Denoising DNA Encoded Library Screens with Sparse Learning

Péter Kómár*,† and Marko Kalinić*,‡

†*Totient, Inc., 1 Alewife Center, Cambridge MA, 02140 USA*
‡*Totient, Inc., Sinđelićeva 9, 11000 Belgrade, Serbia*

E-mail: peter.komar@totient.bio; marko.kalinic@totient.bio

Abstract

DNA-encoded libraries (DELs) are large, pooled collections of compounds in which every library member is attached to a stretch of DNA encoding its complete synthetic history. DEL-based hit discovery involves affinity selection of the library against a protein of interest, whereby compounds retained by the target are subsequently identified by next-generation sequencing of the corresponding DNA tags. When analyzing the resulting data, one typically assumes that sequencing output (i.e. read counts) is proportional to the binding affinity of a given compound, thus enabling hit prioritization and elucidation of any underlying structure-activity relationships (SAR). This assumption, though, tends to be severely confounded by a number of factors, including variable reaction yields, presence of incomplete products masquerading as their intended counterparts, and sequencing noise. In practice, these confounders are often ignored, potentially contributing to low hit validation rates, and universally leading to loss of valuable information. To address this issue, we have developed a method for comprehensively denoising DEL selection outputs. Our method, dubbed "deldenoiser", is based on sparse learning and leverages inputs that are commonly available within a DEL generation and screening workflow. Using simulated and publicly available DEL affinity selection data, we show that "deldenoiser" is not only able to recover and rank true binders much more robustly than read count-based approaches, but also that it yields scores which accurately capture the underlying SAR. The proposed method can, thus, be of significant utility in hit prioritization following DEL screens.

Keywords

DNA encoded library, affinity selections, denoising, sparse inference, machine learning

Introduction

Since the seminal paper by Clark et al.[1] that saw the concept of DNA encoded libraries (DELs) reduced to practice, the technology has been gaining popularity as a novel hit discovery[2–4] and, more recently, target prioritization tool.[5] DELs represent large combinatorial libraries of small molecules that are typically generated using a split-and-pool methodology, with 3 to 4 cycles of chemistry providing routine access to millions or even billions of unique compounds. Unlike traditional chemical libraries, though, each compound in a DEL is attached to a sequence of DNA – the tag or barcode – that stores information on its complete synthetic history, and all DEL members are kept in a mixture. Once a DEL is prepared, affinity selection experiments utilizing immobilized protein targets can be used to capture high-affinity DEL binders from these mixtures, whose chem-

ical identity can subsequently be determined by amplification and sequencing of the DNA tags.[6–8] We illustrate this process in Fig. 1. The scale and efficiency at which resolution of retained compound identities' can be made has been greatly improved with the introduction of next-generation sequencing[9] and associated statistical analysis[10] into DEL affinity selection workflows.

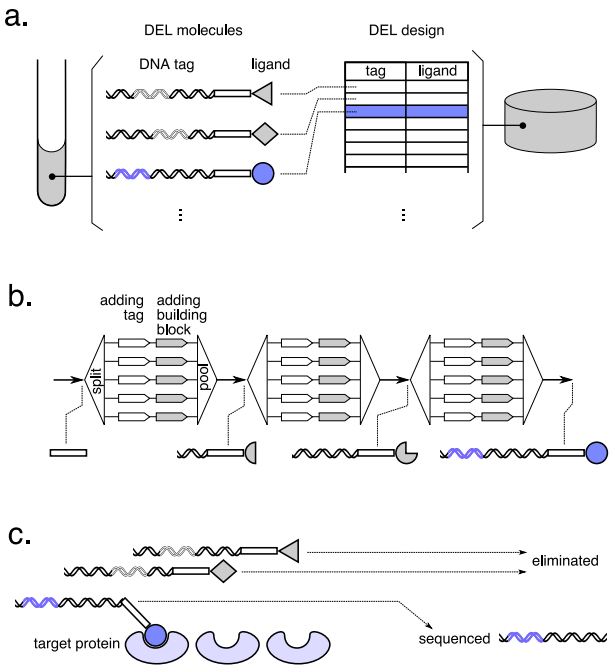

Figure 1: **a.** DNA encoded library consists of a mixture of molecules and a record of its design that provides one-to-one mapping between chemical identity and DNA tag sequence. **b.** Synthesis is performed in consecutive split-and-pool cycles, in each of which tags get extended and building blocks are attached to the ligand. **c.** The target protein binds compounds with high affinity, which are subsequently sequenced to reveal their identity.

In "DNA recorded" DEL preparation,[2,11] which is the most commonly used approach, the tag sequence is incrementally built up by ligating short oligonucleotides to the nascent tag in each cycle of chemistry. A unique oligonucleotide sequence thus identifies each building block that went into the library. While, in theory, this DNA encoding scheme should enable one to unambiguously resolve the chemical identity of any DEL member by simply sequencing the corresponding tag, in practice, the correspondence between a given sequence and the chemical composition of a compound is not one-to-one.[12] Specifically, no chemical reaction leads to complete conversion of reactants to the desired product, instead yielding a mixture of the starting reactants, side products, and the expected product. Since tags are extended irrespective of the true chemical composition of compounds they are attached to, it follows that one tag sequence will be associated with more than one specific compound; the converse, products of failed ligation, are removed with HPLC purification.[13] While this issue can be mitigated by optimizing reaction conditions, and profiling building blocks on mock scaffolds so that only high-yielding reactions (e.g. conversion greater than 75-85%) are used in actual DEL generation, it is commonly accepted that every DEL will contain some proportion of truncates – compounds with one or more building blocks missing, when compared to the intended full-cycle products; and that these truncated products are indistinguishable from their full-cycle counterparts on basis of the DNA tag.[12]

While presence of truncates in DEL mixtures is an inevitable artefact of the methodology used to prepare these libraries, understanding how they affect the results of DEL affinity screens is of significant relevance. Namely, identifying hits from DEL screens typically leverages the assumption that number of reads mapped to one tag is proportional to the binding affinity of a DEL member associated with that tag.[14,15] This assumption theoretically allows the investigator not only to identify the most potent binders in the library, but also derive structure-activity relationships (SARs) from affinity selection data, when suitable patterns emerge. As such, it has served as basis for interpreting the results of most selection experiments reported in the literature.[16–23]

Yet this assumption is, implicitly, subject to a number of constraints, as can be gleaned from previously reported results of computational simulations of DEL affinity selections.[24] Notably, for the count-affinity relationship to robustly hold: (i) all DEL members must be represented in equimolar amounts at the start

of a screen or, alternatively, the starting concentration of each DEL member must be known; (*ii*) each tag must unambiguously resolve to a single molecular entity; (*iii*) there must be sufficient sequencing coverage and low level of sequencing noise present; (*iv*) experimental conditions (e.g. concentration of the target protein, number of affinity selection cycles, number of wash steps per cycle etc.) must be carefully matched to the desired affinity of any binders one wishes to recover.

In practice, the above listed requirements will seldom be satisfied. Reactions used to generate individual DEL members have unequal yields, and it is not tractable to analytically determine them in complex mixtures. Likewise, as described above, presence of truncates that have tags identical to full-cycle products violates the assumption that compounds are uniquely tagged. Importantly, these truncates can have high binding affinities themselves. Furthermore, amplification and sequencing can introduce additional noise to the final output, constituting another important confounder.

It is appropriate to note that DEL affinity screens, even without any advanced post-processing of read counts, are able to recover high-affinity binders.[16–23] However, most of the time, leveraging raw data would lead to erroneous ranking of hits, and yield spurious, low-fidelity SARs.[25] In turn, this could potentially lead to undue experimentation in the follow up to a DEL screen.

This fact has motivated others to propose specific methods for processing affinity selection outputs. Satz[25] demonstrated that truncated products constitute a major confounder when analyzing binding assays based on raw read counts, and proposed a data aggregation approach to more robustly identify true patterns in selection outputs. Kuai et al.[26] performed a large-scale replicate selection experiment, and demonstrated how DEL selection outputs – of an identical library – tend to be intrinsically noisy, especially at low read counts. The authors further suggested that random noise in these experiments can reliably be modelled using a Poisson distribution, and proposed a specific normalization approach to transform raw counts into an enrichment metric with associated confidence intervals. Similarly, Faver et al.[27] modelled selection data using a binomial distribution, and developed a normalized z-score enrichment metric, demonstrating its utility in quantitative comparison of results from parallel selection experiments. More recently, Gerry et al.[28] described an analysis framework that also takes into account non-uniform abundance of individual DEL members in screened libraries. Leveraging pre- and post-selection read counts modelled by a superposition of multiple Poisson distributions, the authors developed a normalized fold-change score they subsequently used to rank binders.

Each of the described approaches attempts to correct for a single, or a few of the factors contributing to noise associated with DEL affinity screens. For example, accounting for random noise in sequencing outputs to normalize read counts still does not correct for representation imbalance in a DEL, nor does it account for the fact some of the counts attributed to full-cycle products may be inflated due to binding of identically tagged truncates. Likewise, aggregating read counts over cycles does help identify genuine enrichment in a noisy selection output, but only in a largely qualitative manner.

Here, we propose a method for processing DEL affinity selection outputs, which can be used to obtain high-fidelity binding affinity estimates for DEL members. Unlike previously reported approaches, our method seeks to account for all the major sources of noise in selection data simultaneously: truncated products bearing tags equivalent to those of their full-cycle counterparts, representation imbalance, and sequencing noise. We base our approach on a previously demonstrated, formally derivable relation between read counts and binding association constants,[24] which we extend leveraging two key assumptions. These two assumptions are both simple and well founded in DEL practice, as will later be discussed. One, under what we term the "null-block model", we assume the majority of DEL members that are not full-cycle products can be treated as simple truncates, i.e. full-cycle product analogues missing one or more building blocks at respective

diversity points. Two, we assume most DEL members will exhibit negligible affinity for the target of interest, with only a minor proportion binding with high association constants. With these two assumptions, and using data commonly available within a DEL generation and screening workflow, we developed a sparse learning method that, as demonstrated on simulated and publicly available data, can robustly recover true affinity rankings of DEL binders from inherently noisy selection data. Moreover, since our method also estimates truncates' binding affinity, it enables comprehensive evaluation of SAR. We also describe the implementation details of the method and make it freely accessible to the scientific community as a Python package and command line tool.

Results and Discussion

First, we introduce the basic notation used throughout this section. Then we show how a naive Bayes model can reduce sequence bias. Next, we proceed to our main result: first, describing our model, then showing how much noise suppression can be achieved on simulated and real experimental data. We finish this section with extensions to our model and details of its software implementation.

Notation

Each synthesized compound $\mathrm{L}_{r,q}$ is uniquely identified by its DNA tag $r = (r_1, r_2, \ldots r_C)$, which records the sequence of reactions in which the molecules have taken part, and the list of attached building blocks $q = (q_1, q_2, \ldots q_C)$. Here, C is the number of synthesis cycles, r_c is the index of the chemical reaction in cycle c, and q_c is the index of the building block which is actually attached in that cycle. Ideally, $q_c = r_c$, indicating a successful synthesis step (which happens with probability Y_{c,r_c}), but if a truncation happens (with probability $1 - Y_{c,r_c}$), we denote it with $q_c = 0$, marking the molecules that are missing the corresponding building block. The amount of each compound relative to the ideal amount of the corresponding full-cycle product, $\mathrm{L}_{r,r}$, can be written as a product over reaction yields,

$$\frac{[\mathrm{L}_{r,q}]_{\text{synthesized}}}{[\mathrm{L}_{r,r}]_{\text{ideal}}} = \prod_{c=1}^{C} y_c(r_c, q_c) \quad (1)$$

$$y_c(r_c, q_c) = \begin{cases} Y_{c,r_c}, & \text{if } q_c = r_c \\ 1 - Y_{c,r_c}, & \text{if } q_c = 0 \\ 0, & \text{otherwise.} \end{cases}$$

We denote the set of full-cycle building block combinations q, for which all $q_c = r_c$, with $\mathcal{F}$, and the set of truncated building block combinations q, where there is at least one cycle for which $q_c = 0$, with $\mathcal{T}$.

The binding assay depletes compounds whose building block combination q is unsuitable for interacting with the target protein. The fraction of molecules that survive C_{sel} number of selection cycles, (each of which consists of equilibrating the library with anchored proteins, washing away free ligands and dissociating the bound complexes), can be written as

$$\frac{[\mathrm{L}_{r,q}]_{\text{survived}}}{[\mathrm{L}_{r,q}]_{\text{synthesized}}} = \left(\frac{K_q[\mathrm{P}]}{1 + K_q[\mathrm{P}]}\right)^{C_{\text{sel}}} =: S_q, \quad (2)$$

where [P] is the concentration of the target protein and K_q is the association constant of the ligand $\mathrm{L}_{r,q}$ and the protein in the reaction $\mathrm{L}_{r,q}{:}\mathrm{P} \rightleftharpoons \mathrm{L}_{r,q} + \mathrm{P}$. (For derivation, see Supporting Information 1.2.) We call S_q the survival rate of building block combination q.

Sequencing the tags of the surviving molecules produces reads that can be mapped to the set of pre-defined tag sequences. After dropping low-confidence and noisy reads, we can summarize the data in a form of read counts N_r, each associated with a particular r tag. Taking into account the PCR amplification factor A, we can write the expected read counts as

$$\langle N_r \rangle = N_{\text{tot}} k A \sum_{q \in \mathcal{F} \cup \mathcal{T}} [\mathrm{L}_{r,q}]_{\text{survived}}, \quad (3)$$

where expectation value is denoted by $\langle \ldots \rangle$, N_{tot} is the total number of cleaned reads, and k is a protocol- and apparatus-specific normalization constant that mathematically converts

DNA concentration to number of sequencing reads.

Tag imbalance

Presence of DNA tags attached to compounds in a DEL is generally assumed not to govern the outcome of affinity selection assays. However, tags can be unevenly represented in sequencing outputs[28] even in absence of selection pressure. In other words even the $[\mathrm{L}_{r,r}]_{\mathrm{ideal}}$ concentrations can be different for different r tags, which, if disregarded, can adversely affect downstream analysis. Fortunately, such an imbalance would be apparent from results of sequencing performed *before* selection, and such data can be used to correct this bias.

We were able to investigate the extent of tag imbalance in pre-selection read counts thanks to data published by Gerry et al.,[28] which includes the full set of read counts from a $8 \times 114 \times 119$ DEL. In agreement with their findings, we identified two cycle-2 sequences and three cycle-3 sequences associated with higher than 2-fold tag imbalance, (see Supplementary Information 2.1). While the authors have attributed the imbalance to differential tag ligation efficiency, it is pertinent to note inhibition of the ligase (e.g. by leftover building blocks or catalysts), or PCR amplification bias can lead to similar artefacts. The latter can typically be avoided by including a degenerate region in the closing segment of a DNA tag,[13] which ultimately enables PCR deduplication. The tagging strategy employed by Gerry et al. seemingly does not make use of this approach. Irrespective of the source, however, this implies the need to accurately model pre-selection read counts. By assuming that the same experimental protocol is used for obtaining pre-selection results, we can write the expected read counts as

$$\begin{aligned}\langle N_r^{\mathrm{pre}} \rangle &= N_{\mathrm{tot}}^{\mathrm{pre}} k^{\mathrm{pre}} A^{\mathrm{pre}} \sum_{q \in \mathcal{F}\cup\mathcal{T}} [\mathrm{L}_{r,q}]_{\mathrm{synthesized}} \\ &= N_{\mathrm{tot}}^{\mathrm{pre}} k^{\mathrm{pre}} A^{\mathrm{pre}} [\mathrm{L}_{r,r}]_{\mathrm{ideal}}, \end{aligned} \tag{4}$$

where A^{pre} and k^{pre} are the amplification factor and the normalization constant, respectively, specific to the pre-selection sequencing experiment. By dividing Eq. 3 by Eq. 4, we can write the post-selection read counts as

$$\langle N_r \rangle = N_{\mathrm{tot}} \frac{kA}{k^{\mathrm{pre}} A^{\mathrm{pre}}} \frac{\langle N_r^{\mathrm{pre}} \rangle}{N_{\mathrm{tot}}^{\mathrm{pre}}} \sum_{q \in \mathcal{F}\cup\mathcal{T}} \frac{[\mathrm{L}_{r,q}]_{\mathrm{survived}}}{[\mathrm{L}_{r,r}]_{\mathrm{ideal}}}. \tag{5}$$

Directly estimating the $\langle N_r^{\mathrm{pre}} \rangle / N_{\mathrm{tot}}^{\mathrm{pre}}$ bias factor from the read counts N_r^{pre} is possible only if sequencing depth is at least 10, i.e. for libraries where complexity is lower than the total number of reads by at least one order of magnitude. Since many reported DELs have a numerical size between 10^7 and 10^9, sequencing to the required coverage would be highly impractical, and resource-intensive, even on contemporary NGS platforms. To enable tag imbalance bias correction for larger libraries, we use a naive Bayes model, where we assume that tags associated with one cycle contribute bias factors independent of tags of other cycles. With this assumption, we write the expectation value of N_r^{pre} (denoted by λ_r^{pre}) as a product of (much fewer) b_{c,r_c} bias parameters

$$\lambda_r^{\mathrm{pre}} := \langle N_r^{\mathrm{pre}} \rangle = N_{\mathrm{tot}}^{\mathrm{pre}} \prod_{c=1}^{C} b_{c,r_c}. \tag{6}$$

The maximum likelihood estimates of the bias factors (see Supporting Information 1.1),

$$\hat{b}_{c,r_c} = \frac{1}{N_{\mathrm{tot}}^{\mathrm{pre}}} \sum_{\substack{\rho \in \mathcal{F} \\ \text{s.t. } \rho_c = r_c}} N_\rho^{\mathrm{pre}}, \tag{7}$$

are robust even if the total number of reads is lower than library complexity. Since the complexity of a *single cycle* typically does not exceed 3×10^3, a mere total of one million reads is enough to estimate the b factors with at most 5% uncertainty. Substituting estimates of b from Eq. 7 to Eq. 6 yields the estimates $\hat{\lambda}_r^{\mathrm{pre}}$.

Using the data of Gerry et al.,[28] we verified that the naive Bayes model can provide significant noise reduction by comparing the dispersion (variance divided by the mean) of the raw read counts N_r^{pre} with the variance of the normalized residuals $z_r := (N_r^{\mathrm{pre}} - \hat{\lambda}_r^{\mathrm{pre}})/\sqrt{\hat{\lambda}_r^{\mathrm{pre}}}$, where $\hat{\lambda}_r^{\mathrm{pre}}$ is the prediction of the naive Bayes model, in Fig. 2. We found $\mathrm{Var}(N_r^{\mathrm{pre}}) \,/\, \overline{N_r^{\mathrm{pre}}} =$

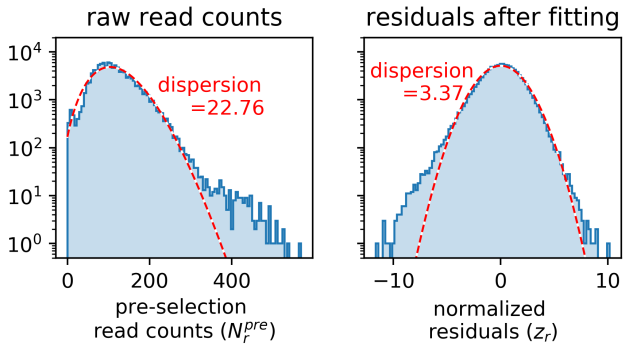

Figure 2: Noise reduction achieved by the naive Bayes model on pre-selection sequencing data from Gerry et al.[28] **Left:** Distribution of raw read counts and fitted dispersed Poisson distribution (dashed line), the variance of which is 22.76 times higher than its mean. **Right:** Distribution of normalized residuals $z_r = (N_r - \hat{\lambda}_r)/\sqrt{\hat{\lambda}_r}$, and the best fitting normal distribution (dashed line), which has a variance of 3.37.

22.76 and $\mathrm{Var}(z_r) = 3.37$, suggesting that the naive Bayes model is capable of reducing the variance due to tag imbalance by a factor of 7. Furthermore, while the distribution of the read counts has a heavier tail than what a dispersed Poisson distribution can explain, the high values of the residuals are accurately accounted for by the corresponding normal distribution. This demonstrated the capability of the naive Bayes model to suppress noise due to tag bias.

Truncated compounds

Ideally, all Y_{c,r_c} yields are 100%, the concentration of all truncates $[\mathrm{L}_{r,q\neq r}]$ are zero, and each expected read count $\langle N_r \rangle$ is affected by only one compound, $\mathrm{L}_{r,r}$. In reality, most $Y_{c,r_c} < 100\%$, which, combined with the possibility that some truncates q can exhibit significant binding to the target protein, i.e. $S_q > 0$, allows truncates to survive the selection process and masquerade as full-cycle products. Here, we construct a statistical model capable of estimating S_q up to an unknown global constant from the individual reaction yields Y_{c,r_c} and pre- and post-selection read counts, N_r^{pre}, N_r. It is worth noting that Y_{c,r_c} are typically unknown, as separation and analytical quantification of individual DEL members is largely impractical. However, as we demonstrate later, yields of individual reactions on mock scaffolds, routinely obtained during building block validation, are a suitable surrogate for actual yields.

Combining Eqs. 1, 2 and 5 leads to the following linear relationship between $\langle N_r \rangle$ and the survival probabilities S_q.

$$\begin{aligned} \langle N_r \rangle &= \sum_{q \in \mathcal{F} \cup \mathcal{T}} X_{r,q} F_q, \quad \text{where} \\ X_{r,q} &= N_{\mathrm{tot}} \frac{\langle N_r^{\mathrm{pre}} \rangle}{N_{\mathrm{tot}}^{\mathrm{pre}}} \prod_{c=1}^{C} y(r_c, q_c), \\ F_q &= \frac{kA}{k^{\mathrm{pre}} A^{\mathrm{pre}}} S_q, \end{aligned} \tag{8}$$

where we separated the different factors so that X contains the known variables, and F the unknowns. The F_q values, by virtue of being proportional to the survival probabilities S_q, indicate how well each building block combination q withstands the selection step; we call F_q the "fitness" of q. We assume independent, Poisson-distributed sequencing noise,

$$P(N_r = n \,|\, \langle N_r \rangle = \lambda) = e^{-\lambda} \frac{\lambda^n}{n!} \tag{9}$$

which, together with Eq. 8, completes the main definition of our model. One may note that our model can be described as a generalized linear model with Poisson noise and identity link function over non-negative coefficients.

This model is under-determined: there are fewer data points N_r ($r \in \mathcal{F}$) than unknown coefficients F_q ($q \in \mathcal{F} \cup \mathcal{T}$). This comes at no surprise, because there is no read count which would directly inform us about truncated compounds. To enable robust inference, despite this difficulty, we turn to sparse machine learning techniques,[29] which work well under the assumption that many coefficients are exactly zero. Such an assumption is naturally fitting for DEL screens, where many of the full-cycle products are expected to not bind to the target protein, i.e. their association constant $K_q \ll [\mathrm{P}]^{-1}$, leading to $S_q \approx 0$ and $F_q \approx 0$.

From the long list of published sparse regular-

ization strategies,[30] we chose the one used by LASSO[31] because of its mathematical compatibility with the Poisson distribution. Namely, we assume a common exponential prior density for the fitness coefficients,

$$P(F_q) = \alpha e^{-\alpha F_q}. \tag{10}$$

As we show below, this choice allows us to average over the fitness of full-cycle products and compute the marginal posterior of the fitness of truncates in closed form. The new parameter, α, controls how strongly the model favors sparse solutions. Practical considerations (see Supporting Information 1.3) suggest that the optimal value is one tenth of sequencing depth. To simplify our exposition, we stick to this choice, but we also give the more general version of all following mathematical formulas, as well as their derivations, in Supporting Information 1.4 and 1.6.

According to Bayes theorem, the posterior of F, $P(F \,|\, N)$, is equal to the product of the likelihood $P(N \,|\, F)$ and the prior $P(F)$, up to a normalization constant.

$$\begin{aligned} P(F \,|\, N) &\propto P(N \,|\, F)\, P(F) \\ &\propto \left[\prod_{r\in\mathcal{F}} P(N_r \,|\, F)\right] \left[\prod_{q\in\mathcal{F}\cup\mathcal{T}} P(F_q)\right] \\ &\propto \left[\prod_{r\in\mathcal{F}} e^{-\lambda_r} \lambda_r^{N_r}\right] \left[\prod_{q\in\mathcal{F}\cup\mathcal{T}} e^{-\alpha F_q}\right], \end{aligned} \tag{11}$$

where $\lambda_r = \langle N_r \rangle$ is a function of F, defined in Eq. 8. Due to the specific structure of $y_c(r_c, q_c)$ (see Eq. 1), only one full-cycle product affects each N_r read count, the one where $q = r$, i.e. $X_{r,q} \neq 0$ only if $q \in \{r\} \cup \mathcal{T}$. This allows us to write the expected read count as

$$\lambda_r = \sum_{q\in\mathcal{F}} X_{r,q} F_q + \sum_{q\in\mathcal{T}} X_{r,q} F_q = X_{r,r} F_r + B_r$$

where $B_r = \sum_{q\in\mathcal{T}} X_{r,q} F_q$ denotes the contribution of truncates. Our main goal here is to estimate B_r background for all r tags. After that, we will be able to estimate F_r directly from N_r and $X_{r,r}$.

Due to the compatibility of the $e^{-\lambda_r}$ and the $e^{-\alpha F_q}$ factors, we can average over the full-cycle fitness coefficients, $F^{(\mathcal{F})} = \{F_q\}_{q\in\mathcal{F}}$, and obtain a closed-form expression for the posterior of truncate fitness coefficients, $F^{(\mathcal{T})} = \{F_q\}_{q\in\mathcal{T}}$,

$$\begin{aligned} P(F^{(\mathcal{T})} \,|\, N) &= \int_{F^{(\mathcal{F})}} P(F \,|\, N) \\ &\propto \left[\prod_{r\in\mathcal{F}} \Gamma_r\, e^{\frac{B_r}{X_{r,r}}}\right] \left[\prod_{q\in\mathcal{T}} e^{-\alpha F_q}\right], \end{aligned} \tag{12}$$

where $\Gamma_r = \Gamma\big(N_r + 1, (1 + 1/X_{r,r})B_r\big)$, and $\Gamma(s, x) = \int_x^\infty z^{s-1} e^{-z}\, dz$ is the upper incomplete gamma function, which is efficiently implemented in numerical software libraries.

The expression of the marginal posterior $P(F^{(\mathcal{T})} \,|\, N)$ in Eq. 12 can be numerically maximized with coordinate descent[32] that minimizes the cost function, $f(F^{(\mathcal{T})}) = -\log P(F^{(\mathcal{T})} \,|\, N)$. We terminate the optimization once all B_r background contributions change less than 0.1 between consecutive iterations. We found that it requires fewer than 100 iteration cycles to converge. Any optimum found is guaranteed to be the global optimum because the cost function f is convex everywhere (see Supplementary Information 1.5 for proof).

Once the fitness of truncates $F_q \in F^{(\mathcal{T})}$ are estimated and the background $\hat{B}_r$ is computed, we can estimate each full-cycle fitness, $F_r \in F^{(\mathcal{F})}$,

$$\hat{F}_r = \max\left(0,\ \frac{N_r}{X_{r,r} + 1} - \frac{\hat{B}_r}{X_{r,r}}\right). \tag{13}$$

This takes the prior of F into account, which is apparent from the $+1$ term in the denominator of the first term, which would be missing from the maximum likelihood result. This completes our quest to estimate F.

From the estimates $\hat{F}$, we can compute the most likely breakdown of post-selection read counts. This splits each observed read count N_r into different $N_{r,q}$ contributions ($N_r = \sum_q N_{r,q}$) each counting how many reads come specifically from ligand $\mathrm{L}_{r,q}$. Under the assumption of Poisson sequencing noise, the posterior of read count breakdown $\{N_{r,q}\}_{q\in\mathcal{F}\cup\mathcal{T}}$ is multinomial.

We use its conditional expectation value to estimate it:

$$\hat{N}_{r,q} = \langle N_{r,q} \rangle_{|N_r, \hat{F}} = N_r \frac{X_{r,q} \hat{F}_q}{\sum_{q'} X_{r,q'} \hat{F}_{q'}}. \qquad (14)$$

The estimates $\hat{N}_{r,q=r}$ can be regarded as a denoised version of the input data N_r, from which the effects of truncates have been statistically subtracted. Although it is tempting to consider $N_{r,q}$ the final output of our model, fitness F is a more reliable metric for distinguishing compounds by affinity, as we show next.

Benchmarking on simulated data

We investigated the accuracy and robustness of our model on data simulated under a large variety of realistic conditions. First, we describe the settings of the simulation, then we summarize our findings.

Libraries from 1 million to 1 billion compounds were simulated, the default size being 100 million. The number of synthesis cycles were chosen to be 2, 3, 4, 5 and 6, with default being 3. Reaction yields Y_{c,r_c} were sampled from a beta distribution with mean from the range $[40\%, 80\%]$ (default: 70%) and fixed standard deviation of 10 percentage points. The association constants K_q were chosen relative to the inverse protein concentration $[\mathrm{P}]^{-1}$. Their logarithms, $\log_{10} K_q$, were drawn from a half-normal distribution (with $\sigma = 1.5$, to match realistic spread[25]) whose minimum was chosen to result in a pre-defined "density", i.e. the fraction of K_q values being larger than $[P]^{-1}$. We simulated libraries with densities between 10^{-6} and 10^{-2}, default being 10^{-4}. Fig. 3 shows the distribution of simulated yields and association constants with the default choice of parameters. Sequencing depth, i.e. the average number of reads per DNA tag, were chosen to be between 10^{-4} and 10, with default being 0.1. We also investigated how robust our model is against measurement uncertainty of the yields, considering the fact that we expect yields from mock scaffolds to be used in actual computation. To do this, we fed noisy yield values to the model, where noise standard deviation was

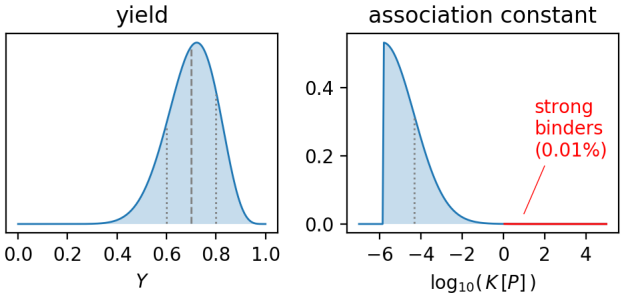

Figure 3: Distributions of simulated yields and association constants. **Left:** Yields were drawn from a beta distribution, the default parameters of which were chosen to give a mean of 0.7 and standard deviation of 0.1. The mean and the $\pm\sigma$ values are marked by dashed and dotted lines, respectively. **Right:** Logarithm of association constants (relative to the inverse protein concentration $[\mathrm{P}]^{-1}$) were drawn from a half-normal distribution with $\sigma = 1.5$. By default, the minimal value of $\log_{10}(K[\mathrm{P}])$ was chosen to result in 0.01% of the compounds be strong binders, i.e. have $K > [\mathrm{P}]^{-1}$. Shading marks this region, and the dashed line marks the value 1σ above the minimum.

ranging from 0 to 25 percentage points, default being 15. Finally, we tested how much accuracy we lose if we neglect to correct tag imbalance by not providing pre-sequencing data to the model, forcing it to assume the absence of any bias. The logarithm of tag imbalances $\log_{10} \lambda_r^{\mathrm{pre}}$ were sampled from a normal distribution centered at 0, the standard deviation of which we increased gradually from 0 to $\log_{10}(300\%)$, the default being $\log_{10}(200\%)$. We set the number of selection cycles to $C_{\mathrm{sel}} = 2$. After computing $X_{r,q}$ and F_q, from the simulated yields and association constants, we drew each read count $N_{r,q}$ (broken down by q) from a Poisson distribution with mean $X_{r,q} F_q$. The simulated "observed" read counts were computed by the sum $N_r = \sum_q N_{r,q}$.

We conducted seven sequences of numerical experiments, one for each parameter (library size, cycles, mean yield, density, sequencing depth, yield noise, tag imbalance), where we change the value of the selected parameter while keeping all other parameters at default value. To evaluate how well the model performed on

the simulated data, we visualize the estimated number of full-cycle reads $\hat{N}_{r,r}$ and the estimated fitness values $\hat{F}_q$ vs the true read counts and association constants. We include the plots here for the simulation where each parameter was set to its default value (100 million compounds, 3 cycles, 70% mean yield, 10^{-4} density, 0.1 sequencing depth, 15% yield noise, tag imbalance of 1.0), and provide plots from all runs in Supporting Information 2.2. Square roots of observed N_r and estimated $N_{r,r}$ are plotted against the true $N_{r,r}$ and $\log_{10}(K[\mathrm{P}])$ in Fig. 4, showing that $\hat{N}_{r,r}$ is confounded by less noise than raw N_r. Fig. 5 shows the fitness of full-cycle products and truncates as functions of

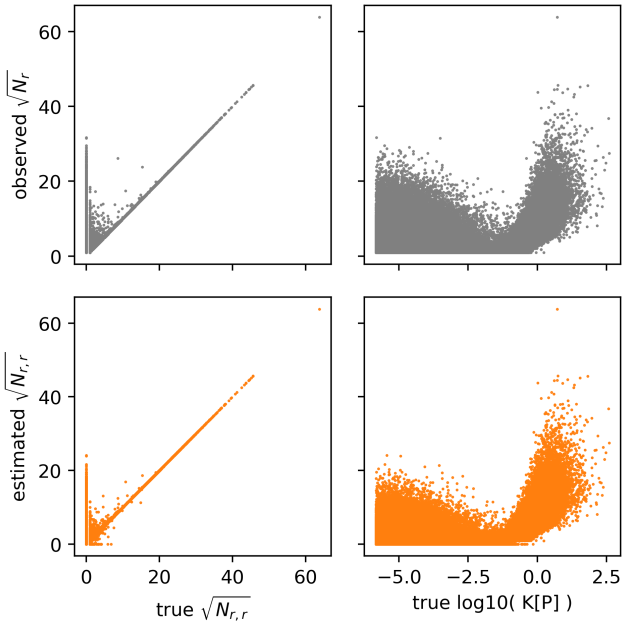

Figure 4: Comparison of the square roots of the observed read counts N_r and the estimated and true full-cycle counts $N_{r,r}$ on data simulated with default parameters. Association between N_r and binding strength is poor because truncates masquerade as full-cycle products and inflate their counts. The estimates $\hat{N}_{r,r}$ predict the true $N_{r,r}$ with reduced noise on the lower end. (We plot the square root, instead of the actual counts, because they are subject to a Poisson noise, meaning that $\mathrm{Var}(N) \approx N$, but $\mathrm{Var}(\sqrt{N}) \approx 0.5$, i.e. $\sqrt{N}$ is approximately homoscedastic, providing a more intuitive comparison of statistical significance at different levels of N.)

true association constant. Fitness of truncates are estimated accurately, and high fitness values of full-cycle products are enriched among strong binders. Comparison with Fig. 4 suggests that $\hat{F}_r$ is as good or better predictor of strong binding as $\hat{N}_{r,r}$.

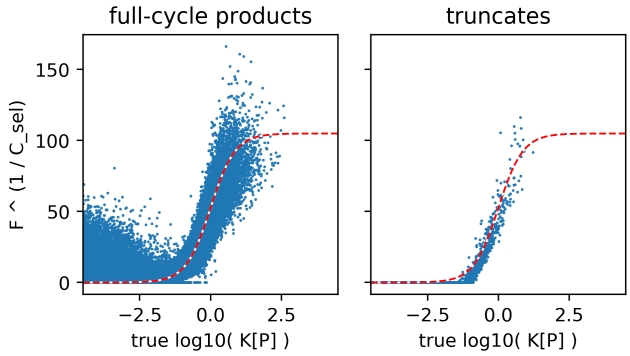

Figure 5: Estimated and true fitness F_q (scattered points and dashed lines, respectively) for full-cycle products and truncates as functions of $\log_{10}(K_q[\mathrm{P}])$. Association between K_q and $\hat{F}_q$ is strong for both groups. Low fitness values are overestimated. For high fitness values, truncates are accurately recovered, but full-cycle products retain some of the original noise. (We plot the transformed fitness values, i.e. $(F_q)^{1/C_{\mathrm{sel}}}$, because their true curve is symmetric around its center point at $\log_{10}(K[\mathrm{P}]) = 0$, and independent of C_{sel}.)

The main purpose of the DEL screen is to distinguish strongly-binding ligands (which we define with the condition $K_q > [\mathrm{P}]^{-1}$) from weakly-binding ones. We compute how well different metrics perform in this binary classification problem. We compare the classification performance, i.e. false discovery rate (FDR) and false negative rate (FNR), of three metrics: observed read count N_r, estimated full-cycle read counts $\hat{N}_{r,r}$, and estimated fitness coefficients $\hat{F}_r$. We define FDR and FNR with

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} \tag{15}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TP}}, \tag{16}$$

where false positives (FP) refer to compounds that are above a chosen threshold of the metric

but bind weakly, false negatives (FN) are compounds that are below the threshold but bind strongly, and true positives (TP) are the ones where the metric correctly predicts strong binding. One may note that FDR = 1 − Precision and FNR = 1 − Recall, using the usual definitions of Precision and Recall.[33]

By plotting corresponding FDR and FNR values for a sequence of different thresholds, we visualize the detection error trade-off[34] (DET) curves of the three metrics. This is shown on Fig. 6 for the simulation with default parameters. At the point, where FDR = 5%, the

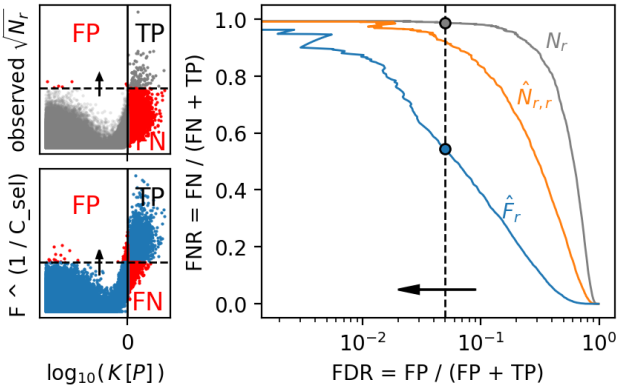

Figure 6: Detection error trade-off curve for three metric: observed read count N_r, estimated full-cycle read counts $\hat{N}_{r,r}$ and estimated fitness $\hat{F}_r$. The vertical line marks FDR = 5%, and the two smaller plots on the left shows the origin of the false positives (FP), false negatives (FN) and true positives (TP) for two metric (N_r and $\hat{F}_r$). Increasing the acceptance threshold of the metrics, a change depicted by the arrows, lowers FDR but raises FNR.

false negative rates of the three metrics are 99%, 92% and 54%, for N_r, $\hat{N}_{r,r}$ and $\hat{F}_r$, respectively. This means, if one had access to only the raw read counts but needed to keep false discovery rate below 5%, then the acceptance threshold would have to be set so high that 99% of strong binders would be missed. By using the estimated fitness coefficients to assess binding, only 54% of the strong binders would be missed at the same FDR level, a 50-fold increase in Recall.

To determine under what circumstances a particular metric performs well, we selected two indicators: First, the FNR at the point where FDR = 5%, which indicates how reliably a metric is able to distinguish strong binders from weak binders. Second, the Spearman-correlation between the metric and K_q among strong binders, which measures how accurately can the metric rank compounds by their binding strength. We plot these metrics for six of the seven sequences of numerical experiments in Fig. 7, and for tag imbalance in Fig. 8.

For the observed counts N_r, FNR stays above 90%, whereas $\hat{F}_r$ can achieve significantly lower FNR under all tested library sizes, cycles, mean yields, density values highlights the usefulness of the null-block model as a post-processing step.

Although FNR and Spearman rank correlation reflect different facets of model accuracy, their trends are mirror images of each other: when FNR is low, Spearman correlation is high for $\hat{N}_{r,r}$ and $\hat{F}_r$ across all seven parameters.

Metric $\hat{N}_{r,r}$ fails to improve Spearman correlation beyond the level already achievable with N_r, and improves FNR only marginally. Because computing $\hat{F}_r$ is possible from $\hat{N}_{r,r}$ only if we know the yields with some degree of accuracy, this highlights that knowing the yields is important for being able to predict strongly-binding compounds. This is also highlighted by the 6th sub-figure column in Fig. 7, which shows that FNR is increasing and Spearman correlation is decreasing quickly as the measurement uncertainty of the yields grows beyond 0.1.

The most difficult conditions, where all three metrics have a difficult time estimating strong binders, are low mean yield ($\leq$ 0.5), low sequencing depth ($\leq$ 2), high density of strong binders ($\geq$ 3%), and high yield noise ($\geq$ 0.15). It is encouraging to see that the performance of $\hat{N}_{r,r}$ and $\hat{F}_r$ are constant with increasing library size, suggesting that libraries much larger than the ones we simulated can be accurately analyzed with our model. As the number of synthesis cycles is increasing from 3 to 6, the null-block model loses some accuracy, but this is less of a concern, because the number of cycles is limited to 4 in most DELs.

Sensitivity to tag imbalance shows a similar trend as sensitivity to yield noise, but here we

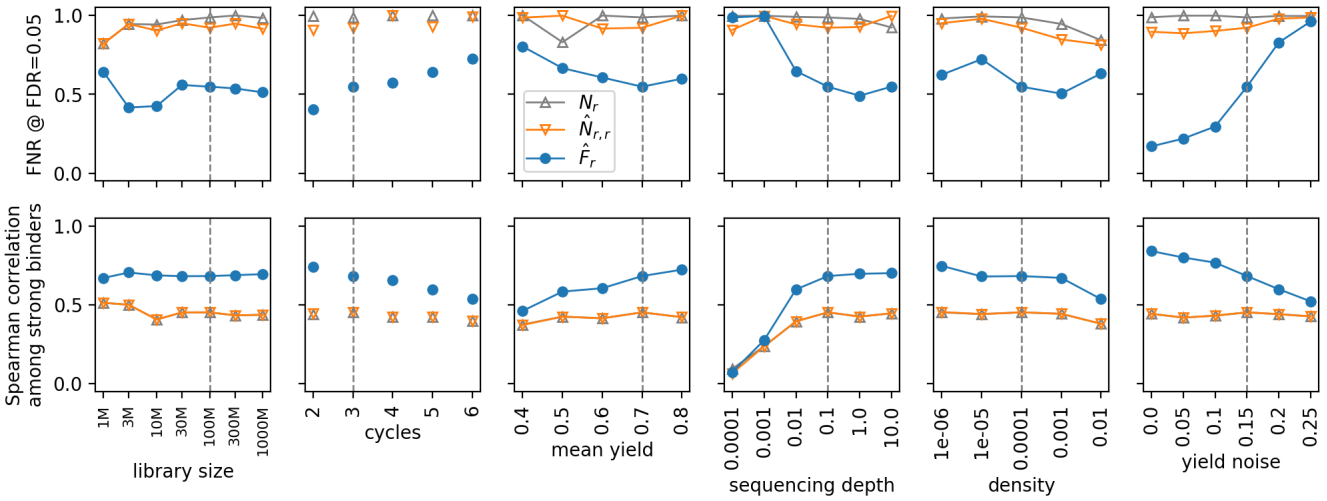

Figure 7: False negative rate (at the point where false discovery rate is 5%) for three different metrics (N_r, $\hat{N}_{r,r}$, $\hat{F}_r$) and Spearman's rank correlation between the metric and the true association constant K_r among strong binders, plotted as functions of six simulation parameters: library size, number of synthesis cycles, mean of reaction yields, sequencing depth, density (fraction of strong binders), and noise on measured yields. Vertical dashed lines mark the default values of each parameter, which was used in all other sequences of simulations where other parameters were changed.

distinguish between two scenarios. First, we do not use pre-selection sequencing data, and implicitly assume that there is no tag imbalance. Second, we fit the naive Bayes model to pre-sequencing read count data and correct tag imbalance. Fig. 8 shows that correcting tag imbalance greatly improves FNR and Spearman correlation for the metric $\hat{F}_r$, but does nothing for $\hat{N}_{r,r}$. This is expected, since the true read count breakdown $N_{r,r}$ (which $\hat{N}_{r,r}$ estimates) is subject to the same tag imbalance bias as the observed read counts N_r.

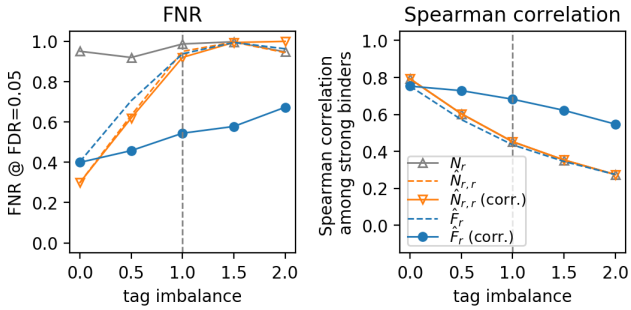

Figure 8: False negative rate and Spearman correlation for the three metrics, N_r, $\hat{N}_{r,r}$ and $\hat{F}_r$, as functions of tag imbalance. Dashed lines indicate uncorrected results. (Here, "tag imbalance" is defined by the formula $10^\sigma - 1$, where σ is the standard deviation of $\log_{10} \lambda_r^{\text{pre}}$, which is normally distributed, e.g. tag imbalance of 0.5 means that $\sigma = \text{std}(\log_{10} \lambda_r^{\text{pre}}) = \log_{10}(1 + 0.5) = 0.17$.)

Benchmarking on experimental data

We show that high fitness values, estimated by the null-block model, are good indicators of strong binding on real experimental data. Once again, we make use of the supplementary data published by Gerry et al.[28]

We take advantage of the fact that index 23 denotes a "null" reaction in cycle 3, as designed by the authors. All read counts $N_{(r_1,r_2,r_3)}$ where $r_3 = 23$ actually measure the fitness of one of the truncates $q = (r_1, r_2, 0)$, where r_1 and r_2 are non-zero cycle-1 and cycle-2 indexes, respectively. First, we run our model on the pre- and post-sequencing data, and obtain estimates of all truncate fitness values F_q, for $q = (r_1, r_2, 0)$, without informing our model about the existence of a "null" reaction (in fact, using $Y_{3,23} = 0.99$, which implies that $r_3 = 23$

is a proper reaction). Then we compute the observed fold change, N_r/N_r^{pre} for all r of the from $(r_1, r_2, 23)$. In Fig. 9, we compare the two

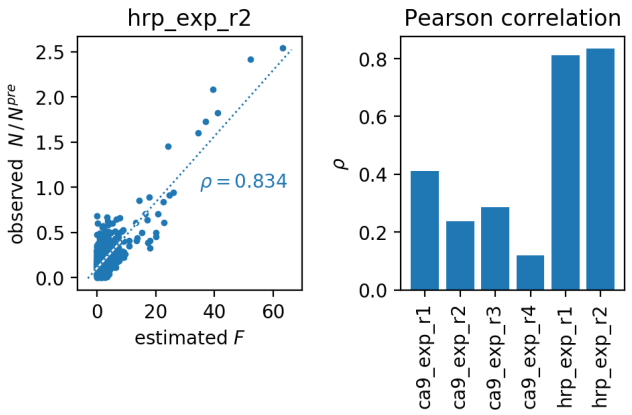

Figure 9: **Left:** Comparison of estimated fitness $\hat{F}_q$ and observed fold change N_r/N_r^{pre} for cycle-3 truncates, i.e. $q = (r_1, r_2, 0)$, and $r = (r_1, r_2, 23)$, because reaction 23 in cycle 3 is a "null" reaction. The Pearson correlation of the two from the experiment "hrp_exp_r2" is 0.834. **Right:** Person correlation from all six experiments. (Scatter plots of all six experiments are included in Supporting Information 2.3.)

metrics, and compute their Pearson correlation. Their correlation is statistically significant in all six experiments, and especially strong in the two "hrp_exp" data sets. This shows that estimated fitness of truncated compounds, which is central for denoising, is consistent with what one can obtain with a more direct measurement of the truncates.

We further evaluated the concordance of results obtained using our method to those discussed in Gerry et al.[28] For horseradish peroxidase (HRP), the authors observed a strong dependency between enrichment and the electrophilic character of the N-capping group (building block 1, BB1), further highlighting three sulfonyl chloride-derived Michael acceptors, ranked according to their electrophilicity, which furnished compounds of highest affinity for HRP. As can be seen in Fig. 10, results obtained using "deldenoiser" consistently recapitulate the aforementioned dependency, which is broadly observable across most scaffolds and the second diversity point, suggesting the latter two play only a minor role in governing binding. Similarly, for carbonic anhydrase IX (CA9), "deldenoiser" successfully recovered aryl sulfonamide-based building blocks at the second diversity point (building block 2, BB2) as privileged substructures, particularly in *trans* stereoisomers of azetidine derivatives (e.g. scaffold ID 8); the activity of these products was experimentally validated off-DNA by Gerry et al. Finally, the preference for *para*-sulfonamide building blocks over their *meta*- positional isomers is also captured (Fig. 10). It is pertinent to note these trends do not become readily appreciable when using a simple count-based enrichment metric (see Figure 10 in Supporting Information), primarily due to numerous outliers; however, an aggregate visualization does render them obvious.

Discovering SARs

Structure-activity relationships (SARs) manifest themselves as elevated read counts along one or more reaction indexes. E.g. if the combination of reaction 47 in cycle-1 and reaction 38 in cycle-2 already produces a structure that binds strongly to the target, irrespective of the cycle-3 reaction, then most $r = (47, 38, r_3)$ tags will have high read counts. Unfortunately, such a data set is similar to what a high-affinity truncate $q = (47, 38, 0)$ would produce. Still, we can clean the read count data from the effect of potential truncates and faithfully retain the SAR signature, if the design of the DEL includes "null" reactions, i.e. synthesis steps where nothing is done to the compounds (see details in "Extensions" subsection). To continue our example, let us imagine that $r_3 = 101$ is a "null" reaction. The, usually low, read count $N_{(47,38,101)}$ prevents the fitness of the affecting truncate, $F_{(47,38,0)}$, from being overestimated. Accurate estimation of the truncate fitness, even if it is much smaller than all full-cycle fitness values, enables accurate estimation of the entire SAR series.

We demonstrate this capability of our model by first simulating DEL results with default parameter settings, (see "Benchmarking on simu-

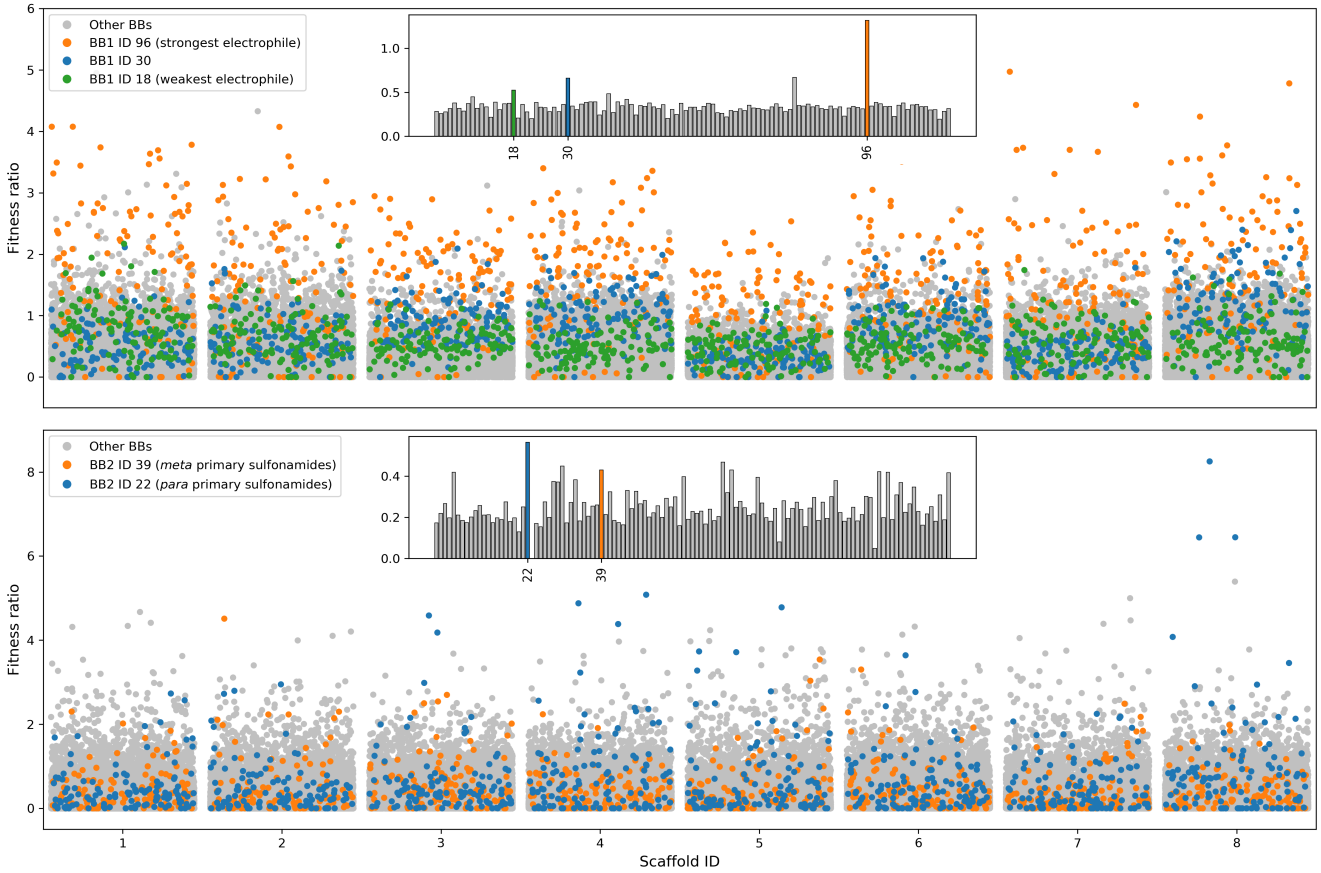

Figure 10: Results from the Gerry et al.[28] affinity screens, reanalyzed using "deldenoiser". **Top:** Ratio of mean fitness values (across replicates) for protein-loaded vs. beads-only affinity selections against horseradish peroxidase (HRP), grouped according to scaffold, and colored by the presence of a specific building block, as summarized in the legend. Inset bar chart provides an aggregate representation of the data, grouped by the identity of BB1, with height of the bars reflecting the mean fitness ratio across all species with the corresponding building block. **Bottom:** Corresponding results for carbonic anhydrase IX. Inset bar chart reflects grouping on the identity of BB2.

lated data" subsection), and adding ten artificial SAR series to each data set. We randomly selected ten 1-index SAR series, where values of all except one r_c reaction indexes are fixed, and drew their association constants from the $K_q \geq [\mathrm{P}]^{-1}$ tail of their distribution. We chose all $K_q[\mathrm{P}]$ of the confounding truncates to be a fixed value from 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, and 100. We used the direct estimation formulas (see Supporting Information 1.10) to fit our model. To benchmark our model, we compute the same metrics as before, but this time, we focus on the compounds that are part of the SAR series: i.e. Spearman correlations and false negative rates are evaluated only on the set of these compounds. Fig. 11 shows the false negative rate at the threshold where the false discovery rate is 5% (which we compute among *all* full-cycle compounds), and Spearman correlation of the three metrics ($N_r, \hat{N}_{r,r}$ and $\hat{F}_r$) among SAR compounds, plotted as a function of $\log_{10}(K_q[\mathrm{P}])$ of the truncate q that directly confounds the SAR series. For $\hat{F}_r$, false negative rate hovers around 0.5, independent of truncate binding strength, while Spearman correlation slowly decreases from 0.75 to 0.6 as $\log_{10}(K[\mathrm{P}])$ increases from -2 to +2, a significant improvement over what is achievable with the two other metrics N_r and $\hat{N}_{r,r}$. This suggests that our model, with the help of read counts of "null" reactions, can accurately separate the effects of truncates, no matter their affinity, and recover

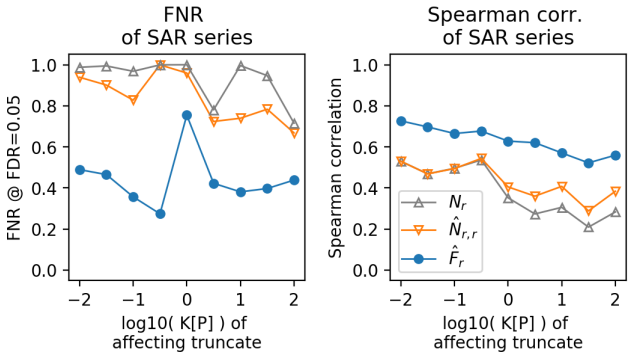

Figure 11: False negative rate and Spearman correlation of full-cycle products that are part of SAR series. The log-association constant $\log_{10}(K_q[\mathrm{P}])$ of the truncate that directly affects the series is plotted on the x axis.

the true SARs. More details and plots of these numerical benchmarking experiments are available in Supporting Information 2.4.

Extensions

The null-block model, as presented above, can be extended in several ways. Here we give a summary of the different avenues, and details can be found in the Supporting Information.

First, if the DEL is designed to include "null" reactions, as has commonly been reported,[17,20,21,35] then the null-block model can be adapted to take this information into account and estimate truncate fitness not only from the background contributions B_r, but also from the read counts corresponding to these "null" reactions. Setting $Y_{c,r_c} = 0$ for the "null" reaction indexed by r_c in cycle c results in $X_{r',r'} = 0$ for all r' where $r'_c = r_c$. This forces the model to try to explain the read count $N_{r'}$ using only truncates, providing a boost to the accuracy of estimating truncate fitness values, which translates to more accurate estimates of the full-cycle products. Such an input data also enables direct fitting of our model, as we explain in Supporting Information 1.10.

Second, although we found that the naive Bayes model can account for the majority of the effects resulting in overdispersion of the read counts, one may wish to model the remaining dispersion (which we found to be about 3.37 for data from Gerry et al.). This can be efficiently done by replacing the Poisson distribution in Eq. 9 with a dispersed Poisson distribution,

$$P(n \mid \lambda, d) = Z(\lambda, d)\frac{(\lambda/d)^{(n/d)}}{\Gamma(n/d+1)} \qquad (17)$$

defined on $n = 0, 1, 2, \ldots$, where $Z(\lambda, d) \approx e^{\lambda/d}/d$, and Γ is the gamma function. The dispersion parameter d determines the ratio of variance and expectation value, i.e. $\mathrm{Var}(n) \approx d\langle n\rangle \approx d\lambda$. Because the dispersed Poisson distribution has the same mathematical compatibility with the exponential prior as the Poisson distribution, the integral prescribed in Eq. 12 can still be evaluated in closed form, and the algorithm remains computationally efficient. Formulas can be found in Supplementary Information 1.4 and 1.6.

Third, so far we were dealing with the problem of estimating the fitness F_q, and we did not discuss what is needed to estimate the survival chance S_q and association constant K_q. The added difficulty stems from not knowing the factor k/k^{pre}. To calibrate this value, we need additional information. Sequencing on Illumina machines is performed with an additional DNA spike-in to be used as positive control, and which helps maintain diversity in libraries originating from DELs.[15] Sequences from the virus PhiX are added to the prepared DNA library in controlled amount. Taking note of the added amounts and the number of reads mapping to PhiX genome provide sufficient information to estimate the ratio k/k^{pre}. Alternatively, one could use a compound with known binding affinity, resynthesized on-DNA, and spiked into the library at a reasonable concentration. Combined with the PCR amplification rates A and A^{pre}, which can be estimated from the experimental protocol, one can compute the proportionality constant between F_q and S_q, enabling direct computation of $\hat{S}_q$ from the estimated fitness $\hat{F}_q$. Then, the protein concentration [P] and the number of selection cycles C_{sel} can be used to estimate K_q. Formulas can be found in Supporting Information 1.7. Note that depending on the details of the experimental protocol, equations more involved than Eq. 2 may

be needed to establish the relation between S_q and K_q, but this is beyond the scope of this article.

Fourth, our model is equally applicable for pooled DEL screens. After separating the data from a pooled screen by libraries, our algorithm can be run on each part separately to obtain fitness estimates. Added difficulty is created by the fact that concentrations of the pooled libraries may be uncertain. To overcoming this, pre-selection sequencing data is crucial. Previously we used pre-selection sequencing data to correct tag imbalance due to unequal sequence generation, but the exact same method corrects the effect of unequal concentrations between libraries. More details can be found in Supporting Information 1.8. Furthermore, analyzing pooled libraries is directly parallelizable, saving wall clock time.

Finally, our formalism allows developing more complex truncation models. Whenever the exact same side product get produced alongside different full-cycle products, correlated noise is present in the observed data, which opens the possibility of deconvolving the side product from the full-cycle products. Only the formula for computing the X matrix needs to be changed to incorporate reaction branches, and the rest of the machinery will function without change.

Limitations

An inherent difficulty in developing a method for analysis of DEL affinity selections is the lack of experimental data that can serve as a robust ground truth dataset in method validation. Ideally, one would desire having association constants for several hundred compounds on-DNA, along with the sequencing data produced after the affinity selection of the parent library. In this scenario, one would not need to account for factors contributing to differences in binding affinity when hits are resynthesized off-DNA, facilitating a straightforward comparison. In reality, however, published data typically includes only biochemical or biophysical affinity measurements for off-DNA compounds, with the number of data points seldom exceeding 10-20, and with no or incomplete sequencing data released. In validating the described method, we therefore resorted to heavily leveraging simulated data. While we made every effort to ensure this data was generated by incorporating reasonable assumptions pertaining to the affinity selection and sequencing steps, these involve complex thermodynamic processes, and are subject to numerous potential experimental artifacts. This renders any simulated data only a modest surrogate for real selection outputs. Correspondingly, prospective validation of our method is highly warranted, and will be instrumental in confirming its practical utility. None the less, the results we obtained by analyzing data from Gerry et al. – which, to date, presents the most comprehensive publicly available DEL selection dataset – provide encouraging evidence to this end. It should be noted, though, that even this dataset contains only 8 IC50 measurements for library members synthesized off-DNA, precluding any statistically robust assessment. Further advances in algorithmic developments supporting DEL screens, as well as systematic cross-method benchmarks, would benefit greatly from broader availability of additional datasets akin to that released by Gerry et al.

Beyond real-world validation of the method we described, its underpinnings features a key limiting assumption that is worth further underlining. The null-block model assumes all library members are either full-cycle (intended) products or their truncated counterparts. Yet, chemistry involved in many library designs is clearly conducive to the generation of highly diverse arrays of side products. These are challenging to account for on several grounds: (*i*) their presence does not typically result in correlated noise; (*ii*) they render library complexity an almost arbitrary parameter; (*iii*) no estimates on yields of individual side products can be obtained. Although a more complex model can be envisioned to address these challenges, we opted against pursuing it. For one, most practitioners will choose to include only high-yielding building blocks into the final library, and will optimize designs and reaction conditions to minimize the generation of alternate

products. Secondly, the bulk of DEL literature cites truncates as key contributors to spurious SAR, with significant utility coming from determining their putative affinity. Finally, a more complex model would ultimately be of questionable utility, given the inherent difficulties in fitting it, and validating its performance. Prospective adopters should, none the less, be aware of this limitation, as libraries that are suspected to contain complex side product mixtures in non-negligible quantities may not be suited for analysis by this approach.

Software implementation

The null-block model can be fitted with our software implementation "deldenoiser". We developed a python package and command line tool under the same name, and made them available under GNU General Public License v3.0 at `https://github.com/totient-bio/deldenoiser`.

The input data about reaction yields Y_{c,r_c}, pre- and post-selection read counts N_r^{pre}, N_r, are processed and output files are created that contain the estimated read count breakdown $\hat{N}_{r,r}$ and fitness coefficients $\hat{F}_q$. The logical flow is shown in Fig. 12.

Our implementation, based on python's numpy framework,[36] takes 12 minutes to run for a library of 100 million compounds, using 8 CPUs and 1.5 GB of memory. Analyzing bigger libraries require more computational resources. Fig. 13 shows how running time and peak memory usage increases with increasing library size, reaching 5.5 hours and almost 2.6 GB for a library of 1 billion compounds.

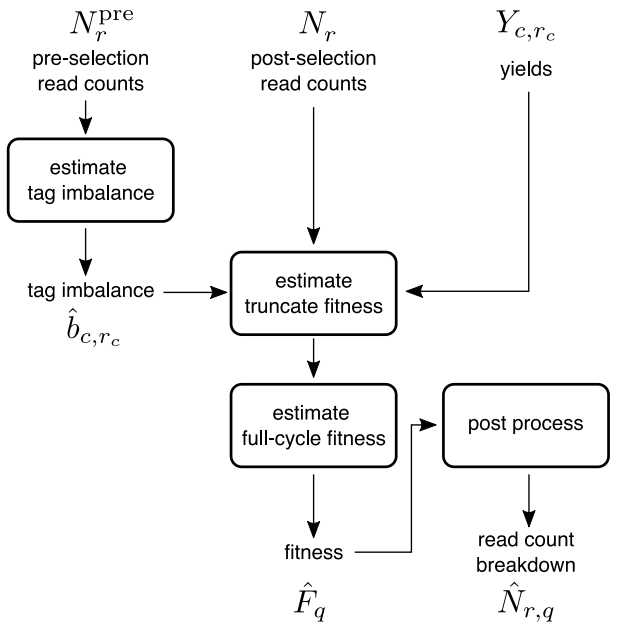

Figure 12: Logical flow of "deldenoiser" algorithm. From the input data the tag imbalances b are estimated. Then the fitness of the truncates is estimated, which is the most computationally intensive step, followed by estimation of the fitness of the full-cycle products. Finally, the fitness values $\hat{F}_q$ are used to estimate the read count breakdown $\hat{N}_{r,q}$.

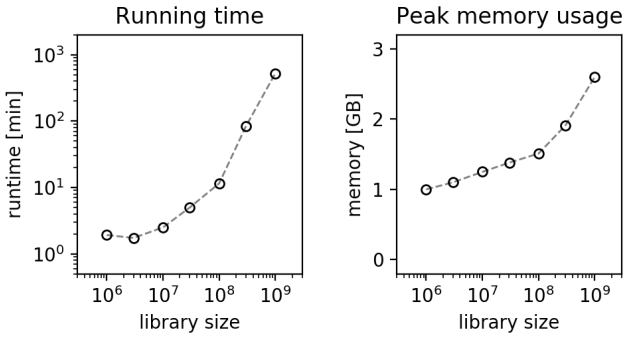

Figure 13: Running time and maximum memory usage of "deldenoiser" command line tool running on 8 CPUs, as functions of library size.

Conclusion

We developed and benchmarked a statistical method capable of reducing the noise affecting sequencing results of DNA encoded libraries due to truncated compounds. Numerical experiments conducted with simulated data showed that one can select and rank strongly binding compounds more reliably with the metric produced by our model, compared to using the raw read counts of the DEL screen. Comparison with experimental methods that directly measure the effects of truncated compounds confirmed the validity of our model.

Author Information

M.K. identified the problem, set the scope of the work, and advised on numerical experiments.

P.K. designed, implemented the benchmarked the statistical model. P.K. and M.K. co-wrote the manuscript and Supporting Information.

Acknowledgement The authors thank Yilong Li for advice on developing the statistical model, and reviewing the manuscript.

Supporting Information Available

deldenoiser-SupportingInformation.pdf: The Supporting Information contains mathematical formulas and derivations, detailed plots for all benchmarking results, and user documentation for the command line tool "deldenoiser".

References

(1) Clark, M. A. et al. Design, synthesis and selection of DNA-encoded small-molecule libraries. *Nature Chemical Biology* **2009**, *5*, 647–654.

(2) Goodnow, R. A.; Dumelin, C. E.; Keefe, A. D. DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nature Reviews Drug Discovery* **2017**, *16*, 131–147.

(3) Kunig, V.; Potowski, M.; Gohla, A.; Brunschweiger, A. DNA-encoded libraries – an efficient small molecule discovery technology for the biomedical sciences. *Biological Chemistry* **2018**, *399*, 691–710.

(4) Satz, A. L. What Do You Get from DNA-Encoded Libraries? *ACS Medicinal Chemistry Letters* **2018**, *9*, 408–410.

(5) Machutta, C. A. et al. Prioritizing multiple therapeutic targets in parallel using automated DNA-encoded library screening. *Nature Communications* **2017**, *8*, 16081.

(6) Franzini, R. M.; Neri, D.; Scheuermann, J. DNA-Encoded Chemical Libraries: Advancing beyond Conventional Small-Molecule Libraries. *Accounts of Chemical Research* **2014**, *47*, 1247–1255.

(7) Chan, A. I.; McGregor, L. M.; Liu, D. R. Novel selection methods for DNA-encoded chemical libraries. *Current Opinion in Chemical Biology* **2015**, *26*, 55–61.

(8) Shi, B.; Zhou, Y.; Huang, Y.; Zhang, J.; Li, X. Recent advances on the encoding and selection methods of DNA-encoded chemical library. *Bioorganic & Medicinal Chemistry Letters* **2017**, *27*, 361–369.

(9) Mannocci, L.; Zhang, Y.; Scheuermann, J.; Leimbacher, M.; De Bellis, G.; Rizzi, E.; Dumelin, C.; Melkko, S.; Neri, D. High-throughput sequencing allows the identification of binding molecules isolated from DNA-encoded chemical libraries. *Proceedings of the National Academy of Sciences* **2008**, *105*, 17670.

(10) Buller, F.; Zhang, Y.; Scheuermann, J.; Schäfer, J.; Bühlmann, P.; Neri, D. Discovery of TNF Inhibitors from a DNA-Encoded Chemical Library based on Diels-Alder Cycloaddition. *Chemistry & Biology* **2009**, *16*, 1075–1086.

(11) Favalli, N.; Bassi, G.; Scheuermann, J.; Neri, D. DNA-encoded chemical libraries – achievements and remaining challenges. *FEBS Letters* **2018**, *592*, 2168–2180.

(12) Satz, A. L. *A Handbook for DNA-Encoded Chemistry*; John Wiley & Sons, Ltd, 2014; Chapter 5, pp 99–121.

(13) Creaser, S. P.; Acharya, R. A. *A Handbook for DNA-Encoded Chemistry*; John Wiley & Sons, Ltd, 2014; Chapter 6, pp 123–151.

(14) P. Hale, S. *A Handbook for DNA-Encoded Chemistry*; John Wiley & Sons, Ltd, 2014; Chapter 13, pp 281–317.

(15) Decurtins, W.; Wichert, M.; Franzini, R. M.; Buller, F.; Stravs, M. A.; Zhang, Y.; Neri, D.; Scheuermann, J. Automated screening for small organic

ligands using DNA-encoded chemical libraries. *Nature Protocols* **2016**, *11*, 764–780.

(16) Deng, H. et al. Discovery, SAR, and X-ray Binding Mode Study of BCATm Inhibitors from a Novel DNA-Encoded Library. *ACS Medicinal Chemistry Letters* **2015**, *6*, 919–924.

(17) Ding, Y.; O'Keefe, H.; DeLorey, J. L.; Israel, D. I.; Messer, J. A.; Chiu, C. H.; Skinner, S. R.; Matico, R. E.; Murray-Thompson, M. F.; Li, F.; Clark, M. A.; Cuozzo, J. W.; Arico-Muendel, C.; Morgan, B. A. Discovery of Potent and Selective Inhibitors for ADAMTS-4 through DNA-Encoded Library Technology (ELT). *ACS Medicinal Chemistry Letters* **2015**, *6*, 888–893.

(18) Litovchick, A.; Dumelin, C. E.; Habeshian, S.; Gikunju, D.; Guié, M.-A.; Centrella, P.; Zhang, Y.; Sigel, E. A.; Cuozzo, J. W.; Keefe, A. D.; Clark, M. A. Encoded Library Synthesis Using Chemical Ligation and the Discovery of sEH Inhibitors from a 334-Million Member Library. *Scientific Reports* **2015**, *5*, 10916.

(19) Samain, F.; Ekblad, T.; Mikutis, G.; Zhong, N.; Zimmermann, M.; Nauer, A.; Bajic, D.; Decurtins, W.; Scheuermann, J.; Brown, P. J.; Hall, J.; Gräslund, S.; Schüler, H.; Neri, D.; Franzini, R. M. Tankyrase 1 Inhibitors with Drug-like Properties Identified by Screening a DNA-Encoded Chemical Library. *Journal of Medicinal Chemistry* **2015**, *58*, 5143–5149.

(20) Yang, H. et al. Discovery of a Potent Class of PI3Kα Inhibitors with Unique Binding Mode via Encoded Library Technology (ELT). *ACS Medicinal Chemistry Letters* **2015**, *6*, 531–536.

(21) Deng, H. et al. Discovery and Optimization of Potent, Selective, and in Vivo Efficacious 2-Aryl Benzimidazole BCATm Inhibitors. *ACS Medicinal Chemistry Letters* **2016**, *7*, 379–384.

(22) Ahn, S. et al. Allosteric "beta-blocker" isolated from a DNA-encoded small molecule library. *Proceedings of the National Academy of Sciences* **2017**, *114*, 1708–1713.

(23) Cuozzo, J. W.; Centrella, P. A.; Gikunju, D.; Habeshian, S.; Hupp, C. D.; Keefe, A. D.; Sigel, E. A.; Soutter, H. H.; Thomson, H. A.; Zhang, Y.; Clark, M. A. Discovery of a Potent BTK Inhibitor with a Novel Binding Mode by Using Parallel Selections with a DNA-Encoded Chemical Library. *ChemBioChem* **2017**, *18*, 864–871.

(24) Satz, A. L. DNA Encoded Library Selections and Insights Provided by Computational Simulations. *ACS Chemical Biology* **2015**, *10*, 2237–2245.

(25) Satz, A. L. Simulated Screens of DNA Encoded Libraries: The Potential Influence of Chemical Synthesis Fidelity on Interpretation of Structure-Activity Relationships. *ACS Combinatorial Science* **2016**, *18*, 415–424.

(26) Kuai, L.; O'Keeffe, T.; Arico-Muendel, C. Randomness in DNA Encoded Library Selection Data Can Be Modeled for More Reliable Enrichment Calculation. *SLAS Discovery* **2018**, *23*, 405–416.

(27) Faver, J. C.; Riehle, K.; Lancia, D. R.; Milbank, J. B. J.; Kollmann, C. S.; Simmons, N.; Yu, Z.; Matzuk, M. M. Quantitative Comparison of Enrichment from DNA-Encoded Chemical Library Selections. *ACS Combinatorial Science* **2019**, *21*, 75–82.

(28) Gerry, C. J.; Wawer, M. J.; Clemons, P. A.; Schreiber, S. L. DNA Barcoding a Complete Matrix of Stereoisomeric Small Molecules. *Journal of the American Chemical Society* **2019**, *141*, 10225–10235.

(29) Qiao, L.-b.; Zhang, B.-f.; Su, J.-s.; Lu, X.-c. A systematic review of structured sparse learning. *Frontiers of Information Technology & Electronic Engineering* **2017**, *18*, 445–463.

(30) Piironen, J.; Vehtari, A. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing* **2017**, *27*, 711–735.

(31) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **1996**, *58*, 267–288.

(32) Wright, S. J. Coordinate descent algorithms. *Mathematical Programming* **2015**, *151*, 3–34.

(33) Powers, D. M. W. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies.* **2011**, *2*, 37–63.

(34) Martin, A.; Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M.; Przybocki, M. The DET Curve in Assessment of Detection Task Performance. Proc. Eurospeech '97, Rhodes, Greece. 1997; pp 1895 – 1898.

(35) Ding, Y.; Chai, J.; Centrella, P. A.; Gondo, C.; DeLorey, J. L.; Clark, M. A. Development and Synthesis of DNA-Encoded Benzimidazole Library. *ACS Combinatorial Science* **2018**, *20*, 251–255.

(36) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **2011**, *13*, 22–30.

Graphical TOC Entry