypothetical experimental scenario, we present a theoretical exploration of interval censoring, followed by an in-depth analysis of simulated HTS data. We then assess the performance of ML models trained on censored data using mutant-based normalized LCs, highlighting challenges, and proposing solutions for reconstructing continuous fitness distributions from censored data. Our findings indicate that optimizing HTS parameters, such as gate widths and sample sizes, can significantly reduce errors and enhance the learning efficiency of ML models. We propose the introduction of a dimensionless number, R_D, to guide these optimizations, ensuring a balance between minimizing censoring errors and maintaining experimental feasibility. This study aims to offer practical guidelines for researchers, ultimately improving the accuracy and efficiency of HTS methodologies.

2 Results

2.1 Parameter Set and System Overview

HTS experiments based on FACS/NGS are influenced by several parameters (Fig. 1A, Tab. S2). In this study, we focused on the effect of (i) gate width (h); (ii) gate positioning (z_0 or h_0); (iii) the average number of samples acquired ($\bar{N}_s$); and (iv) the distribution of mutagenized variants in the screened library (p_{WT}). Additionally, we kept the following parameters constant: the wild-type (WT) sequence and its length (n), the mutation vocabulary (v), the functional form of the continuous fitness landscape ($f(\cdot)$), along with the corresponding variability in fitness measurements (σ_y). Even though such parameters are also crucial factors that could influence HTS outcomes, we kept them constant to simplify and focus our study.

The type of HTS experiments we consider here typically begin with construction of a DNA library encoding the variants to be screened. This library is constructed starting from a WT sequence where mutations are introduced from a chosen vocabulary (v). The vocabulary refers to the set of amino acids eligible to be substituted at each position in the WT protein sequence. As a concrete example, we focus on a segment of the human fibrinogen beta chain (WT, *FFFSARG*, $n = 7$), and we allow mutations to be selected from the amino acids that are already present within the WT sequence ($v = 5$, amino acids: A, F, G, R and S). This peptide binds to a partner protein called SdrG from *S. epidermidis* [32]. The structure of the complex between the WT peptide and SdrG is known (PDB: 1r17), enabling us to use computationally predicted binding energy as a fitness measure for the mutagenized peptides. Binding energy was calculated via an energy force field function (EvoEF [31, 21], Supporting Information). In Fig. 1B, we present examples of mutagenized variants in the library, each shown with the corresponding number of mutations (m) and their associated position along the combinatorial normalized scale used in this work (N_{norm}, Supporting Information).

Another critical aspect of HTS outcomes is the frequency of each variant in the library. The number of mutations ($P(m)$) is assumed to follow a binomial distribution, where a mutation at each position is equally likely. Under this assumption, the probability of any given variant with a known number of mutations can be calculated ($P(x_i|m)$, Fig. 1C). However, this binomial distribution is just one possible model for mutation frequency. Depending on the library design strategy, other distributions may also be relevant. The binomial model is used here for simplicity, but alternative approaches could still yield similar analytical outcomes.

In an ideal experimental scenario, the DNA library is transformed into a host cell line and protein expression is induced using standard biochemical techniques. The variant library is then characterized and screened for a specific fitness of interest using FACS-based HTS. In Fig. 1D, we illustrate the hypothetical scenario for this study, where a library of mutagenized peptides is screened for binding activity against a target protein. The fitness parameter of interest is the binding affinity between the peptide variant (Fig. 1D, a) and the target protein. Each variant in the combinatorial library is displayed on the surface of a host cell (black dots), typically *E. coli* or *S. cerevisiae*, in a multivalent (i.e. many thousands of copies per cell) manner, with one variant type per cell and multiple cells per variant type. A fluorescently labeled (i.e., green) antibody (Fig. 1D, b) can bind to the displayed protein to estimate the quantity of molecular receptors (the displayed peptide) expressed on the cell

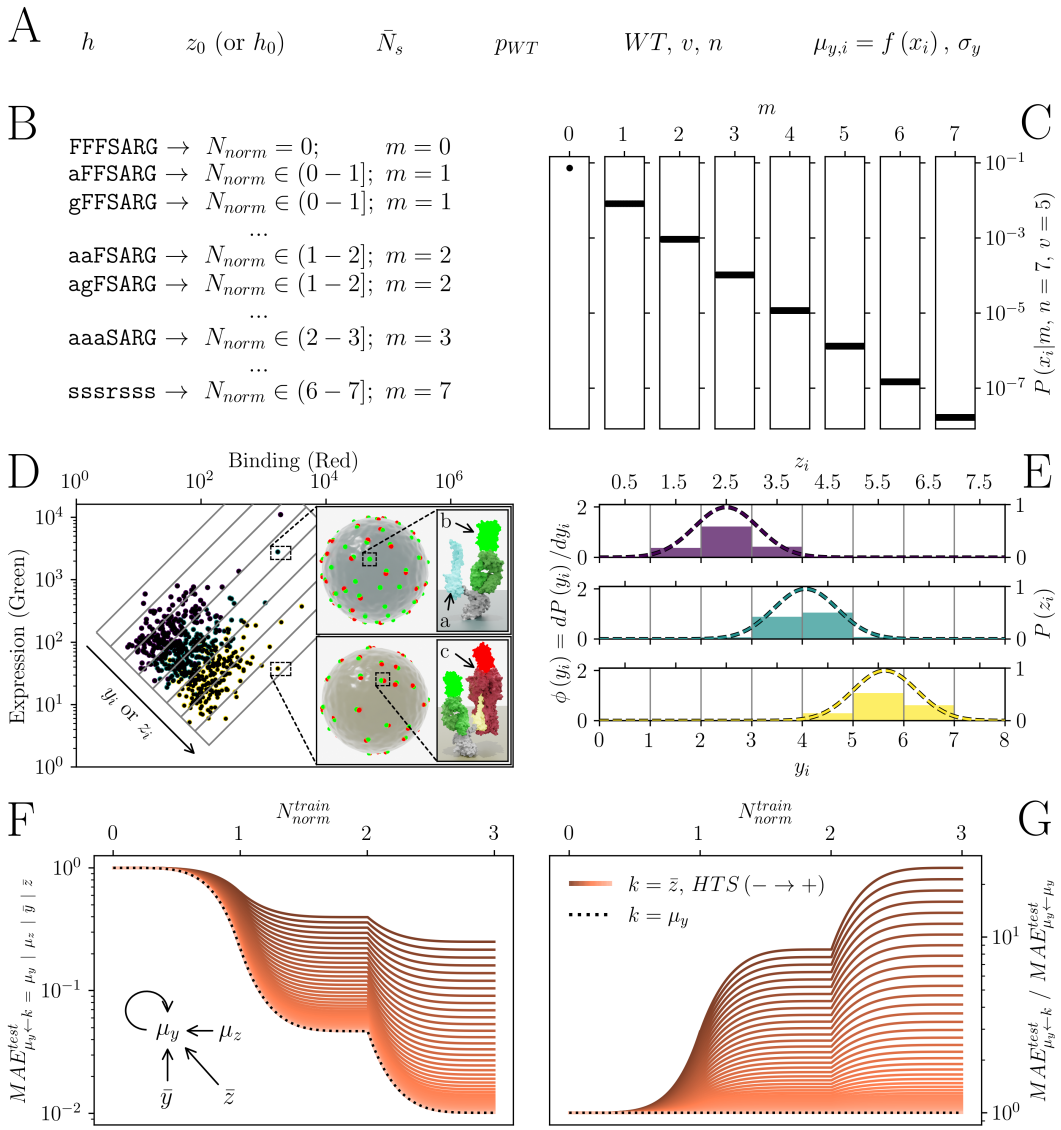

Figure 1: Overview of mutagenized library generation, FACS-based sorting and ML evaluation with interval-censored data. (A) Key parameters influencing interval censoring in FACS experiments. (B) Design of a combinatorial variant library derived from the parent peptide sequence FFSARG, with different mutation positions leading to variant distributions. (C) Example distribution of the combinatorial library, following a binomial distribution pattern based on the mutant frequency, where the probability of retaining the wild-type sequence (p_{WT}) was approximately 0.69. (D) Hypothetical FACS experiment designed to evaluate expression levels (green fluorescence) and binding affinity (red fluorescence) across variants. The first level zoom insets show two labeled cells expressing two variants (yellow and blue). The second level zoom insets show: (a) an unbound mutant peptide (blue), (b) labeled (green) expression antibody (dark green) bound to its tag (grey), and (c) labeled (red) target (dark red) bound to a mutant variant (yellow). (E) Example of population distributions of variants as determined by FACS. Dashed lines represent theoretical continuous distributions for each variant (true distributions), while the solid bars represent the binned/censored distributions (observed data). (F) Example comparison of mutant-based normalized LCs for models trained on interval-censored data ($\bar{z}_i$) versus models trained on true continuous data ($\mu_{y,i}$), with varying HTS parameters (color gradient). The inset depicts the possible approximation/prediction-training pairings. (G) Ratio of LCs of models trained on censored averages ($\bar{z}_i$) versus those trained on true means ($\mu_{y,i}$).

surface. Subsequently, a fluorescently labeled (i.e., red) protein target (Fig. 1D, c) would be screened against the variant peptide library using FACS. Measuring both the display level and binding level is crucial to correctly normalize the binding affinity to expression levels. In such an experiment, FACS sorting gates corresponding to peptides exhibiting equal affinity (y or z) with the target protein can be drawn obliquely (Fig. 1D). The cells contained within one of these slanted bins are then sorted by the FACS instrument, and processed by NGS to obtain the bin-wise interval-censored distribution of each variant (Fig. 1E).

An efficient data analysis algorithm would reconstruct the continuous distribution of each variant from its censored distribution and ultimately obtain an accurate estimate of true fitness values (i.e., the true means of the continuous distributions, $\mu_{y,i}$) from the binned data (the averages of the censored distributions, $\bar{z}_i$). Since we generated our datasets via simulations rather than experiments, we had access to two additional approximations, namely the censored means ($\mu_{z,i}$) and the continuous averages ($\bar{y}_i$). Later in this study, we used these values to evaluate the divergence of each approximation level (Tab. S3) from $\mu_{y,i}$ by calculating their mean absolute error (MAEs) using different binning and sorting strategies. Subsequently, we utilized LCs to evaluate ML performance on predicting $\mu_{y,i}$ from models trained with data coming from the different approximations ($\mu_{y,i}$, $\mu_{z,i}$, $\bar{y}_i$ and $\bar{z}_i$) for different HTS parameter sets. Such LCs were normalized using a combinatorial scaling based on the number of mutations (N_{norm}^{train}, Supporting Information). Moreover, we also report the ratio of LCs (Fig. 1G) by normalizing LCs to the ones corresponding to the same models trained on continuous means ($\mu_{y,i}$).

2.2 Theoretical distribution of a single variant ($N_s = \infty$)

We began by analyzing the impact of censoring on a single Gaussian distribution (Fig. 2A), focusing on the effects of the normalized gate width (h/σ_y) and gate position (z_0/h). This analysis involved measuring the difference between the means of the continuous and censored distributions ($\mu_y - \mu_z$, Fig. 2B) and the censored standard deviation (σ_z, Fig. 2C). Both quantities were normalized by the standard deviation of the continuous distribution (σ_y).

The normalized difference in means (Fig. 2B) exhibited an antisymmetric relationship with respect to the y-axis ($z_0/h = 0$). This difference converged to zero when the continuous distribution was located exactly in the middle of a gate ($z_0/h = \pm 0.5$), when the population was equally divided in two parts ($z_0/h = 0$), or when the gate width was below a certain threshold. The first observation was related to the fact that when the gate width was sufficiently large, the entire population could be contained within a single gate. Conversely, when the population was divided into a sufficient number of gates, the continuous and censored populations converged (Kullback–Leibler divergence approached zero, Fig. S1A), resulting in equal means.

The normalized censored standard deviation (Fig. 2C) exhibited a symmetric behavior, with a peak occurring at a normalized position of zero ($z_0/h = 0$). At high gate width, the magnitude of this peak increased with the growth of h/σ_y following the formula for a binary discrete distribution (Eq. 1).

$$\left.\frac{\sigma_z}{\sigma_y}\right|_{h/\sigma_y > h_{norm}^{crit};\ z_0=0} = \frac{h/\sigma_y}{2} \tag{1}$$

Otherwise, if the entire continuous population was contained within a single gate, the censored standard deviation collapsed to zero. This outcome did not imply that the general uncertainty was zero, but rather that the precise position within the gate could not be determined. Similarly, as observed for the difference in means, below a certain gate width threshold, the normalized censored standard deviation became independent of the gate position and converged. In this case, it did not converge to zero but to the Sheppard's correction of the second moment (Eq. 2).

$$\left.\frac{\sigma_z}{\sigma_y}\right|_{h/\sigma_y < h_{norm}^{crit}} = \sqrt{1 + \frac{1}{12} \cdot (h/\sigma_y)^2} \tag{2}$$

We could then define the normalized critical gate width (h_{norm}^{crit}) as the point at which the two previous expressions (Eq. 2 and 1) intersected (Eq. 3 and Fig. S2).

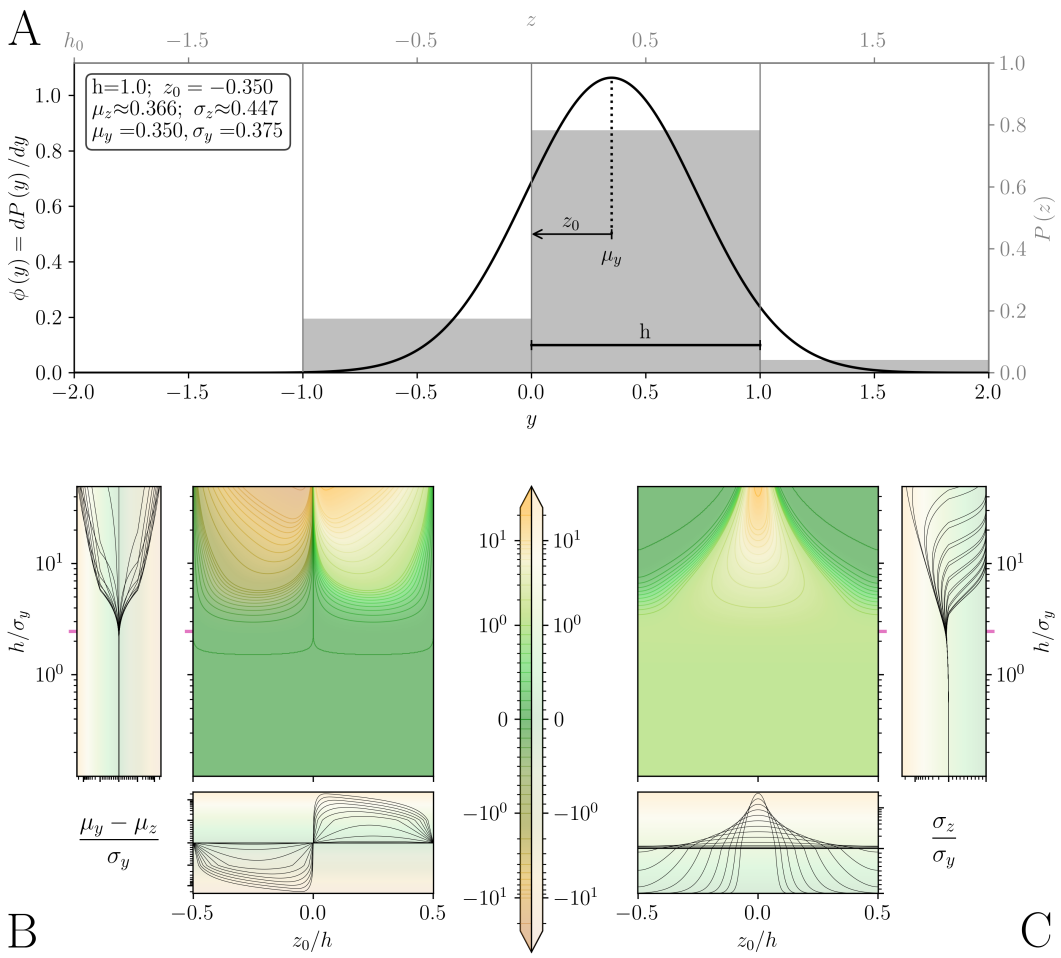

Figure 2: Impact of interval censoring on a single variant with ideally distributed fitness (gaussian), changing the normalized gate width (h/σ_y) and the normalized gate position (z_0/h). (A) Example of continuous gaussian distribution (black line) and censored distribution associated with it (grey bars). (B) Contour plots display the difference in normalized continuous mean (μ_y/σ_y) and censored mean (μ_z/σ_y). The side and bottom panels provide cross-sectional views along the x-axis (z_0/h) and y-axis (h/σ_y). (C) Similar to (B), but showing the ratio of censored to continuous standard deviations (σ_z/σ_y). The colormap is shared between panels (B) and (C).

$$h^{crit}_{norm} := h/\sigma_y \Big| \frac{h/\sigma_y}{2} = \sqrt{1 + \frac{1}{12} \cdot (h/\sigma_y)^2} \rightarrow h/\sigma_y = \pm\sqrt{6} \tag{3}$$

Similar observations could be made by examining the Shannon's entropy of the censored population (Fig. S1B). In this case, h^{crit}_{norm} could be interpreted as the point below which the information was uniformly distributed regardless of the gate position. Beyond this point, the entropy linearly increased with the exponential decay of h/σ_y.

2.3 Realistic distribution of a single variant ($N_s < \infty$)

We extended the previous analysis by introducing the effect of sampling. To estimate the standard error of both the continuous and censored distributions, we relied on the central limit theorem (CLT, $\sigma_{\bar{x}} = \sigma_x/\sqrt{N_{samples}}$). Our goal was to decouple the censoring divergence into two parts.

At first, we focused on the pure censoring error, which arises from the expected normalized absolute difference between the continuous and censored means (μ_y and μ_z). We refer to this quantity as "$Bias$" or "B." (Eq. 4, Fig. 3A).

$$Bias := \mathbb{E}_{z_0}\left[\frac{|\mu_y - \mu_z|}{\sigma_y}\right] \tag{4}$$

Where z_0 was marginalized by calculating its expected value. $Bias$ started from zero and grew exponentially with h/σ_y, and it represented the error that one could not overcome by simply sampling the distribution more heavily (e.g., in our HTS/FACS example, by sorting more cells).

Secondly, we measured the normalized absolute difference between the censored mean and average, referring to this as "$Var.$" or "$V.$" ($Variance$, Eq. 5, Fig. 3B).

$$Var. := \mathbb{E}_{z,z_0}\left[\frac{|\mu_z - \bar{z}|}{\sigma_y}\right] = \mathbb{E}_z\left[\frac{|\mathcal{N}\left(0,\, \mathbb{E}_{z_0}\left[\sigma_{\bar{z}}\right]\right)|}{\sigma_y}\right] = \sqrt{\frac{2}{\pi}} \cdot \frac{\mathbb{E}_{z_0}\left[\sigma_z\right]}{N_s} \cdot \frac{1}{\sigma_y} \tag{5}$$

Where $\sigma_{\bar{z}}$ was the standard error of the censored variable accordingly to the CLT ($\sigma_{\bar{z}} = \sigma_z/\sqrt{N_s}$) and N_s was the number of samples acquired.

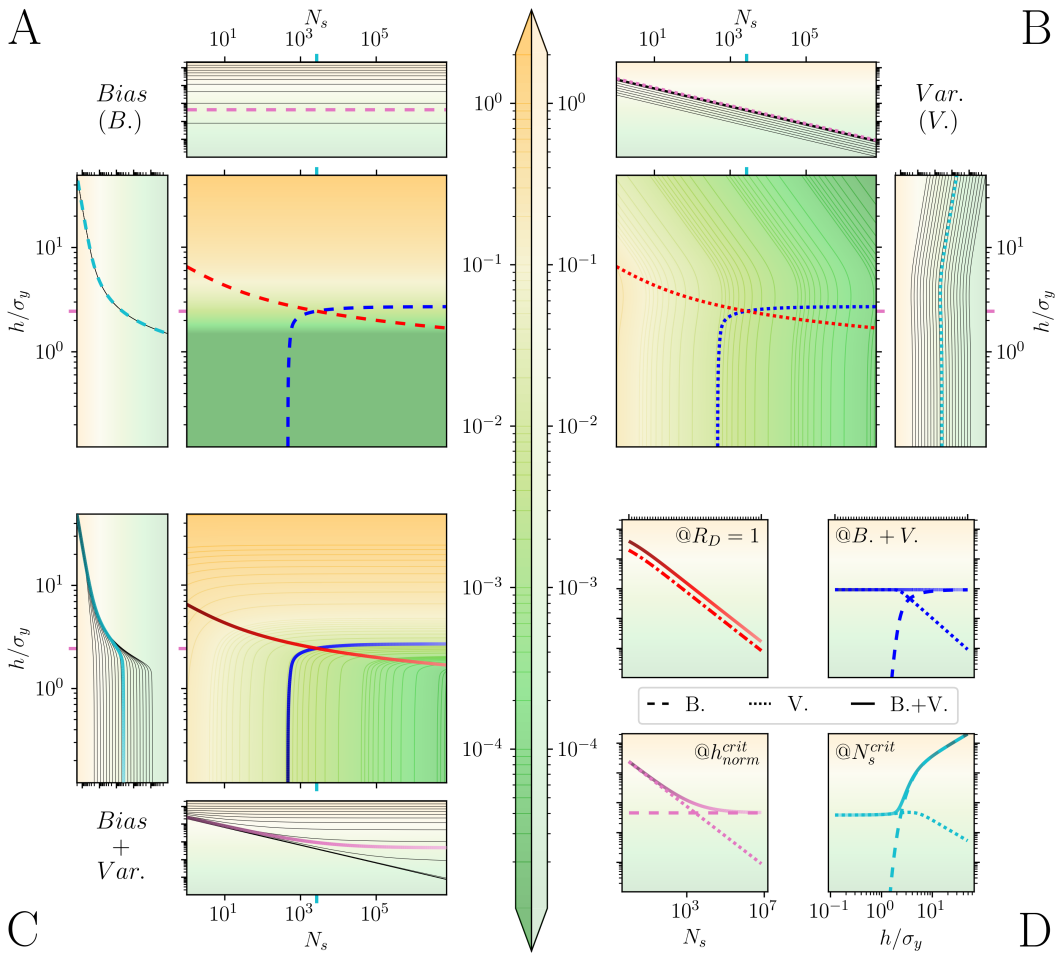

Figure 3: Impact of interval censoring on a single variant with realistically distributed fitness (sampled from gaussian), changing the normalized gate width (h/σ_y) and the number of samples (N_s). (A) Expected value of the normalized absolute difference between continuous and censored means ($Bias$, $B.$). (B) Expected value of the normalized absolute difference between censored mean and average ($Var.$, $V.$). (C) Sum of Bias and Variance ($Bias + Var.$). (D) Contour cross-sections at different critical parameters: $R_D = 1$ (red), constant $B. + V.$ (blue), h_{norm}^{crit} (magenta), and N_s^{crit} (cyan). The colormap is shared between all panels.

In contrast to $Bias$, $Var.$ was generally dependent on both h/σ_y and N_s. Nevertheless, at low gate widths ($h/\sigma_y < \sqrt{6}$), it could be approximated as dependent only on the number of samples. By our definition, $Var.$ could be seen as an estimation of the expected deviation from the biased censored mean, arising from the stochastic nature of the sampling process.

After defining $B.$ and $V.$, we summed them up (Fig. 3C) to estimate the maximum total divergence of the experimental censored average ($\bar{z}$) from the true continuous mean (μ_y). Note that, due to the absolute value function, the sum of these quantities corresponded to the absolute difference between the experimental and the true values ($|\mu_y - \bar{z}|$) only in the worst-case scenario, where both errors had the same sign. This sum clearly showed the existence of a trade-off between h/σ_y and N_s to reach a certain maximal error, illustrated in the form of contour lines (e.g., blue line, Fig. 3C).

Finally, inspired by the chemical concept of rate-limiting reaction step, we defined a dimensionless number to measure whether the censoring error was $Bias$ or $Variance$ driven (Eq. 6).

$$R_D := \left| \frac{\mu_y - \mu_z}{\mu_z - \bar{z}} \right| \approx \frac{Bias}{Var.} \tag{6}$$

In our case, we approximated R_D as the ratio of $Bias$ to $Variance$ by calculating its expected value with respect to z and z_0. By setting this number to be equal to one (i.e. the point at which $B. = V.$) we defined a solution to the h/σ_y vs. N_s trade-off. We subsequently used this solution to define a critical number of samples by calculating its intersection with the critical gate width (Eq. 7).

$$N_s^{crit} := N_s \big| R_D = 1 \cap h/\sigma_y = h_{norm}^{crit} \rightarrow N_s \approx 2659 \tag{7}$$

We note that, in comparison to h_{norm}^{crit}, N_s^{crit} did not rely on a transition between different phases but simply on how R_D is calculated. Furthermore, we did not account for changes in expected experimental cost associated with decreasing the gate widths relative to scaling up the number of samples, which could be problem/experiment-specific, making the proposed solution at $R_D = 1$ potentially cost-suboptimal. Nevertheless, to show the error-optimality of $R_D = 1$, we reported $Bias$, $Variance$ and the sum of the two at different conditions (3D, dashed, dotted and solid lines, respectively). At $R_D = 1$ (3D, @$R_D = 1$), a power-law-like decay showed the advantage of progressing in this direction. Following a contour line (3D, @$B. + V.$) confirmed that $B. + V.$ remained constant, revealing how $B.$ increased non-linearly while $V.$ was stable until it started to decay exponentially. Moving towards the h_{norm}^{crit} line (3D, @h_{norm}^{crit}) and the N_s^{crit} line (3D, @N_s^{crit}) showed how $B. + V.$ was constrained by one of the two quantities. This behavior is analogous to what could be observed in reactions when a shift in a rate-limiting regime occurs.

2.4 Realistic distribution of multiple variants

Until this point, our investigation focused on the distribution of fitness values of a single variant. However, our ultimate goal was to study a scenario involving a group (or pool) of distributions, specifically a pooled mutagenized library of variants displayed on cells undergoing the censoring process in a FACS experiment. Typically, the distribution of variants containing a certain number of point mutations in a combinatorial library (Fig. 1B) can be modelled using a binomial distribution. This concept could be easily extended (Eq. 8, Fig. 1C) to model the probability of a specific variant (x_i) by knowing its number of mutations (m), the vocabulary size (v) and the length of its construct (n).

$$P_{bin}(x_i|m,\, n,\, v,\, p_{WT}) = \frac{\binom{n}{m} \cdot {p_{WT}}^{n-m} \cdot (1 - p_{WT})^m}{\binom{n}{m} \cdot (v-1)^m} \tag{8}$$

Where p_{WT} represented the probability of drawing the WT amino acid at any position, and assuming that the vocabulary included at least all amino acids present in the WT sequence.

By employing Eq. 8 along with Shannon entropy (H) and the Jensen Inequality, one could optimize p_{WT} to maximize the information contained in the combinatorial library (Eq. 8).

$$p^*_{WT} := \underset{p_{WT}\in[0-1]}{\arg\max} \sum_i H\left(P_{bin}\left(x_i|m,\, n,\, v\right)\right) = 1/v \rightarrow P_{bin}\left(x_i\right) = P_{uni}\left(x_i\right) = 1/v^n \quad (9)$$

This approach corresponded to transitioning from the binomial distribution to a uniform distribution. Moreover, since it would be unfeasible to generate a library containing all possible mutations, one could use a similar approach to find a p_{WT} to maximize the information up to a specific number of mutations (M, Eq. 10).

$$p^{\#}_{WT}(M) := \underset{p_{WT}\in[0-1]}{\arg\max} \sum_i H\left(\frac{P_{bin}\left(x_i|m \in (0, M < n]\,,\, n,\, v\right)}{\sum_i P_{bin}\left(x_i|m \in (0, M < n]\,,\, n,\, v\right)}\right) \quad (10)$$

We then utilized these concepts to generate a uniformly distributed restricted library ($p_{WT} = 1/v = 0.2$, Supporting Information), containing all variant sequences with the number of mutations m $\leq$ 3, based on the fitness data described in our previous work [15]. We reported the mean absolute error between the continuous means ($\mu_{y,i} = f_{EvoEF}(x_i)$) and: (i) the censored means ($\mu_{z,i}$, Fig. S3A); (ii) the continuous averages ($\bar{y}_i$, Fig. S3B); and (iii) the censored averages ($\bar{z}_i$, Fig. S3C). We observed the same trends previously recorded in the cases of $Bias$ (Fig. 3A), $Variance$ (Fig. 3B) and $B. + V.$ (Fig. 3C), confirming the extendibility of the concepts initially presented for a single distribution to a combinatorial library of variants.

2.5 ML performance on reconstructing the continuous distribution from censored training data

To evaluate the performance of ML algorithms on censored data, we utilized kernel ridge regressions (KRRs) [34] to predict the continuous means of various test sets drawn from the uniformly distributed restricted library generated in the previous section. The task for the KRR models was to reconstruct such continuous mean values of variants absent from the training set, using training data that contained varying levels of approximation. The validation and test sets comprised only continuous mean values ($\mu_{y,i}$) from variants with more than three mutations. In contrast, the training set included only variants with a maximum of three mutations and variable labels ($\mu_{y,i}$, $\mu_{z,i}$, $\bar{y}_i$, $\bar{z}_i$).

We reported the median ratio of test MAE, calculated across different shuffles and averaged samples (Supporting Information). The ratios were calculated between models trained using censored means (Fig. S4A, $\mu_{z,i}$), continuous averages (Fig. S4B, $\bar{y}_i$), and censored averages (Fig. S4C, $\bar{z}_i$), against models trained on continuous means ($\mu_{y,i}$). These results consistently demonstrated the same trends observed in the single population study ($B.$, $V.$, and $B. + V.$, Fig. 3) and in the error calculation of library of variants (Fig. S3).

To further study the impact of the training dataset dimension, we plotted the mutant-based LCs at the critical values of $R_D = 1$ (Fig. 4A and S5A), at constant $B. + V.$ (Fig. 4B and S5B), at h^{crit}_{norm} (Fig. 4C and S5C) and the N^{crit}_s (Fig. 4D and S5D). Among such conditions, $R_D = 1$ was the only one able to achieve the same level of error as models trained with true continuous means across all possible training set sizes. This demonstrated the optimality of scaling experimental complexity according to the sets of gate widths and number of samples leading to $R_D = 1$. As expected, working at high gate widths significantly impacted the stability of learning (Fig. 4D and S5D), sometimes leading to no learning (i.e., flat LCs) or even deleterious learning, where increasing training data decreased the model performance. The data acquired from a contour line with constant $B. + V.$ (Fig. 4B and S5B) did not demonstrate any clear advantages of high $Var.$ conditions over high $Bias$ conditions. Finally, scaling up the number of samples collected per variant (Fig. 4C and S5C) consistently improved mean absolute error (MAE) at each training set size. Interestingly, at low $\bar{N}_s$, the learning curves began to lose their characteristic discontinuities, resembling a normal asymptotic power law decay.

In the case of a binomial distribution of variants (Fig. S6 and S7, $p_{WT} \approx 0.69$), similar trends were observed. However, due to the uneven distribution of samples, the uncertainty of each variant increased with the number of mutations and, consequently, with the training size. As a result, the divergence between the LCs from models trained with censored averages and those obtained with continuous means was initially lower than in the uniformly distributed case (Fig. 4 and S5). In some instances, this divergence increased with the training size until a point was reached where adding more data decreased the accuracy of the models (deleterious learning).

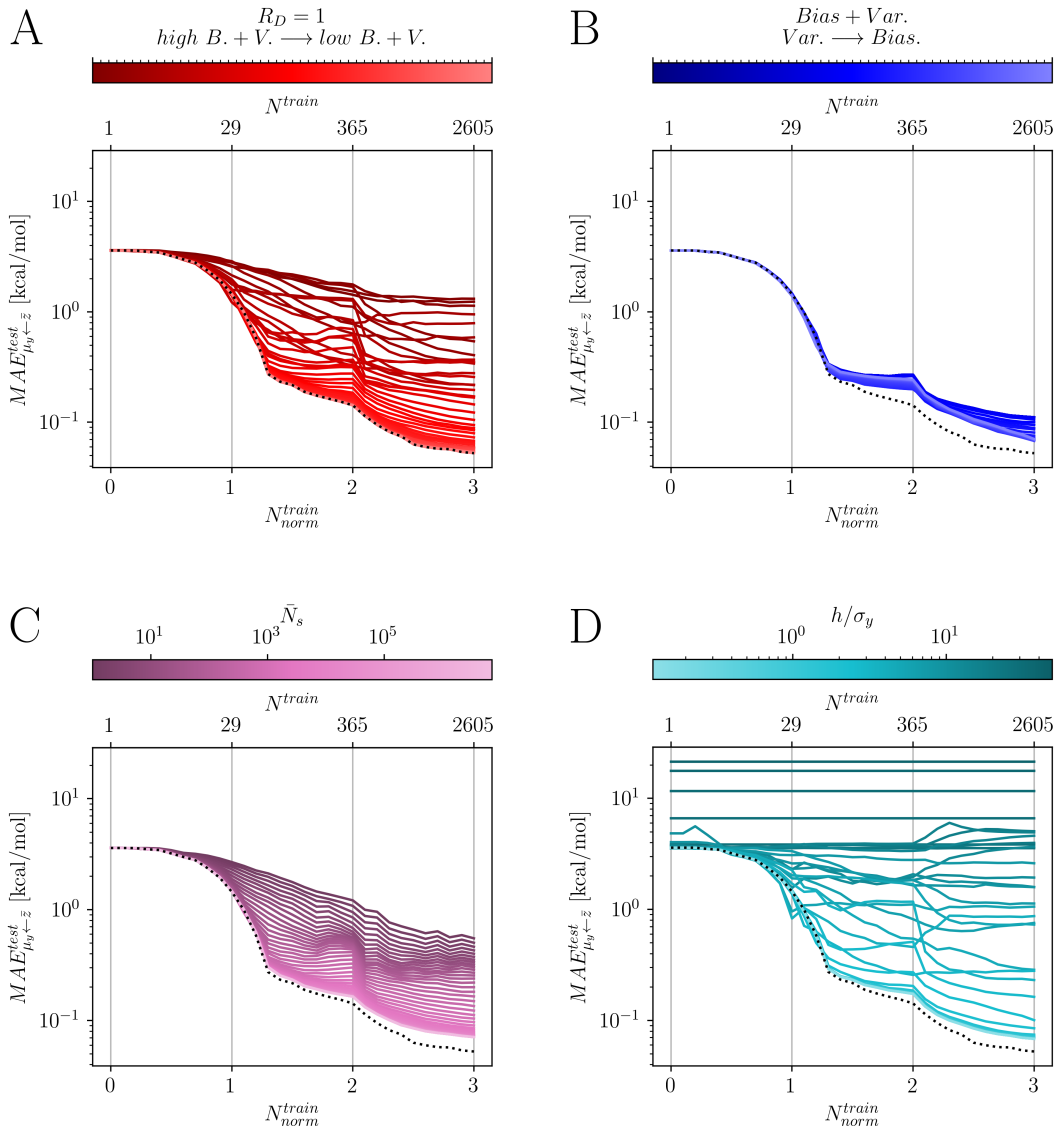

Figure 4: Impact of interval censoring on LCs of models trained with up to triple mutants ($N^{train}_{norm} \in [0-3]$) as a function of the mutant-based normalized number of training points (N^{train}_{norm}), and using a uniform distributed combinatorial library ($P_{WT} = 0.2$). The plots illustrate the test performance under different conditions: (A) $R_D = 1$ with a gradient from high $B + V$ to low $B + V$ (red), (B) constant $B + V$ with varying B-V contributions (blue), (C) constant normalized gate width ($h/\sigma_y = h^{crit}_{norm}$) with varying average sample size ($\bar{N}_s$, magenta), (D) constant average sample size ($\bar{N}_s = \bar{N}^{crit}_s$) with varying normalized gate width (h/σ_y, cyan).

In general, the analysis of these LCs demonstrated that, under certain conditions, it would be more beneficial to focus on improving the acquisition parameters, such as gate width and the number of samples (to jump vertically to a different LC), rather than increasing the library size and consequently the training set size (to move right on a specific LC). This observation can significantly inform experimental design since increasing library size (and therefore screening time) can be hugely costly. The optimal gating strategy can therefore be used to maximize useful information at a given level of library complexity and screening cost.

3 Discussion

In this work, we investigated the impact of interval data censoring in HTS experiments where data are sorted into different bins, as is commonly the case in protein engineering campaigns relying on FACS-based single-cell sorting, next-generation DNA sequencing and machine learning. We examined increasingly complex scenarios including (i) an ideal distribution (sorting an infinite number of cells of a single variant); (ii) a sampled distribution (sorting a finite number of cells of a single variant); and (iii) multiple sampled distributions (sorting finite cells of a combinatorial variant library). In the last case, we extended our analysis by evaluating the ability of ML models to accurately reconstruct the continuous means (true fitness values) when trained on censored averages (HTS data). This approach simulated the scenario of a HTS campaign where ML would be trained on experimentally observable variants and used to predict fitness values of unseen variants not contained in the original library. As expected, it was generally possible to accurately estimate true fitness values from our synthetic HTS datasets, but at the cost of increasing the complexity of the experiments (increase number of bins / decrease bin width and increase the number of samples per variant).

We found that narrowing the sorting gate width and increasing the number of samples (i.e., the number of cells sorted) generally reduced the reconstruction error. However, modifying only one of these parameters beyond a certain threshold leads to an asymptotic behavior in the reconstruction error reduction. This asymptotic behavior was limited by either the censoring error ($Bias$) or by the sampling error ($Variance$). We found that the preferred conditions for HTS experiments would be where neither error is dominant, and we developed a dimensionless number (R_D) to estimate the distance of a given set of conditions from the optimal ones ($R_D = 1$). Thus, R_D can be utilized by experimental scientists to optimize HTS parameters, thereby minimizing the intrinsic errors arising from sampling and binning and maximizing the utility of costly experimental data.

We also quantified the critical normalized gate width ($h/\sigma_y = \sqrt{6}$) below which the differences between the censored populations (binned data) and the continuous populations began to converge to Sheppard's correction, becoming independent of gate position. This insight can be utilized by researchers to plan their experiments, particularly in complex scenarios where R_D is difficult to determine or where the cost trade-off of changing the gate widths compared to the number of samples is not easily estimated. Specifically, researchers can first estimate the experimental standard deviation (σ_y) by focusing on a single variant (e.g., WT), and then calculate the necessary gate width that should not be exceeded ($h < \sigma_y\sqrt{6}$).

Our analysis has certain limitations associated with it. In real-world scenarios, the number of gates required to screen an entire variant library using optimal gate widths may exceed current technological or instrument capabilities. Additionally, our assumption of a continuously normal distribution could be invalid for some systems due to various factors, such as cell non-homogeneity or skewed fitness distributions. To address the first issue, we propose to divide the experiment into multiple sorting runs with different gate offsets. For each sorting run, the number of gates should be set to the maximum allowed by the hardware, typically between 4 and 12. Gate widths should be set to ensure they cover the entire library in each screen. Subsequently, all gates should be sequentially shifted by a specific value for each sorting run. When the data from these sorting runs are combined, this approach will simulate a single sorting run with the necessary gate width. For the second issue (instrument laser/temperature fluctuations), using the median as a measure of the population fitness instead of the mean is a valid strategy. Specifically, the median should be calculated by assuming that the probability of a specific bin is continuously distributed within the gate boundaries, rather than being entirely located in the middle as in a purely discrete distribution. Note that if the assumption of normality holds, the median calculated as suggested reverts to the mean. However, in the presence of outliers that skew the distribution of some variants, the median would be by definition more robust.

We showed that the trade-off between HTS parameters, training set size and variant library distribution is a complex process that can potentially be both hardware- and problem-specific. In some situations, increasing the library size may be advantageous, while in others, optimizing the HTS parameters could be more beneficial. Rather than providing a one-size-fits-all solution, we aimed to equip researchers working with discretized HTS data with additional theoretical and practical tools. These tools are intended to help scientists better understand how binning their fitness values through FACS interval censoring could potentially impact experimental results, and thereby affect ML performance of models trained on such data.

Acknowledgments and Disclosure of Funding

This work was supported by the University of Basel, the Swiss Federal Institute of Technology in Zürich (ETH Zürich) and the Swiss Nanoscience Institute (SNI, project P1802). The authors express their gratitude to Paolo Minelli for his insightful comments. The authors declare that some text was edited using ChatGPT (https://chat.openai.com). The editing focused solely on rearranging, enhancing, or correcting the syntax of existing sentences without creating new content. Portions of this work were performed at sciCORE scientific computing center at the University of Basel (https://scicore.unibas.ch).

References

[1] Jeremy J. Agresti, Eugene Antipov, Adam R. Abate, Keunho Ahn, Amy C. Rowat, Jean-Christophe Baret, Manuel Marquez, Alexander M. Klibanov, Andrew D. Griffiths, and David A. Weitz. Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences*, 107(9):4004–4009, March 2010. doi: 10.1073/pnas.0910781107. URL https://doi.org/10.1073/pnas.0910781107.

[2] Shun-ichi Amari, Naotake Fujita, and Shigeru Shinomoto. Four types of learning curves. *Neural Comput*, 4(4):605–618, July 1992. ISSN 0899-7667. URL https://doi.org/10.1162/neco.1992.4.4.605.

[3] Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione, and David Baker. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-021-04184-w.

[4] Frances H Arnold. Design by directed evolution. *Accounts of chemical research*, 31(3):125–131, 1998.

[5] Elizabeth L. Bell, Ross Smithson, Siobhan Kilbride, Jake Foster, Florence J. Hardy, Saranarayanan Ramachandran, Aleksander A. Tedstone, Sarah J. Haigh, Arthur A. Garforth, Philip J. R. Day, Colin Levy, Michael P. Shaver, and Anthony P. Green. Directed evolution of an efficient and thermostable pet depolymerase. *Nature Catalysis*, 5(8):673–681, 2022. ISSN 2520-1158. URL https://doi.org/10.1038/s41929-022-00821-3.

[6] Vincent Blay, Bhairavi Tolani, Sunita P. Ho, and Michelle R. Arkin. High-throughput screening: today's biochemical and cell-based approaches. *Drug Discovery Today*, 25(10):1807–1821, 2020. ISSN 1359-6446. URL https://www.sciencedirect.com/science/article/pii/S1359644620303056.

[7] Eric T. Boder, Katarina S. Midelfort, and K. Dane Wittrup. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proceedings of the National Academy of Sciences*, 97(20):10701–10705, September 2000. doi: 10.1073/pnas.170297297. URL https://doi.org/10.1073/pnas.170297297.

[8] Marshall Case, Matthew Smith, Jordan Vinh, and Greg Thurber. Machine learning to predict continuous protein properties from binary cell sorting data and map unseen sequence space. *Proceedings of the National Academy of Sciences*, 121(11):e2311726121, March 2024. doi: 10.1073/pnas.2311726121. URL https://doi.org/10.1073/pnas.2311726121.

[9] Kai Chen and Frances H. Arnold. Engineering new catalytic activities in enzymes. *Nature Catalysis*, 3(3):203–213, 2020. ISSN 2520-1158. URL https://doi.org/10.1038/s41929-019-0385-5.

[10] Lin Chen, Zehong Zhang, Zhenghao Li, Rui Li, Ruifeng Huo, Lifan Chen, Dingyan Wang, Xiaomin Luo, Kaixian Chen, Cangsong Liao, and Mingyue Zheng. Learning protein fitness landscapes with deep mutational scanning data from multiple sources. *Cell Systems*, 14(8): 706–721.e5, 2023. ISSN 2405-4712. URL https://www.sciencedirect.com/science/article/pii/S2405471223002107.

[11] Corinna Cortes, L. D. Jackel, Sara Solla, Vladimir Vapnik, and John Denker. Learning curves: Asymptotic values and rate of convergence. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1994. URL https://proceedings.neurips.cc/paper/1993/file/1aa48fc4880bb0c9b8a3bf979d3b917e-Paper.pdf.

[12] Cecil C. Craig. Sheppard's corrections for a discrete variable. *The Annals of Mathematical Statistics*, 7(2):55–61, 1936. ISSN 00034851. URL http://www.jstor.org/stable/2957618.

[13] P. J. Crosland-Taylor. A device for counting small particles suspended in a fluid through a tube. *Nature*, 171(4340):37–38, 1953. ISSN 1476-4687. URL https://doi.org/10.1038/171037b0.

[14] Andrew Currin, Neil Swainston, Philip J. Day, and Douglas B. Kell. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.*, 44(5):1172–1239, 2015. ISSN 0306-0012. URL http://dx.doi.org/10.1039/C4CS00351A.

[15] Vanni Doffini, O. Anatole von Lilienfeld, and Michael A. Nash. Data-error scaling in machine learning on natural discrete combinatorial mutation-prone sets: Case studies on peptides and small molecules, May 2024. URL https://arxiv.org/abs/2405.05167.

[16] Girum Fitihamlak Ejigu and Jaehee Jung. Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology*, 9, Sep 2020.

[17] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016. ISSN 1471-0064. URL https://doi.org/10.1038/nrg.2016.49.

[18] Kyrin R. Hanning, Mason Minot, Annmaree K. Warrender, William Kelton, and Sai T. Reddy. Deep mutational scanning for therapeutic antibody engineering. *Trends in Pharmacological Sciences*, 43(2):123–135, February 2022. ISSN 0165-6147. doi: 10.1016/j.tips.2021.11.010. URL https://doi.org/10.1016/j.tips.2021.11.010.

[19] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021. ISSN 0198-8859. URL https://www.sciencedirect.com/science/article/pii/S0198885921000628.

[20] Bing Huang, Guido Falk von Rudorff, and O. Anatole von Lilienfeld. The central role of density functional theory in the ai age. *Science*, 381(6654):170–175, July 2023. doi: 10.1126/science.abn3445. URL https://doi.org/10.1126/science.abn3445.

[21] Xiaoqiang Huang, Robin Pearce, and Yang Zhang. Evoef2: accurate and fast energy function for computational protein design. *Bioinformatics (Oxford, England)*, 36(31588495):1135–1142, February 2020. ISSN 1367-4803. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7144094/.

[22] S. A. Hunter, J. R. Cochran, and Vincent L. Pecoraro. Chapter two - cell-binding assays for determining the affinity of protein-protein interactions: Technologies and considerations. In *Methods in Enzymology*, volume 580, pages 21–44. Academic Press, January 2016. URL http://www.sciencedirect.com/science/article/pii/S0076687916300374.

[23] M. G. Kendall. The conditions under which sheppard's corrections are valid. *Journal of the Royal Statistical Society*, 101(3):592–605, 1938. ISSN 09528385. URL http://www.jstor.org/stable/2980630.

[24] Justin R. Klesmith, John-Paul Bacik, Emily E. Wrenbeck, Ryszard Michalczyk, and Timothy A. Whitehead. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences*, 114(9):2265–2270, February 2017. doi: 10.1073/pnas.1614437114. URL https://doi.org/10.1073/pnas.1614437114.

[25] Yi Li, Ruth Moysey, Peter E. Molloy, Anne-Lise Vuidepot, Tara Mahon, Emma Baston, Steven Dunn, Nathaniel Liddy, Jansen Jacob, Bent K. Jakobsen, and Jonathan M. Boulter. Directed evolution of human t-cell receptors with picomolar affinities by phage display. *Nature Biotechnology*, 23(3):349–354, 2005. ISSN 1546-1696. URL https://doi.org/10.1038/nbt1070.

[26] Dina Listov, Casper A. Goverde, Bruno E. Correia, and Sarel Jacob Fleishman. Opportunities and challenges in design and optimization of protein function. *Nature Reviews Molecular Cell Biology*, 25(8):639–653, 2024. ISSN 1471-0080. URL https://doi.org/10.1038/s41580-024-00718-y.

[27] Ulrich Markel, Khalil D. Essani, Volkan Besirlioglu, Johannes Schiffels, Wolfgang R. Streit, and Ulrich Schwaneberg. Advances in ultrahigh-throughput screening for directed enzyme evolution. *Chem. Soc. Rev.*, 49(1):233–262, 2020. ISSN 0306-0012. URL http://dx.doi.org/10.1039/C8CS00981C.

[28] Derek M. Mason, Simon Friedensohn, Cédric R. Weber, Christian Jordi, Bastian Wagner, Simon M. Meng, Roy A. Ehling, Lucia Bonati, Jan Dahinden, Pablo Gainza, Bruno E. Correia, and Sai T. Reddy. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature biomedical engineering*, 5:600–612, Jun 2021.

[29] K.-R. Müller, M. Finke, N. Murata, K. Schulten, and S. Amari. A numerical study on learning curves in stochastic multilayer feedforward networks. *Neural Comput*, 8(5):1085–1106, July 1996. ISSN 0899-7667. URL https://doi.org/10.1162/neco.1996.8.5.1085.

[30] Evangelos-Marios Nikolados, Arin Wongprommoon, Oisin Mac Aodha, Guillaume Cambray, and Diego A. Oyarzún. Accuracy and data efficiency in deep learning models of protein expression. *Nature Communications*, 13(1):7755, 2022. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-022-34902-5.

[31] Robin Pearce, Xiaoqiang Huang, Dani Setiawan, and Yang Zhang. Evodesign: Designing protein-protein binding interactions using evolutionary interface profiles in conjunction with an optimized physical energy function. *Journal of molecular biology*, 431(30851277):2467–2476, June 2019. ISSN 0022-2836. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6589126/.

[32] Karthe Ponnuraj, M. Gabriela Bowden, Stacey Davis, S. Gurusiddappa, Dwight Moore, Damon Choe, Yi Xu, Magnus Hook, and Sthanam V. L. Narayana. A "dock, lock, and latch" structural model for a staphylococcal adhesin binding to fibrinogen. *Cell*, 115(2):217–228, 2003. ISSN 0092-8674. URL https://www.sciencedirect.com/science/article/pii/S0092867403008092.

[33] Yuchi Qiu, Jian Hu, and Guo-Wei Wei. Cluster learning-assisted directed evolution. *Nature Computational Science*, 1(12):809–818, 2021. ISSN 2662-8457. URL https://doi.org/10.1038/s43588-021-00168-y.

[34] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[35] J. P. Robinson, Raluca Ostafe, Sharath N. Iyengar, Bartek Rajwa, and Rainer Fischer. Flow cytometry: The next revolution, 2023. ISSN 2073-4409.

[36] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature reviews. Molecular cell biology*, 10:866–76, Dec 2009.

[37] A. Romero Philip, Krause Andreas, and H. Arnold Frances. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110 (3):E193–E201, January 2013. doi: 10.1073/pnas.1215251110. URL https://doi.org/10.1073/pnas.1215251110.

[38] Gelman Sam, A. Fahlberg Sarah, Heinzelman Pete, A. Romero Philip, and Gitter Anthony. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 118(48):e2104878118, November 2021. doi: 10.1073/pnas.2104878118. URL https://doi.org/10.1073/pnas.2104878118.

[39] Mateo I. Sanchez and Alice Y. Ting. Directed evolution improves the catalytic efficiency of tev protease. *Nature Methods*, 17(2):167–174, 2020. ISSN 1548-7105. URL https://doi.org/10.1038/s41592-019-0665-7.

[40] Heena Satam, Kandarp Joshi, Upasana Mangrolia, Sanober Waghoo, Gulnaz Zaidi, Shravani Rawool, Ritesh P. Thakare, Shahid Banday, Alok K. Mishra, Gautam Das, and Sunil K. Malonia. Next-generation sequencing technology: Current trends and advancements. *Biology*, 12(7), 2023. ISSN 2079-7737. doi: 10.3390/biology12070997. URL https://www.mdpi.com/2079-7737/12/7/997.

[41] Niklas E. Siedhoff, Ulrich Schwaneberg, Mehdi D. Davari, and Dan S. Tawfik. Chapter twelve - machine learning-assisted enzyme engineering. In *Methods in Enzymology*, volume 643, pages 281–315. Academic Press, 2020. URL https://www.sciencedirect.com/science/article/pii/S0076687920302275.

[42] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5): 1295–1310.e20, September 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.08.012. URL https://doi.org/10.1016/j.cell.2020.08.012.

[43] Steven A. Sundberg. High-throughput and ultra-high-throughput screening: solution- and cell-based approaches. *Current Opinion in Biotechnology*, 11(1):47–53, 2000. ISSN 0958-1669. URL https://www.sciencedirect.com/science/article/pii/S0958166999000518.

[44] Joseph M. Taft, Cédric R. Weber, Beichen Gao, Roy A. Ehling, Jiami Han, Lester Frei, Sean W. Metcalfe, Max D. Overath, Alexander Yermanos, William Kelton, and Sai T. Reddy. Deep mutational learning predicts ace2 binding and antibody escape to combinatorial mutations in the sars-cov-2 receptor-binding domain. *Cell*, 185(21):4008–4022.e14, 2022. ISSN 0092-8674. URL https://www.sciencedirect.com/science/article/pii/S0092867422011199.

[45] Neil Thomas and Lucy J. Colwell. Minding the gaps: The importance of navigating holes in protein fitness landscapes. *Cell Systems*, 12(11):1019–1020, 2021. ISSN 2405-4712. URL https://www.sciencedirect.com/science/article/pii/S2405471221004154.

[46] Derek Vallejo, Ali Nikoomanzar, Brian M. Paegel, and John C. Chaput. Fluorescence-activated droplet sorting for single-cell directed evolution. *ACS Synth. Biol.*, 8(6):1430–1440, June 2019. doi: 10.1021/acssynbio.9b00103. URL https://doi.org/10.1021/acssynbio.9b00103.

[47] Rosario Vanella, Gordana Kovacevic, Vanni Doffini, Jaime Fernández de Santaella, and Michael A. Nash. High-throughput screening, next generation sequencing and machine learning: advanced methods in enzyme engineering. *Chem. Commun.*, 58(15):2455–2467, January 2022. ISSN 1359-7345. URL http://dx.doi.org/10.1039/D1CC04635G.

[48] Rosario Vanella, Christoph Küng, Alexandre A. Schoepfer, Vanni Doffini, Jin Ren, and Michael A. Nash. Understanding activity-stability tradeoffs in biocatalysts by enzyme proximity sequencing. *Nature Communications*, 15(1):1807, 2024. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-024-45630-3.

[49] T. Viering and M. Loog. The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022. ISSN 1939-3539.

[50] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, July 2022. doi: 10.1126/science.abn2100. URL https://doi.org/10.1126/science.abn2100.

[51] Yajie Wang, Pu Xue, Mingfeng Cao, Tianhao Yu, Stephan T. Lane, and Huimin Zhao. Directed evolution: Methodologies and applications. *Chem. Rev.*, 121(20):12384–12444, October 2021. ISSN 0009-2665. doi: 10.1021/acs.chemrev.1c00260. URL https://doi.org/10.1021/acs.chemrev.1c00260.

[52] Bruce J. Wittmann, Yisong Yue, and Frances H. Arnold. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems*, 12(11):1026–1045.e7, 2021. ISSN 2405-4712. URL https://www.sciencedirect.com/science/article/pii/S2405471221002866.

[53] Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, August 2019. ISSN 1548-7105. URL https://doi.org/10.1038/s41592-019-0496-6. (read).

[54] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J. Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z. Zhang, Ivan Anishchenko, Brian Coventry, Longxing Cao, Justas Dauparas, Samer Halabiya, Michelle DeWitt, Lauren Carter, K. N. Houk, and David Baker. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023. ISSN 1476-4687. URL https://doi.org/10.1038/s41586-023-05696-3.

[55] Weizhu Zeng, Likun Guo, Sha Xu, Jian Chen, and Jingwen Zhou. High-throughput screening technology in industrial biotechnology. *Trends in Biotechnology*, 38(8):888–906, 2020. ISSN 0167-7799. URL https://www.sciencedirect.com/science/article/pii/S0167779920300019.

[56] Zhigang Zhang and Jianguo Sun. Interval censoring. *Stat Methods Med Res*, 19(1):53–70, August 2009. ISSN 0962-2802. doi: 10.1177/0962280209105023. URL https://doi.org/10.1177/0962280209105023.