INFORMATION-DENSE SYNTHESIS FOR MOLECULAR DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning can accelerate molecular discovery by designing molecules and planning experiments. However, many scientific challenges demand molecules with very rare properties, and in this sparse setting, existing algorithms offer little gain over random guessing. We propose a method to efficiently search large regions of molecular space using algorithmically controlled stochastic synthesis. Rather than design, make and test individual molecules, we design and make complex mixtures, test them as a pool, then deconvolute the molecule-activity map. We optimize synthesis to encode maximal information. Theoretically, this approach can reduce the number of experiments required to find the optimal molecule among d candidates from $\mathcal{O}(d)$ to $\mathcal{O}(\log d)$ or $\mathcal{O}(1)$. In simulation, on estimated peptide fitness landscapes, it finds active molecules with an order of magnitude fewer experiments than existing Bayesian optimization methods.

1 INTRODUCTION

Lab-in-the-loop ML and AI systems promise to efficiently explore molecular space, by iteratively learning from data and generating new molecules to test (e.g. Romero & Arnold, 2009; Romero et al., 2013; Frey et al., 2025; Rapp et al., 2024; Liu et al., 2025b; Mints et al., 2022; Hastrup et al., 2026). However, ML systems still struggle on hard discovery problems. Very often, the region of molecular space where molecules have a desired property is exceedingly narrow: within the combinatorial vastness of possible molecules, only a small patch catalyzes a reaction or treats a disease. This *sparseness* poses fundamental challenges to existing lab-in-the-loop systems, which design and test candidate molecules individually. Most candidates fail, and therefore provide almost no information about the optimal molecule. On sparse problems, even state-of-the-art algorithms perform no better than random screening (Proposition 1) (Bach, 2020).

In this article we propose lab-in-the-loop learning with *information-dense* synthesis (LIDS), a method that designs, builds and tests complex mixtures of molecules, rather than individual molecules. It is tractable to synthesize complex mixtures, without added laboratory time or cost, by combinatorial chemistry (Szostak, 1997; Houghten et al., 1999). Our idea is to control stochastic synthesis algorithmically, within the framework of Bayesian optimization and experimental design (Rainforth et al., 2024; Shahriari et al., 2016). We use generative models to describe stochastic synthesis (Weinstein et al., 2022; 2026). We test the synthesized mixture as a pool (Liu et al., 2025a). We optimize synthesis so the experiment provides maximum information about the molecule with the highest activity (Hernández-Lobato et al., 2014; Russo & Van Roy, 2018). The method yields rapid acceleration on sparse molecular discovery problems (Fig. 1).

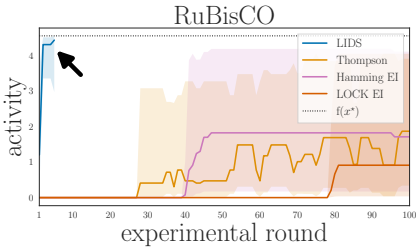


Figure 1: Lab-in-the-loop speedup. Optimizing an *in silico* oracle with LIDS (blue) versus baselines (Sec. 6).

Our contributions are: (1) A recipe to accelerate lab-in-the-loop molecular discovery by controlled stochastic synthesis, rather than deterministic synthesis. (2) Efficient approximate Bayesian computation for designing stochastic syntheses, and learning from assays that test the resulting mixtures as a pool (Sec. 3). (3) Theoretical results showing exponential speedups over methods that design and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

test molecules individually (Sec. 5). (4) Empirical demonstration of order-of-magnitude speedups across 6 simulated DNA and peptide activity landscapes (Sec. 6). Notation is defined where introduced; a table is in App. A.

2 Problem Setting

Our goal is to discover molecules with a desired property. Let $\mathcal{X}$ denote the space of molecules we are exploring, and let $\mathrm{f} : \mathcal{X} \to \mathbb{R}$ denote the mapping from molecule to activity. Our goal is to find $x^\star \in \mathrm{argmax}_x \mathrm{f}(x)$. For example, $\mathcal{X}$ could be the space of peptide sequences of length L, i.e. $\mathcal{X} = \{A, C, \ldots\}^L$, and $\mathrm{f}(x)$ could be activation of a target receptor, e.g. GLP-1. We study lab-in-the-loop optimization methods where we query $\mathrm{f}(x)$ by running an experiment.

Current approach. The usual technique for solving this problem is to iteratively (1) design a molecule x_t, (2) synthesize x_t in the laboratory, (3) test x_t in an assay that returns y_t, a noisy measurement of $\mathrm{f}(x_t)$, and (4) update our knowledge of f and its optimum. Algorithms based on Bayesian optimization or reinforcement learning can be used to learn from the data $\mathcal{D}_{t-1} = \{(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\}$ and propose new molecules x_t (steps 1 and 4) (Romero & Arnold, 2009; Romero et al., 2013; Maus et al., 2026; Groth et al., 2024; Jankowiak et al., 2026). Robotic systems are increasingly used to automate synthesis and testing (steps 2 and 3) (Ha et al., 2023). Compiled together, steps 1-4 have become a basic recipe for automated or "self-driving" labs in the molecular sciences (Rapp et al., 2024; Abolhasani, 2026).

Core challenge. Active molecules often occupy a narrow region of the molecular space. For example: antibody screens turn up one useful sequence among a billion candidates (Skora et al., 2015); flipping a single chiral center in thalidomide makes the drug toxic (Vargesson, 2015); 0.14% of single point mutations in RuBisCO, a CO_2 capturing enzyme, increase activity over baseline (Prywes et al., 2025). The sparseness of molecule-activity maps poses a fundamental challenge to existing lab-in-the-loop discovery methods. To illustrate, consider a simple setting in which we have $d \triangleq |\mathcal{X}|$ candidate molecules, but only one is active, and we have no prior knowledge about which one. Then there does not exist a learning algorithm that can outperform a random screen.

Proposition 1. *Assume* $\mathrm{f}(x) = \mathbb{I}(x = x^\star)$ *where* $x^\star \sim \mathrm{Uniform}(1, \ldots, d)$*, and there is no noise in the assay, so we measure* $y_t = \mathrm{f}(x_t)$*. Define expected regret,* $R \triangleq \mathbb{E}_{x^\star}[\mathbb{E}_\pi[\sum_{t=1}^{\infty} \mathbb{I}(X_t \neq X^\star)]]$*. A policy* $\pi(x_t \mid \mathcal{D}_{t-1})$ *that selects the next candidate* x_t *uniformly at random from among remaining candidates has expected regret* $(d-1)/2$*. There does not exist a policy with lower expected regret.*

Proof in App. B. In short, no algorithm–using Bayesian optimization, reinforcement learning, generative models, agents, or anything else–can find the optimum in less than $(d+1)/2$ experiments on average. The problem is that on sparse problems there is no information to extrapolate from: an experiment only tells us whether or not the tested molecule works.

Our approach. To break the $(d-1)/2$ regret barrier of Proposition 1, we propose designing, synthesizing and testing mixtures of molecules, so that each experiment carries information about many molecules at once (Aldridge et al., 2026; Russo & Van Roy, 2018; Liu et al., 2025a).

3 Method

We optimize $\mathrm{f}(x)$ by synthesizing and testing mixtures of molecules. Mathematically, we specify a mixture as a positive *measure* $q(x)$: a distribution over different molecules, $\bar{q}(x) \in \mathcal{P}(\mathcal{X})$, times a yield $0 < c < \infty$, the total number of molecules that are made, $q(x) = c \cdot \bar{q}(x) \in \mathcal{M}^+(\mathcal{X})$. When we test the mixture, we assume the assay provides a noisy measurement of the average activity,

$$y_t = \int \mathrm{f}(x) q_t(x) \mathrm{d}x + \epsilon_t = \langle \mathrm{f}, q_t \rangle + \epsilon_t \tag{1}$$

where ϵ_t is noise, and $\langle \mathrm{f}, \mathrm{g} \rangle \triangleq \int \mathrm{f}(x)\mathrm{g}(x)\mathrm{d}x$ denotes the inner product on functions. In the case where we synthesize a pure compound, the distribution of molecules $q_t(x)$ reduces to a delta mass at a single point x_t, and we recover the standard problem setting, $y_t = \mathrm{f}(x_t) + \epsilon_t$.

Eq. (1) contains a linearity assumption: it says the experiment's outcome depends on the sum of the activity of each x, $\mathrm{f}(x)$, weighted by how much of x is in the mixture, $q_t(x)$. This is an assumption about the *assay*, not about the activity map $\mathrm{f} : \mathcal{X} \to \mathbb{R}$, which can be arbitrarily complex and nonlinear. For example, in spectroscopy, the total signal depends on the sum of the absorbance or emission of each individual molecule, even if the absorbance and emission depend in a complex way on molecular structure. *In vitro* or *in vivo*, linearity violations are *drug-drug interactions*, which are often lower order contributions to a cocktail's effect (Liu et al., 2025a).

Our recipe for lab-in-the-loop learning consists of:

1. A synthesis model that describes the chemical process of creating a mixture. It specifies which $q(x) \in \mathcal{M}^+(\mathcal{X})$ are practically achievable.
2. An assay model that describes the noise associated with measurement, ϵ_t.
3. A design algorithm that chooses the next mixture q_t to maximize information.
4. An inference algorithm that updates our knowledge of f based on the experimental data.

3.1 Synthesis models

We aim to design a complex mixture of molecules that is tractable to synthesize in the lab. To do so, we build on variational synthesis, which describes the output of stochastic synthesis using a generative model (Weinstein et al., 2022; 2026). Our synthesis model consists of a family of measures $\mathcal{Q} = \{q_\psi : \psi \in \Psi\} \subseteq \mathcal{M}^+(\mathcal{X})$, where (1) ψ denotes experimental parameters, i.e. specific instructions that can be sent to a synthesizer, and (2) $q_\psi(x)$ describes the mixture of molecules that come out of the synthesizer. To design a mixture we choose a $\psi \in \Psi$.

We focus on models for solid phase synthesis, a combinatorial synthesis approach that is widely applied to make DNA, RNA, peptides, carbohydrates and more (Beaucage & Caruthers, 1981; Caruthers, 2013; Merrifield, 1963; Zuckermann et al., 1992). Consider synthesizing peptides of length L. The molecular space is $\mathcal{X} = \mathcal{B}^L$ where $\mathcal{B}$ is the alphabet of amino acids, size $B \triangleq |\mathcal{B}|$. A solid phase peptide synthesizer makes peptides by sequentially adding amino acids. At each step it can add a single amino acid, or it can add a mixture of different amino acids, so that each growing peptide molecule will randomly encounter and react with a different amino acid. The resulting mixture can be described by a site-wise independent distribution $\mathrm{Cat}(x \mid \beta_1) = \prod_{j=1}^{L} \beta_{1jx_j}$ where the probability of amino acid x_j at position j depends on its effective concentration in the jth reaction, β_{1jx_j}, with $\beta_1 \in (\Delta_{B-1})^L$ (Houghten et al., 1999). By repeating the synthesis process S times with different input concentrations, and mixing the products, we build up a distribution,

$$\bar{q}_\psi(x) = \sum_{s=1}^{S} \alpha_s \mathrm{Cat}(x \mid \beta_s), \tag{2}$$

where $\alpha \in \Delta_{S-1}$ specifies the relative concentration of peptides from each synthesis s, and $\beta \in (\Delta_{B-1})^{S \times L}$. This is a discrete mixture model over sequences. We set $\mathcal{Q} = \{q_{\psi=(\alpha,\beta,c)} : c \sum_{s=1}^{S} \alpha_s \mathrm{Cat}(x \mid \beta_s)\}$ as the family of synthesizable distributions, where $\psi = (\alpha, \beta, c)$ are the experimental parameters, including the overall yield c. This synthesis model is applicable to a range of heteropolymers, e.g. for DNA we can set $\mathcal{B}$ to be the nucleotides (Weinstein et al., 2022; 2026).

The laboratory time and resources required to synthesize the mixture distribution in Eq. (2) is effectively the same as that required to synthesize S pure sequences $x_1, \ldots, x_S$. The number of reaction steps and the amount of reagents are the same. The only added complexity is from mixing reagents before or during each synthesis step, a process amenable to automated liquid handling (Torres-Acosta et al., 2022).

3.2 Assay models

We assume a model of the assay's noise. We focus on Gaussian noise, $\epsilon_t \sim \mathrm{Normal}(0, \sigma)$. Though not required, our method exploits prior knowledge of the noise level, e.g. from control experiments, so we assume the standard deviation σ is known. Bernoulli or Poisson noise is also possible.

3.3 DESIGNING MIXTURES

We work in the framework of sequential Bayesian experimental design (Rainforth et al., 2024). We optimize synthesis to maximize the information our next experiment will deliver about the optimum. This acquisition function has been proposed in predictive entropy search and information directed sampling (Hernández-Lobato et al., 2014; Russo & Van Roy, 2018). First, we place a prior on the unknown molecule-activity map,

$$\theta \sim \mathrm{p}(\theta) \tag{3}$$

$$Y_t \sim \mathrm{p}(y \mid \langle \mathrm{f}_\theta, q_t \rangle) \tag{4}$$

After performing $t-1$ experiments, our posterior updates to $\mathrm{p}(\theta \mid \mathcal{D}_{t-1} = \{(y_1, q_1), \ldots, (y_{t-1}, q_{t-1})\})$. To choose the next mixture, we maximize the expected information gain (EIG) about the optimum: $q_t = \operatorname*{argmax}_{q_\psi \in \mathcal{Q}} \mathrm{EIG}_{x^\star}(q_\psi) =$

$$\operatorname*{argmax}_{q_\psi \in \mathcal{Q}} [\mathcal{H}(\mathrm{p}(x_\theta^\star \mid \mathcal{D}_{t-1})) - \mathbb{E}_{\mathrm{p}(y_t \mid q_\psi, \mathcal{D}_{t-1})}[\mathcal{H}(\mathrm{p}(x_\theta^\star \mid \mathcal{D}_{t-1} \cup \{(Y_t, q_\psi)\}))]] \tag{5}$$

where $\mathcal{H}$ denotes the Shannon entropy and $x_\theta^\star \triangleq \operatorname{argmax}_x \mathrm{f}_\theta(x)$. This objective quantifies the reduction in uncertainty about $x^\star$ after observing y_t, on average across predicted y_t. It is equivalent to the mutual information between the new datapoint, (y_t, q_t), and the optimal molecule $x^\star$ (Hennig & Schuler, 2012; Hernández-Lobato et al., 2014; 2015; Jones et al., 1998; Russo & Van Roy, 2018). Assays are often run in parallel, e.g. in a 96 well plate format, in which case we can consider a batch version of the objective, with $\mathcal{D}_{t-1} \cup \{(Y_1^t, q_{\psi_1}), \ldots, (Y_E^t, q_{\psi_E})\}$.

The EIG is critical to exploiting stochastic synthesis. Thompson sampling draws $\theta \sim \mathrm{p}(\theta \mid \mathcal{D}_{t-1})$ and chooses the next design as $q_t = \operatorname{argmax}_{q \in \mathcal{Q}} \langle \mathrm{f}_\theta, q \rangle = \delta_{x_\theta^\star}$, i.e. it reverts to synthesizing and testing individual molecules (Thompson, 1933). The upper confidence bound algorithm and expected improvement produce the same degeneracy (App. C) (Auer et al., 2002; Jones et al., 1998). As a result, these standard acquisition functions fail to provide major speedups on sparse problems, even when given access to stochastic synthesis (Sec. 6) (Russo & Van Roy, 2018).

Approximating the information gain The EIG is difficult to compute explicitly. Instead, we optimize the Barber-Agakov variational bound (Foster et al., 2020; Barber & Agakov, 2003),

$$\mathrm{EIG}_{x^\star}(q_{\psi_{1:E}^t}) \geq \mathcal{H}(\mathrm{p}(x_\theta^\star \mid \mathcal{D}_{t-1})) + \mathbb{E}_{\mathrm{p}(\theta \mid \mathcal{D}_{t-1})\mathrm{p}(\epsilon)}[\log r_\phi(x_\theta^\star \mid \langle \mathrm{f}_\theta, q_{\psi_{1:E}^t} \rangle + \sigma \epsilon_{1:E})] \tag{6}$$

$\triangleq \mathcal{I}_{BA}(x^\star; q_\psi, \phi)$, where $r_\phi(x^\star \mid y)$ is the approximate posterior and $\mathrm{p}(\epsilon) = \mathrm{Normal}(0, 1)$ is the reparameterized noise (App. D). We first draw samples from the posterior $\theta_{1:M} \overset{iid}{\sim} \mathrm{p}(\theta \mid \mathcal{D}_{t-1})$ (Sec. 3.4), then jointly optimize the synthesis q_ψ and the approximate posterior r_ϕ by stochastic gradient ascent, using AdamW (Loshchilov & Hutter, 2018; Foster et al., 2020). Details on architecture and training are in App. E. This variational bound can be understood as an encoder-decoder model (Grover & Ermon, 2019). The laboratory experiment *encodes* the unknown molecule-activity map f_θ into a noisy E-dimensional measurement $y_{1:E}^t = \langle \mathrm{f}_\theta, q_{\psi_{1:E}^t} \rangle + \sigma \epsilon_{1:E}$. The approximate posterior *decodes* the optimal molecule from this measurement.

A tractable sparse function class We need a function class and prior that can describe sparse molecule-activity maps, which can be challenging to model with Gaussian processes (Park, 2022; Meng et al., 2021). We consider a single-layer Bayesian neural network with an exponential nonlinearity, $\mathrm{f}_{\theta=(W,V)}(x) = W^\top \exp(V \cdot x)$ where $x \in \mathcal{B}^L$ is represented as a one-hot encoding, $W \in \mathbb{R}^N$, $V \in \mathbb{R}^{N \times L \times B}$, and the dot product is over the length dimension L and alphabet dimension B. We design a prior that encourages sparsity, so only x that closely match one of the filters $V_1, \ldots, V_N$ produce nontrivial outputs, and assume that we have an informative prior on the maximum activity that the assay can record, $\|\mathrm{f}\|_\infty$, based on e.g. positive control experiments. This function class is expressive, and offers closed-form computation of the integral $\int \mathrm{f}(x) q_\psi(x) \mathrm{d}x$, as well as fast approximation of $x^\star$ (App. F).

3.4 Inference from mixtures

We perform automated approximate Bayesian inference in a probabilistic programming language (NumPyro), adaptively switching from MCMC to stochastic variational inference as the posterior concentrates (App. G). When done gathering data, we generate the maximum *a posteriori* molecule for further testing $\hat{x}^\star = \text{argmax}_x \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}[x = x^\star_{\theta_j}]$. The full procedure is summarized in Alg. 1.

Algorithm 1 **L**earning with **I**nformation-**D**ense **S**ynthesis (LIDS)

Require: batch size $E \in \mathbb{N}$, syntheses per batch $S \in \mathbb{N}$, noise σ, posterior samples $M \in \mathbb{N}$, prior $\pi(\theta)$.
while conducting experiments, $t = 1, 2, 3, \ldots$ **do**
  $q^t_{1:E}, \phi_t = \text{argmax} \mathcal{I}_{BA}(x^\star; q_\psi, \phi)$ ▷ Max EIG
  $y^t_{1:E} = \langle \mathrm{f}, q^t_{1:E} \rangle + \sigma \epsilon^t_{1:E}$ ▷ Run experiments
  $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(y^t_i, q^t_i)_{i=1}^E\}$
  $\theta^t_{1:M} \overset{iid}{\sim} \mathrm{p}(\theta \mid \mathcal{D}_t)$ ▷ Update beliefs
end while
return $\hat{x}^\star_t = \text{argmax}_x \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}[x = x^\star_{\theta^t_j}]$

4 Related Work

Our approach builds on group testing, which searches among many candidates by conducting pooled tests (Aldridge et al., 2026). A key advantage of pooling in chemistry is that it does not require DNA barcoding or otherwise tagging individual molecules. Pooled testing is related to compressed sensing, which exploits sparsity to learn from limited measurements (Atia & Saligrama, 2009). These ideas have a long history in molecular discovery (Kainkaryam & Woolf, 2009). Liu et al. (2025a) propose *compressed screening* to efficiently learn the activity of a set of compounds from pooled tests (Yao et al., 2023; Cleary & Regev, 2020). Weinstein et al. (2025) learn sequence-activity models using variational synthesis and randomized pooling. We optimize sequence-activity maps using designed mixtures of molecules in each pool.

LIDS builds on previous methods for optimization based on the EIG, such as entropy search (MacKay, 1992; Villemonteix et al., 2009; Hennig & Schuler, 2012; Hernández-Lobato et al., 2014). Russo & Van Roy (2018) show the EIG enables rapid optimization of sparse functions, where UCB and TS fail. To operationalize this idea, we apply recent advances in EIG approximation (Rainforth et al., 2024; Foster et al., 2020). Like the uncertainty autoencoder (UAE) (Grover & Ermon, 2019), we design experiments end-to-end by training an encoder and decoder, but our focus is on learning sparse functions $\mathrm{f}(x)$ rather than sparse Euclidean vectors. Like Sharrock (2026) we optimize distributions $q(x)$, but our goal is to design mixtures not ease computation.

5 Theory

We show information-dense synthesis can rapidly search large molecular spaces (Sec. 5.1), but is slowed by experimental limitations (Sec. 5.2)

5.1 Exponential speed ups

In principle, information-dense synthesis can find the optimal molecule with a single experiment.

Example 1 (No noise, discrete map)**.** *Assume activity is discrete, s.t.* $\mathrm{f} : \mathcal{X} \to \mathcal{Y}$ *for* $\mathcal{Y} = \{0, \ldots |\mathcal{Y}| - 1\}$. *Assume the model is well specified,* $\mathrm{f} \in \{\mathrm{f}_\theta : \theta \in \Theta\}$, *and* $\pi(\mathrm{f}_\theta = \mathrm{f}) > 0$, *we can synthesize any* $q \in \mathcal{Q} = \mathcal{P}(\mathcal{X})$, *and our assay has no noise,* $y_t = \langle \mathrm{f}, q \rangle$. *Then, by optimizing the EIG (Eq.* (5)*), we find* $x^\star$ *after one experiment.*

Proof sketch. Set $q(x) = \frac{1}{Z} \frac{1}{|\mathcal{Y}|^x}$ where $Z = \sum_{x=1}^{d} \frac{1}{|\mathcal{Y}|^x}$. Now $Zy = Z\langle \mathrm{f}, q \rangle = \sum_{x=1}^{d} \frac{\mathrm{f}(x)}{|\mathcal{Y}|^x}$, so $\mathrm{f}(x)$ is encoded in the digits of Zy, e.g. for $|\mathcal{Y}| = 10$ we have $Zy = 0.\mathrm{f}(1)\mathrm{f}(2)\ldots\mathrm{f}(d)$. □

Proof in App. H. We assumed discrete activity, e.g. y could be the number of fluorophores that attach to x, but $\mathcal{O}(1)$ optimization can also be achieved for continuous $\mathcal{Y}$ when f_θ is sparse (App. I). Next, instead of continuous readouts $\langle \mathrm{f}, q \rangle \in \mathbb{R}$, consider assays that provide just one bit.

Example 2 (Binary activity test)**.** *Assume* f *is 1-sparse,* $\mathrm{f}(x) \in \{w\mathbb{I}(x = \theta) : \theta \in \mathcal{X},\ w \in \mathbb{R}_+\}$, *we can synthesize any* $q \in \mathcal{Q} = \mathcal{P}(\mathcal{X})$, *and our model is well-specified, with uniform prior* $\pi(\theta) = 1/d$. *The assay reports where any molecule is active,* $y_t = \mathbb{I}[\langle \mathrm{f}, q \rangle > 0]$. *Then, by optimizing the EIG with batch size* E *(Eq.* (6)*), we find* $x^\star$ *after* $\lceil \log_2 d/E \rceil$ *batched experiments.*

Proof sketch. The EIG executes 2^E-ary search. Each q contains half the current possible $x^\star$ values. Observing y tells us whether $x^\star$ is in that half or the other, cutting the search space in half. □

Proof in App. J, with an additional example in App. K. The space of all small molecules is estimated at $d = \sim 10^{60}$ (Ruddigkeit et al., 2012). This suggests it could be searched with three 96-well plates, where testing compounds individually would require $\sim 10^{60}$ experiments (Proposition 1). We next discuss how reality intervenes.

5.2 Limitations and tradeoffs

Our analytic examples assumed no noise, full control over the molecular mixture, and restrictions on the molecule-activity map. In this section we examine the method's behavior as these assumptions are relaxed or changed. We first examine the information gained in a single experiment.

Proposition 2 (Information bound)**.** *Assume Gaussian noise, $y = \langle \mathrm{f}_\theta, q \rangle + \sigma\epsilon$, where $\epsilon \sim \mathrm{Normal}(0,1)$, and $\langle \mathrm{f}_\theta, \bar{q} \rangle \in \mathcal{Z}$ for $\mathcal{Z}$ a finite set. Then,*

$$\mathrm{EIG}_\theta(q) \leq \min \left\{ \mathcal{H}(\langle \mathrm{f}_\theta, \bar{q} \rangle),\ \frac{1}{2} \log \left(1 + \frac{c^2}{\sigma^2} \mathrm{Var}_\pi[\langle \mathrm{f}_\theta, \bar{q} \rangle] \right) \right\}. \tag{7}$$

The proof (App. L) applies the Shannon-Hartley theorem. This bound shows information is limited by an assay's *signal-to-noise ratio* $\mathrm{SNR}(q) \triangleq \frac{c^2}{\sigma^2} \mathrm{Var}_\pi[\langle \mathrm{f}_\theta, \bar{q} \rangle]$. High noise σ can reduce the EIG to zero. Noise is exacerbated by low synthesis yield c: if we do not have enough material to test, we gain very little information. Control over synthesis also limits information: we must achieve a high inner product between what we can make, q, and the molecule-activity map we want to learn, f_θ.

This bound also reveals a tradeoff that comes from using stochastic synthesis. We can typically increase the signal's entropy $\mathcal{H}(\langle \mathrm{f}_\theta, \bar{q} \rangle)$ by testing more spread out mixtures $\bar{q}(x)$. However, the SNR is maximized by concentrating on a single molecule, since $\exists \delta_x \in \mathrm{argmax}_{\bar{q}} \mathrm{Var}_\pi[\langle \mathrm{f}_\theta, \bar{q} \rangle]$, following Bauer's maximum principle (App. L.1).

Example 3. *Assume the molecule activity map is 1-sparse with maximum value one, $\mathrm{f}_\theta(x) \in \{\mathbb{I}(x = \theta) : \theta \in \mathcal{X}\}$, and the prior is uniform, $\pi(\theta) = 1/d$. Assume a synthesis model that creates uniform mixtures, $q_U(x) = \frac{1}{d_q} \mathbb{I}(x \in U)$ for $U \subseteq \mathcal{X}$, $\|q\|_1 = c = 1$ and $d_q \triangleq |U|$. Assume Gaussian noise $y = \langle \mathrm{f}_\theta, q \rangle + \sigma\epsilon$ where $\epsilon \sim \mathrm{Normal}(0,1)$. In this example, we can numerically solve the EIG, and compare with the bounds from Proposition 2, see Fig. 2. As the noise σ increases, the optimal tradeoff between entropy and SNR shifts to smaller values of d_q. See App. M for details.*

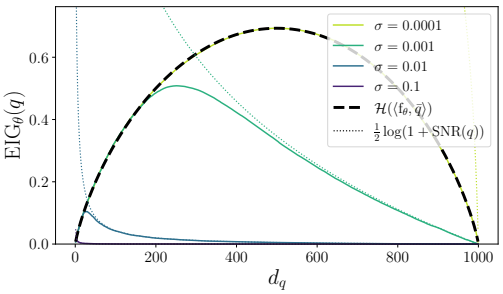
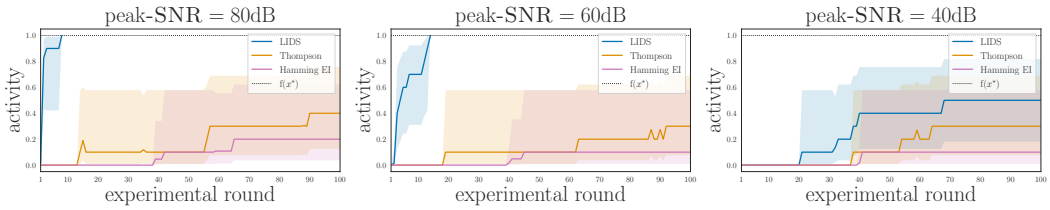


Figure 2: EIG tradeoffs. EIG from Example 3, plotted alongside MI upper bounds from Proposition 2, as a function of $d_q = |\mathrm{supp}(q)|$.

Synthesizing mixtures can offer major gains in early experiments, but the advantage decreases in the limit as the number of experiments goes to infinity.

Proposition 3. *Let $\mathrm{f}_\theta : \mathcal{X} \to \mathbb{R}_+$ be an arbitrary molecule-activity map, parameterized by θ. Consider maximizing the EIG for $\mathrm{f}_\theta = (\mathrm{f}_\theta(x))_{x \in \mathcal{X}}$. Let $\mathcal{M}_c^+(\mathcal{X}) = \{q = c\bar{q} : \bar{q} \in \mathcal{P}(\mathcal{X})\}$ be the set of mixtures and let $\mathcal{S}_c(\mathcal{X}) = \{q = c\delta_x : x \in \mathcal{X}\}$ be the set of delta mixtures (individual synthesis). Define $\sigma^2_{info}(\mathcal{Q}) \triangleq \lim_{t \to \infty} t \cdot \mathrm{Cov}[\mathrm{p}(\mathrm{f}_\theta \mid \mathcal{D}_t)]$ to be the asymptotic covariance of the posterior when we choose $q \in \mathcal{Q}$ at each step t by maximizing the EIG. Then, the asymptotic covariance is equal regardless of whether we test mixtures: $\sigma^2_{info}(\mathcal{S}_c(\mathcal{X})) = \sigma^2_{info}(\mathcal{M}_c^+(\mathcal{X}))$.*

The proof is in App. N; it applies the Bernstein-von Mises theorem for sequential Bayesian design from Paninski (Paninski, 2005). The overall picture is that at the beginning of learning, synthesizing mixtures can offer major advantages, but in the limit as we collect infinite data and approach complete knowledge of f, the advantages diminish.



(a) **(b)** **(c)**

Figure 3: Synthetic f with rare activity. LIDS, Thompson, and Hamming EI, at varying noise-level σ, optimizing a synthetic oracle. The activity of the predicted optimum, $\mathrm{f}(\hat{x}_t^\star)$, is plotted as a function of experimental round t, averaged over 10 duplicates. The 95% CI is calculated using a logit transform and the delta method, so that it stays bounded in the interval $[0, \mathrm{f}(x^\star)]$ (See App. P).

6 Empirical Results

We compare LIDS to individual synthesis empirically. We evaluate in simulations, where the ground truth molecule-activity map f is specified by different oracles, and measurements have Gaussian noise with known variance. We compare to molecular optimization methods that use individual synthesis. First, Thompson sampling using the same prior $\pi(\theta)$ and model f_θ as LIDS (*Thompson*) (Thompson, 1933). Then, Gaussian process (GP) Bayesian optimization with expected improvement (EI), using either the exponential Hamming kernel (*Hamming EI*) or the LOCK kernel (*LOCK EI*), which is SOTA for learning protein activity from limited data (Jankowiak et al., 2026; Amin et al., 2023; Groth et al., 2024). Finally, we ablate LIDS by removing stochasticity from its synthesis design q_t, replacing it with $\bar{q}_t(x) \leftarrow \mathrm{argmax}_{\delta_x} \langle \delta_x, \bar{q}_t \rangle$ (*LIDS-argmax*). To compare LIDS fairly to individual synthesis, we set the number of syntheses equal to the batch size $S = E$ and only vary the weights within a batch $\psi_{1:E}^t = (\alpha_1^t, \beta, c), ..., (\alpha_E^t, \beta, c)$, so the synthesis cost (amount of reagents, number of reaction steps) is the same. (Note when experiments are bottlenecked by assay cost rather than synthesis cost, LIDS can use $S \gg E$ to reach higher EIG.) To monitor performance, after every round t we evaluate the true activity of the predicted optimum, $\mathrm{f}(\hat{x}_t^\star)$. All experiments use a single Nvidia A100 GPU. LIDS is by far the most computationally intensive: EIG maximization takes $\approx$ 30min and posterior updates take $\approx$ 30min for the oracles in Sec. 6.2. Key results:

- On synthetic f with rare activity, LIDS finds the optimum in 2-5 experiments (median) where individual synthesis methods fail even after 100.
- On learned f from a protein language model, LIDS optimizes an order of magnitude faster.
- LIDS is robust to model misspecification and to noise in the assay or synthesis, with performance degrading gradually towards individual synthesis as noise increases.

6.1 Sparse synthetic oracle

We first evaluate LIDS on a small-scale problem where we specify a sparse f as an oracle.

Setup Let $\mathcal{X} = \{A, T, C, G\}^8$ be the set of all $d = |\mathcal{X}| = 65{,}536$ DNA sequences of length $L = 8$. We construct an oracle f where the optimal sequence $x^\star$ has activity one and all sequences more than one mutation away from $x^\star$ have activity zero. Measurement adds Gaussian noise with standard deviation $\sigma \in \{10^{-2}, 10^{-3}, 10^{-4}\}$. We fix the yield $c = 1$. This corresponds to *peak-signal-to-noise ratio* $\mathrm{PSNR} \triangleq 10 \log_{10}(\frac{c^2}{\sigma^2}) \in \{40, 60, 80\}$ dB. At every round, we test a batch of $E = 12$ designs. We run 10 independent repeats for each method. Details in App. O.

Results At low noise-levels (PSNR $= 80$ dB), LIDS identifies the optimum $x^\star$ in less than 8 rounds of experiments across all 10 repeats, with a median of just 2 rounds (Tab. 1 and fig. 3a) In most repeats, the individual synthesis methods (Thompson sampling and Hamming EI) failed to find the optimum within 100 rounds of experiments. As the assay noise increases, the advantage of LIDS diminishes gradually, approaching the individual synthesis methods (Figs. 3b and 3c). We visualized LIDS qualitatively by a low-dimensional projection of its designs q_t and beliefs $\mathrm{p}(x^\star \mid \mathcal{D}_t)$ (Fig. 7). At each round t, LIDS designs complex stochastic syntheses that cover its remaining uncertainty.

Table 1: Sparse oracle. Number of rounds until finding the optimal molecule, $\hat{x}^\star = x^\star$.

Noise level	Method	# of rounds ↓		
		min	median	max
Low	Thompson	14	>100	>100
	Hamming EI	42	>100	>100
	LIDS (ours)	**2**	**2**	**8**
Medium	Thompson	19	>100	>100
	Hamming EI	45	>100	>100
	LIDS (ours)	**3**	**4.5**	**14**
High	Thompson	38	>100	>100
	Hamming EI	41	>100	>100
	LIDS (ours)	**21**	>100	>100

Table 2: PLM oracle. Number of rounds until finding the optimal molecule, $\hat{x}^\star = x^\star$.

Protein	Method	# of rounds ↓		
		min	median	max
PETase	Thompson	>100	>100	>100
	Hamming EI	9	27	88
	LOCK EI	12	27	47
	LIDS-argmax	>5	>5	>5
	LIDS (ours)	**2**	**3**	>5
RuBisCO	Thompson	72	>100	>100
	Hamming EI	42	118	>200
	LOCK EI	81	>100	>100
	LIDS-argmax	>5	>5	>5
	LIDS (ours)	**2**	**2**	>5
SazCA	Thompson	>100	>100	>100
	Hamming EI	>100	>100	>100
	LOCK EI	>100	>100	>100
	LIDS-argmax	>5	>5	>5
	LIDS (ours)	>5	>5	>5

6.2 Protein language model oracle

We next evaluate LIDS on a peptide discovery problem, where the oracle molecule-activity map f is estimated from large scale data.

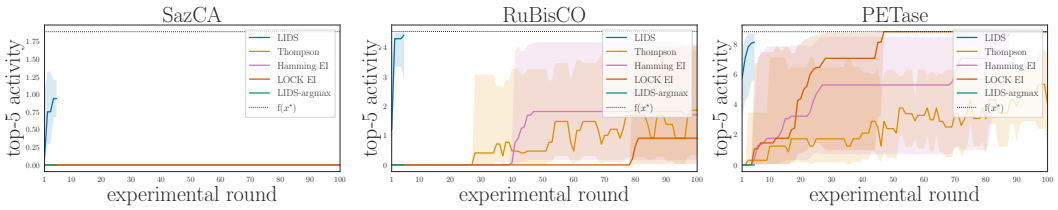
Setup Let $\mathcal{X} = \mathcal{B}^L$ be the set of peptides of length L, where $\mathcal{B} = \{A, C, ...\}$ is the 20 natural amino acids. As an oracle $\mathrm{f} : \mathcal{X} \to \mathbb{R}$, we use an estimate of evolutionary fitness based on a pre-trained protein language model (PLM),

$$\mathrm{f}(x) \triangleq \left(\log \frac{\mathrm{p}_{\mathrm{PLM}}(x \mid x^{wt}_{1:\tau})}{\mathrm{p}_{\mathrm{PLM}}(x^{wt}_{\tau+1:\tau+L} \mid x^{wt}_{1:\tau})} + 1 \right)_+ \tag{8}$$

where we set $\mathrm{p}_{\mathrm{PLM}}(x)$ to ProGen2's likelihood and x^{wt} to a wild-type protein sequence (Nijkamp et al., 2023). The likelihood ratio is the *evolutionary index*, which compares the probability of evolution generating x versus a natural sequence; it has been shown to reliably correlate with diverse measures of protein activity (Hopf et al., 2017; Riesselman et al., 2018). The conditional $\mathrm{p}_{\mathrm{PLM}}(x \mid x^{wt}_{1:\tau})$ is predictive of whether a peptide x will interact with $x^{wt}_{1:\tau}$ (Chen et al., 2026). Sequences below a threshold evolutionary index are typically dysfunctional or pathogenic, so we apply a cutoff (Frazer et al., 2021). Unlike benchmarks based on raw experimental data, $\mathrm{p}_{\mathrm{PLM}}(x)$ is defined over all of sequence space, allowing us to compute the activity of a mixture (Notin et al., 2023). Crucially, we can approximate the integral $\langle \mathrm{f}, q \rangle = \mathbb{E}_{x \sim q(x)}[\mathrm{f}(x)]$ even when active sequences are rare under f, by generating samples from the same PLM whose likelihood defines f. We use importance sampling with a defensive mixture as the proposal, $s(x) = \frac{1}{2}q(x) + \frac{1}{2}\mathrm{p}_{\mathrm{PLM}}(x \mid x^{wt}_{1:\tau})$, ensuring we evaluate $q(x)\mathrm{f}(x)$ in regions where either q or f are high (Hesterberg, 1995) (App. Q.1).

We consider three wild-type proteins x^{wt}: the CO_2-anhydrase SazCA, the CO_2-capturing RuBisCO, and the plastic-degrading PETase. We prompt with the first $\tau = 95$ amino acids of the wild-type, and design the next 5, so the molecular space $\mathcal{X} = \mathcal{B}^5$ is size $d = |\mathcal{X}| = 3{,}200{,}000$ molecules. The LIDS model is given an unbiased prior $\pi(\theta)$, with no evolutionary or physicochemical data that could be predictive of $\mathrm{p}_{\mathrm{PLM}}(x)$. LOCK uses BLOSUM, a model of amino acid correlations based on evolutionary data (Jankowiak et al., 2026; Henikoff & Henikoff, 1992). At every experimental round, we test a batch of $E = 96$ designs. We run 5 repeats, with independent initialization and measurement noise with PSNR $= 80$dB. We run the individual synthesis methods for 100 rounds and, due to computational expense, LIDS for 5. Details in App. Q

Results LIDS finds high-activity molecules faster than methods based on individual synthesis (Fig. 4). LIDS' improvement is smallest on the PETase landscape, where active sequences are more common: here LIDS finds the global optimum in 3 rounds (median across repeats), whereas the GP methods take 27 rounds (median) (Tab. 2). On the RuBisCO landscape, LIDS finds the global optimum in 2 rounds (median), whereas individual synthesis methods did not find it within



(a) **(b)** **(c)**

Figure 4: PLM oracle. LIDS, Thompson, Hamming EI, and LOCK EI, applied to a PLM oracle, with low noise (PSNR $=$ 80dB). The maximum activity of the 5 latest test molecules $\max \mathrm{f}(\hat{x}^\star_{t-5}), ..., \mathrm{f}(\hat{x}^\star_t)$, is plotted as a function of the experimental round t, averaged over 5 duplicates. The 95% CI is calculated using a logit transform and the delta method, so that it stays bounded in the interval $[0, \mathrm{f}(x^\star)]$ (See App. P).

100 rounds, in the majority of repeats (Tab. 2). We kept running Hamming EI for an additional 100 rounds; even so, it only found the optimum in 3/5 repeats. On the SazCA landscape, LIDS does not find the unique global optimum, but identifies sequences with non-zero activity within 5 rounds, while the alternatives find nothing after 100 rounds (Fig. 4a). When LIDS failed to find the global optimum, it appeared to be due to inefficient posterior sampling. Reducing the number of posterior samples M degraded LIDS' performance (Fig. 6).

LIDS' advantage is from information-dense stochastic synthesis. Ablating stochastic synthesis, by replacing each design q_t with its optimum, ablated performance entirely (*LIDS-argmax*). So did ablating stochastic synthesis by using Thompson sampling with the same model f_θ and prior $\pi(\theta)$ as LIDS (*Thompson*). Using a SOTA GP model, containing additional prior information about f, did not compensate for the advantages of stochastic synthesis (*LOCK*).

LIDS is robust: its model f_θ is misspecified for these oracle f. Adding synthesis errors, so that the tested q_t differs from the design, began reducing performance only when 40% of every mixture β_{sl} was corrupted (Fig. 5).

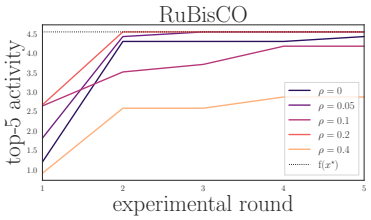


Figure 5: LIDS with synthesis errors. We add random errors to the designed mixtures q_t by replacing β_{sl} with $(1-\rho)\beta_{sl} + \rho\omega$ where $\omega \sim \mathrm{Dir}(\vec{1}_B)$.

7 Discussion

We showed that using information-dense stochastic synthesis, instead of synthesizing and testing individual molecules, can lead to massive speedups on hard molecular discovery problems.

Limitations. First, LIDS has only been evaluated in simulation, not in the wetlab. Second, LIDS shifts costs from experimentation to computation: it requires fewer wetlab experiments but much more compute than individual synthesis. The challenge is LIDS requires (a) calculating the predicted activity of a complex mixture rather than a single molecule, (b) optimizing the EIG, and (c) large numbers of posterior samples. Scaling LIDS to larger molecules and more complex surrogate models f_θ will require accelerating Bayesian experimental design, e.g. using amortization (Foster et al., 2020; 2021; Hedman et al., 2025). The mixture's activity $\langle \mathrm{f}_\theta, q \rangle$ could remain tractable if the surrogate f_θ is a smooth and decomposable probabilistic circuit (Vergari et al., 2021).

Outlook. Scaling molecular machine learning requires scaling laboratory feedback. One axis for scaling is performing more experiments. When testing individual molecules, this is essentially the only hope, since no adaptive algorithm can overcome a sparse activity map (Bach, 2020, Proposition 1). LIDS opens another, complementary approach to scaling: achieving precise control over synthesis, higher synthesis yields, and higher assay signal-to-noise ratios. These improvements do not change the number of datapoints, but rather the information they carry. This opens new opportunities and directions for scaling automated molecular science.

AI USE STATEMENT

In this work, we used generative AI tools to: check proofs, and assist in writing code. We have not used generative AI tools to: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, interpret results, assist in the writing of proofs, and implement methods. Propose or refine hypotheses, and clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, create or edit software code, sourcing/searching for information, and identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Full proofs are in the appendix. All code for reproducing simulations and the corresponding plots can be found in the Supplementary Material.

REFERENCES

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://papers.nips.cc/paper_files/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Milad Abolhasani. The lab that learns. *Science*, 393(6813):761–763, August 2026. doi: 10.1126/science.aee2448. URL https://www.science.org/doi/10.1126/science.aee2448.

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group Testing: An Information Theory Perspective. *Foundations and Trends® in Communications and Information Theory*, 23(1-2): 1–221, May 2026. ISSN 1567-2190, 1567-2328. doi: 10.1108/FTCIT-11-2025-0150. URL http://arxiv.org/abs/1902.06002. arXiv:1902.06002 [cs.IT].

Alan Nawzad Amin, Eli Nathan Weinstein, and Debora Susan Marks. Biological Sequence Kernels with Guaranteed Flexibility, April 2023. URL http://arxiv.org/abs/2304.03775. arXiv:2304.03775 [stat].

Michael Riis Andersen, Ole Winther, and Lars Kai Hansen. Bayesian Inference for Structured Spike and Slab Priors. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/8927f9b43dbc13ddfa1edf2103f2191f-Abstract.html.

George Kamal Atia and Venkatesh Saligrama. Boolean compressed sensing and noisy group testing. *arXiv [cs.IT]*, July 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256, May 2002. ISSN 1573-0565. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Francis Bach. Optimization is as hard as approximation. https://francisbach.com/optimization-is-as-hard-as-approximation/, 2020. Accessed: 2025-7-3.

David Barber and Felix Agakov. Information Maximization in Noisy Channels : A Variational Approach. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper/2003/hash/a6ea8471c120fe8cc35a2954c9b9c595-Abstract.html.

Heinz Bauer. Minimalstellen von Funktionen und Extremalpunkte. *Archiv der Mathematik*, 9(4): 389–393, November 1958. ISSN 1420-8938. doi: 10.1007/BF01898615. URL https://doi.org/10.1007/BF01898615.

S. L. Beaucage and M. H. Caruthers. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Letters*, 22(20):1859–1862, January 1981. ISSN 0040-4039. doi: 10.1016/S0040-4039(01)90461-7. URL https://www.sciencedirect.com/science/article/pii/S0040403901904617.

Michael Betancourt. Sparsity Blues, May 2021. URL https://betanalpha.github.io/assets/case_studies/modeling_sparsity.html.

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-Performance Large-Scale Image Recognition Without Normalization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1059–1071. PMLR, July 2021. URL https://proceedings.mlr.press/v139/brock21a.html.

Marvin H. Caruthers. The Chemical Synthesis of DNA/RNA: Our Gift to Science. *The Journal of Biological Chemistry*, 288(2):1420–1427, January 2013. ISSN 0021-9258. doi: 10.1074/jbc.X112.442855. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3543024/.

Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. Handling Sparsity via the Horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pp. 73–80. PMLR, April 2009. URL https://proceedings.mlr.press/v5/carvalho09a.html.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, Jianli Wu, Audrey Stavrand, Mayumi Schaepers-Cheu, Tian Zi Wang, Divya Srijay, Connor Monticello, Pranay Vure, Rishab Pulugurta, Sarah Pertsemlidis, Kseniia Kholina, Shrey Goel, Matthew P. DeLisa, Jen-Tsan Ashley Chi, Ray Truant, Hector C. Aguilar, and Pranam Chatterjee. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, 44(6):1002–1010, June 2026. ISSN 1546-1696. doi: 10.1038/s41587-025-02761-2. URL https://www.nature.com/articles/s41587-025-02761-2.

Brian Cleary and Aviv Regev. The necessity and power of random, under-sampled experiments in biology. *arXiv [q-bio.QM]*, December 2020.

Adam Foster, Martin Jankowiak, Matthew O'Meara, Yee Whye Teh, and Tom Rainforth. A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2959–2969. PMLR, June 2020. URL https://proceedings.mlr.press/v108/foster20a.html.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design, June 2021. URL http://arxiv.org/abs/2103.02438. arXiv:2103.02438 [stat].

Jonathan Frazer, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, November 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04043-8. URL https://www.nature.com/articles/s41586-021-04043-8.

Nathan C. Frey, Isidro Hötzel, Samuel D. Stanton, Ryan Kelly, Robert G. Alberstein, Emily K. Makowski, Karolis Martinkus, Daniel Berenberg, Jack Bevers, Tyler Bryson, Pamela Chan, Yongmei Chen, Alicja Czubaty, Tamica D'Souza, Henri Dwyer, Anna Dziewulska, James W. Fairman, Allen Goodman, Jennifer Hofmann, Henry Isaacson, Aya Ismail, Samantha James, Taylor Joren, Simon Kelow, James R. Kiefer, Matthieu Kirchmeyer, Joseph Kleinhenz, James T. Koerber, Julien Lafrance-Vanasse, Andrew Leaver-Fay, Jae Hyeon Lee, Edith Lee, Donald Lee, Wei-Ching Liang, Joshua Yao-Yu Lin, Sidney Lisanza, Andreas Loukas, Jan Ludwiczak, Sai Pooja Mahajan, Omar Mahmood, Homa Mohammadi-Peyhani, Santrupti Nerli, Ji Won Park, Jaewoo Park, Stephen Ra, Sarah Robinson, Saeed Saremi, Franziska Seeger, Imee Sinha, Anna M. Sokol, Natasa Tagasovska, Hao To, Edward Wagstaff, Amy Wang, Andrew M. Watkins, Blair Wilson, Shuang Wu, Karina Zadorozhny, John Marioni, Aviv Regev, Yan Wu, Kyunghyun Cho, Richard Bonneau, and Vladimir Gligorijević. Lab-in-the-loop therapeutic antibody design with deep learning, February 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.02.19.639050.

Peter Mø rch Groth, Mads Herbert Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29514–29565. Curran Associates, Inc., 2024. doi: 10.52202/079017-0929. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/34547650b2ca69d91f3b3c3ae8b21962-Abstract-Conference.html.

Aditya Grover and Stefano Ermon. Uncertainty Autoencoders: Learning Compressed Representations via Variational Information Maximization. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pp. 2514–2524. PMLR, April 2019. URL https://proceedings.mlr.press/v89/grover19a.html.

Taesin Ha, Dongseon Lee, Youngchun Kwon, Min Sik Park, Sangyoon Lee, Jaejun Jang, Byungkwon Choi, Hyunjeong Jeon, Jeonghun Kim, Hyundo Choi, Hyung-Tae Seo, Wonje Choi, Wooram Hong, Young Jin Park, Junwon Jang, Joonkee Cho, Bosung Kim, Hyukju Kwon, Gahee Kim, Won Seok Oh, Jin Woo Kim, Joonhyuk Choi, Minsik Min, Aram Jeon, Yongsik Jung, Eunji Kim, Hyosug Lee, and Youn-Suk Choi. AI-driven robotic chemist for autonomous synthesis of organic molecules. *Science Advances*, 9(44):eadj0461, November 2023. doi: 10.1126/sciadv.adj0461. URL https://www.science.org/doi/10.1126/sciadv.adj0461.

Bjarke Hastrup, Francois Cornet, Tejs Vegge, and Arghya Bhowmik. AtomComposer: Discovering Chemical Space from First Principles with Reinforcement Learning, May 2026. URL http://arxiv.org/abs/2605.28287. arXiv:2605.28287 [cs.LG].

Marcel Hedman, Desi R Ivanova, Cong Guan, and Tom Rainforth. Step-DAD: Semi-amortized policy-based bayesian experimental design. In *Forty-second International Conference on Machine Learning*, June 2025.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, 89(22):10915–10919, November 1992.

Philipp Hennig and Christian J. Schuler. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 13(null):1809–1837, June 2012. ISSN 1532-4435. URL https://dl.acm.org/doi/10.5555/2188385.2343701.

José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/6488484c982e9af5c35689523ba1abfe-Abstract.html.

José Miguel Hernández-Lobato, Michael Gelbart, Matthew Hoffman, Ryan Adams, and Zoubin Ghahramani. Predictive Entropy Search for Bayesian Optimization with Unknown Constraints. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1699–1707. PMLR, June 2015. URL https://proceedings.mlr.press/v37/hernandez-lobatob15.html.

Tim Hesterberg. Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics*, 37(2):185–194, May 1995. ISSN 0040-1706. doi: 10.1080/00401706.1995.10484303. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1995.10484303.

Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017. ISSN 1546-1696. doi: 10.1038/nbt.3769.

Richard A. Houghten, Clemencia Pinilla, Jon R. Appel, Sylvie E. Blondelle, Colette T. Dooley, Jutta Eichler, Adel Nefzi, and John M. Ostresh. Mixture-Based Synthetic Combinatorial Libraries. *Journal of Medicinal Chemistry*, 42(19):3743–3778, September 1999. ISSN 0022-2623. doi: 10.1021/jm990174v. URL https://doi.org/10.1021/jm990174v.

Martin Jankowiak, Yerdos Ordabayev, Rudraksh Tuwani, Henry N. Ward, Hunter Nisonoff, James M. McFarland, and Gevorg Grigoryan. Flexible Kernels for Protein Property Prediction, June 2026. URL http://arxiv.org/abs/2606.11057. arXiv:2606.11057 [cs.LG].

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, December 1998. ISSN 1573-2916. doi: 10.1023/A:1008306431147. URL https://doi.org/10.1023/A:1008306431147.

Raghunandan M Kainkaryam and Peter J Woolf. Pooling in high-throughput drug screening. *Curr. Opin. Drug Discov. Devel.*, 12(3):339–350, May 2009.

Nuo Liu, Walaa E. Kattan, Benjamin E. Mead, Conner Kummerlowe, Thomas Cheng, Sarah Ingabire, Jaime H. Cheah, Christian K. Soule, Anita Vrcic, Jane K. McIninch, Sergio Triana, Manuel Guzman, Tyler T. Dao, Joshua M. Peters, Kristen E. Lowder, Lorin Crawford, Ava P. Amini, Paul C. Blainey, William C. Hahn, Brian Cleary, Bryan Bryson, Peter S. Winter, Srivatsan Raghavan, and Alex K. Shalek. Scalable, compressed phenotypic screening using pooled perturbations. *Nature Biotechnology*, 43(8):1324–1336, August 2025a. ISSN 1546-1696. doi: 10.1038/s41587-024-02403-z. URL https://www.nature.com/articles/s41587-024-02403-z.

Yuanbin Liu, Ata Madanchi, Andy S. Anker, Lena Simine, and Volker L. Deringer. The amorphous state as a frontier in computational materials design. *Nature Reviews Materials*, 10(3):228–241, March 2025b. ISSN 2058-8437. doi: 10.1038/s41578-024-00754-2. URL https://www.nature.com/articles/s41578-024-00754-2.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. September 2018. URL https://openreview.net/forum?id=Bkg6RiCqY7.

David J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, July 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590. URL https://doi.org/10.1162/neco.1992.4.4.590.

Natalie Maus, Yimeng Zeng, Haydn Thomas Jones, Yining Huang, Gaurav Ng Goel, Alden Rose, Kyurae Kim, Hyun-Su Lee, Marcelo Der Torossian Torres, Fangping Wan, Cesar de la Fuente-Nunez, Mark Yatskar, Osbert Bastani, and Jacob R. Gardner. Purely Agent-Driven Black-Box Optimization for Biological Design, May 2026. URL http://arxiv.org/abs/2601.22382. arXiv:2601.22382 [cs.LG].

Qun Meng, Songhao Wang, and Szu Hui Ng. Combined Global and Local Search for Optimization with Gaussian Process Models, July 2021. URL http://arxiv.org/abs/2107.03217. arXiv:2107.03217 [stat.ML].

R. B. Merrifield. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *Journal of the American Chemical Society*, 85(14):2149–2154, July 1963. ISSN 0002-7863. doi: 10.1021/ja00897a025. URL https://doi.org/10.1021/ja00897a025.

Vladislav A. Mints, Jack K. Pedersen, Alexander Bagger, Jonathan Quinson, Andy S. Anker, Kirsten M. Ø. Jensen, Jan Rossmeisl, and Matthias Arenz. Exploring the Composition Space of High-Entropy Alloy Nanoparticles for the Electrocatalytic H_2 /CO Oxidation with Bayesian Optimization. *ACS Catalysis*, 12(18):11263–11271, September 2022. ISSN 2155-5435, 2155-5435. doi: 10.1021/acscatal.2c02563. URL https://pubs.acs.org/doi/10.1021/acscatal.2c02563.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, November 2023. ISSN 2405-4712, 2405-4720. doi: 10.1016/j.cels.2023.10.002. URL https://www.cell.com/cell-systems/abstract/S2405-4712(23)00272-7.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-Scale Benchmarks for Protein

Fitness Prediction and Design. *Advances in Neural Information Processing Systems*, 36:64331–64379, December 2023. doi: 10.52202/075280-2810. URL https://papers.nips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html.

Liam Paninski. Asymptotic Theory of Information-Theoretic Experimental Design. *Neural Computation*, 17(7):1480–1507, July 2005. ISSN 0899-7667, 1530-888X. doi: 10.1162/0899766053723032. URL https://direct.mit.edu/neco/article/17/7/1480-1507/6997.

Chiwoo Park. Jump Gaussian Process Model for Estimating Piecewise Continuous Regression Functions. *Journal of Machine Learning Research*, 23(278):1–37, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/21-1472.html.

Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214508000000337. URL https://www.tandfonline.com/doi/full/10.1198/016214508000000337.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro, December 2019. URL http://arxiv.org/abs/1912.11554. arXiv:1912.11554 [stat.ML].

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2), January 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1337SI. URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue-2/Sparsity-information-and-regularization-in-the-horseshoe-and-other-shrinkage/10.1214/17-EJS1337SI.full.

Noam Prywes, Naiya R. Phillips, Luke M. Oltrogge, Sebastian Lindner, Leah J. Taylor-Kearney, Yi-Chin Candace Tsai, Benoit de Pins, Aidan E. Cowan, Hana A. Chang, Renée Z. Wang, Laina N. Hall, Daniel Bellieny-Rabelo, Hunter M. Nisonoff, Rachel F. Weissman, Avi I. Flamholz, David Ding, Abhishek Y. Bhatt, Oliver Mueller-Cajar, Patrick M. Shih, Ron Milo, and David F. Savage. A map of the rubisco biochemical landscape. *Nature*, 638(8051):823–828, February 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08455-0. URL https://www.nature.com/articles/s41586-024-08455-0.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian Experimental Design. *Statistical Science*, 39(1):100–114, February 2024. ISSN 0883-4237, 2168-8745. doi: 10.1214/23-STS915. URL https://projecteuclid.org/journals/statistical-science/volume-39/issue-1/Modern-Bayesian-Experimental-Design/10.1214/23-STS915.full.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 814–822. PMLR, April 2014. URL https://proceedings.mlr.press/v33/ranganath14.html.

Jacob T. Rapp, Bennett J. Bremer, and Philip A. Romero. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1):97–107, January 2024. ISSN 2948-1198. doi: 10.1038/s44286-023-00002-4. URL https://www.nature.com/articles/s44286-023-00002-4.

Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, October 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0138-4. URL https://www.nature.com/articles/s41592-018-0138-4.

Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, December 2009. ISSN 1471-0080. doi: 10.1038/nrm2805. URL https://www.nature.com/articles/nrm2805.

Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3), January 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1215251110. URL https://pnas.org/doi/full/10.1073/pnas.1215251110.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, October 2012. ISSN 1549-9596. doi: 10.1021/ci300415d. URL https://doi.org/10.1021/ci300415d.

Daniel Russo and Benjamin Van Roy. Learning to Optimize via Information-Directed Sampling. *Operations Research*, 66(1):230–252, February 2018. ISSN 0030-364X. doi: 10.1287/opre.2017.1663. URL https://pubsonline.informs.org/doi/10.1287/opre.2017.1663.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1): 148–175, January 2016. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2015.2494218. URL https://ieeexplore.ieee.org/document/7352306/.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://ieeexplore.ieee.org/document/6773024.

Louis Sharrock. Wasserstein Gradient Flows for Batch Bayesian Optimal Experimental Design, 2026. URL https://arxiv.org/abs/2603.12102. Version Number: 1.

Andrew D. Skora, Jacqueline Douglass, Michael S. Hwang, Ada J. Tam, Richard L. Blosser, Sandra B. Gabelli, Jianhong Cao, Luis A. Diaz, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Shibin Zhou. Generation of MANAbodies specific to HLA-restricted epitopes encoded by somatically mutated genes. *Proceedings of the National Academy of Sciences*, 112 (32):9967–9972, August 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1511996112. URL https://pnas.org/doi/full/10.1073/pnas.1511996112.

Jack W Szostak. Introduction: Combinatorial chemistry. *Chem. Rev.*, 97(2):347–348, April 1997.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. doi: 10.2307/2332286. URL http://www.jstor.org/stable/2332286.

Mario A. Torres-Acosta, Gary J. Lye, and Duygu Dikicioglu. Automated liquid-handling operations for robust, resilient, and efficient bio-based laboratory practices. *Biochemical Engineering Journal*, 188:108713, December 2022. ISSN 1369-703X. doi: 10.1016/j.bej.2022.108713. URL https://www.sciencedirect.com/science/article/pii/S1369703X22003825.

Neil Vargesson. Thalidomide-induced teratogenesis: History and mechanisms. *Birth Defects Research Part C: Embryo Today: Reviews*, 105(2):140–156, 2015. ISSN 1542-9768. doi: 10.1002/bdrc.21096. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bdrc.21096. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdrc.21096.

Antonio Vergari, YooJung Choi, Anji Liu, Stefano Teso, and Guy Van den Broeck. A compositional atlas of tractable circuit operations for probabilistic inference. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pp. 13189–13201, Red Hook, NY, USA, December 2021. Curran Associates Inc. ISBN 978-1-7138-4539-3.

Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, August 2009. ISSN 1573-2916. doi: 10.1007/s10898-008-9354-2. URL https://doi.org/10.1007/s10898-008-9354-2.

Eli N. Weinstein, Alan N. Amin, Will S. Grathwohl, Daniel Kassler, Jean Disset, and Debora Marks. Optimal Design of Stochastic DNA Synthesis Protocols based on Generative Sequence Models. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7450–7482. PMLR, May 2022. URL https://proceedings.mlr.press/v151/weinstein22a.html.

Eli N Weinstein, Andrei Slabodkin, Mattia G Gollub, Kerry Dobbs, Xiao-Bing Cui, Fang Zhang, Kristina Gurung, and Elizabeth B Wood. Lifting biomolecular data acquisition. *arXiv [q-bio.BM]*, December 2025.

Eli N. Weinstein, Mattia G. Gollub, Andrei Slabodkin, Kerry Dobbs, Xiao-Bing Cui, Cameron L. Gardner, Ryan J. Grant, Kristina Gurung, Amira Bailey, Alan N. Amin, George M. Church, and Elizabeth B. Wood. Manufacturing-aware generative models enable petascale synthesis of designed DNA. *Nature Biotechnology*, pp. 1–9, March 2026. ISSN 1546-1696. doi: 10.1038/s41587-026-03020-8. URL https://www.nature.com/articles/s41587-026-03020-8.

David Wingate and Theophane Weber. Automated Variational Inference in Probabilistic Programming, January 2013. URL http://arxiv.org/abs/1301.1299. arXiv:1301.1299 [stat.ML].

Douglas Yao, Loic Binan, Jon Bezney, Brooke Simonton, Jahanara Freedman, Chris J Frangieh, Kushal Dey, Kathryn Geiger-Schuller, Basak Eraslan, Alexander Gusev, Aviv Regev, and Brian Cleary. Scalable genetic screening for regulatory circuits using compressed perturb-seq. *Nat. Biotechnol.*, October 2023.

Ronald N. Zuckermann, Janice M. Kerr, Stephen B. H. Kent, and Walter H. Moos. Efficient method for the preparation of peptoids [oligo(N-substituted glycines)] by submonomer solid-phase synthesis. *Journal of the American Chemical Society*, 114(26):10646–10647, December 1992. ISSN 0002-7863. doi: 10.1021/ja00052a076. URL https://doi.org/10.1021/ja00052a076.

A Notation

Notation	Description
$\mathcal{X}$	The molecular space, usually a Hamming space
$\mathcal{P}(\mathcal{X})$	$= \{\mathrm{p}(\cdot) : \mathrm{p}(x) \geq 0, \int_{\mathcal{X}} \mathrm{p}(x)\mathrm{d}x = 1\}$ The set of probability distributions on $\mathcal{X}$
$\mathrm{f} : \mathcal{X} \to \mathbb{R}_+$	The molecule-activity map
$q : \mathcal{X} \to \mathbb{R}_+$	The design - can be a mixture or a delta. The restriction on q comes from the synthesis model.
$\bar{q} = \frac{q}{c} \in \mathcal{P}(\mathcal{X})$	The normalized design, where c is the yield, the total number of molecules produced by a synthesis.
$c \in \mathbb{R}_+$	The chemical yield. The total number of molecules produced by the synthesis model.
$\mathrm{p}(y \mid \langle \mathrm{f}, q \rangle)$	The assay model
$\mathbb{I}[\cdot]$	The indicator function, which takes value 1 when the input is true, and 0 otherwise.
$\mathrm{Uniform}(z)$	A uniform distribution over a set z
$[d] \triangleq \{1, \ldots, d\}$	A set of d elements
$\Delta_{B-1} = \{v \in \mathbb{R}_+^B : \sum_j v_j = 1\}$	The simplex.
$\mathrm{LSE}_{k=1}^K[x_k] = \log \sum_{k=1}^K \exp[x_k]$	The 'logsumexp' operation.

B Proof of Proposition 1

Assume $\mathrm{f}(x) = \mathbb{I}(x = x^\star)$ where $x^\star \sim \mathrm{Uniform}(1, \ldots, d)$, and there is no noise in the assay, so we measure $y_t = \mathrm{f}(x_t)$. Define expected regret, $R \triangleq \mathbb{E}_{x^\star}[\mathbb{E}_\pi[\sum_{t=1}^\infty \mathbb{I}(X_t \neq X^\star)]]$. Let $\pi_0(x_t \mid \mathcal{D}_{t-1})$

denote a policy that selects the next candidate uniformly at random from among the remaining candidates until the hit is discovered.

$$\pi_0(x_t \mid \mathcal{D}_{t-1}) = \begin{cases} x_t \sim \text{Uniform}(\{x \notin \mathcal{D}_{t-1}\}) & \text{if } x^\star \notin \mathcal{D}_{t-1} \\ x_t = x^\star & \text{otherwise} \end{cases} \tag{9}$$

This policy has expected regret $(d-1)/2$. There does not exist a policy π with lower expected regret.

Proof. We first calculate the regret of the randomized policy π_0. We have

$$R_0 = \mathbb{E}_{x^\star}[\mathbb{E}_\pi[\sum_{t=1}^{\infty} \mathbb{I}(X_t \neq X^\star)]] \tag{10}$$

$$= \sum_{t=1}^{\infty} \mathrm{p}(x_t \neq x^\star) \tag{11}$$

$$= \mathrm{p}(x_1 \neq x^\star) + \mathrm{p}(x_1 \neq x^\star)\mathrm{p}(x_2 \neq x^\star \mid x_1 \neq x^\star) + \mathrm{p}(x_{1:2} \neq x^\star)\mathrm{p}(x_3 \neq x^\star \mid x_{1:2} \neq x^\star) + \cdots \tag{12}$$

$$= \frac{d-1}{d} + \frac{d-1}{d}\frac{d-2}{d-1} + \frac{d-2}{d}\frac{d-3}{d-2} + \cdots \tag{13}$$

$$= \frac{d-1}{d} + \frac{d-2}{d} + \cdots + \frac{1}{d} \tag{14}$$

$$= \frac{d-1}{2} \tag{15}$$

Next consider an arbitrary policy π. We will show that it cannot achieve lower regret than the randomized policy. We assume without loss of generality that π plays $x_t = x^\star$ once $x^\star \in \mathcal{D}$, since if it does not, its regret will only be improved by substituting this rule. Now start with the regret at $t = 1$. We have

$$\mathbb{E}_{x^\star}[\mathbb{E}_\pi[\mathbb{I}[X_1 \neq X^\star]]] = \frac{1}{d}\sum_{x^\star=1}^{d} \pi(x_1 \neq x^\star) = \frac{1}{d}\sum_{x^\star=1}^{d}[1 - \pi(x_1 = x^\star)] = \frac{1}{d}(d-1) \tag{16}$$

Next consider the regret at $t = 2$. We have

$$\mathbb{E}_{x^\star}[\mathbb{E}_\pi[\mathbb{I}[X_{1:2} \neq X^\star]]] = \mathrm{p}(x_1 \neq x^\star)\mathbb{E}_{\mathrm{p}(x_1|x_1\neq x^\star)\mathrm{p}(x^\star|x_1\neq x^\star,x_1)\mathrm{p}(x_2|x_1,x^\star,x_1\neq x^\star)}[\mathbb{I}(x_2 \neq x^\star)] \tag{17}$$

$$= \mathrm{p}(x_1 \neq x^\star)\mathbb{E}_{\mathrm{p}(x_1|x_1\neq x^\star)\mathrm{p}(x^\star|x_1\neq x^\star,x_1)}[\pi(x_2 \neq x^\star \mid x_1, y_1 = 0)] \tag{18}$$

$$= \mathrm{p}(x_1 \neq x^\star)\mathbb{E}_{\mathrm{p}(x_1|x_1\neq x^\star)}[\frac{1}{d-1}\sum_{x^\star \neq x_1} \pi(x_2 \neq x^\star \mid x_1, y_1 = 0)] \tag{19}$$

$$= \mathrm{p}(x_1 \neq x^\star)\mathbb{E}_{\mathrm{p}(x_1|x_1\neq x^\star)}[1 - \frac{1}{d-1}\sum_{x^\star \neq x_1} \pi(x_2 = x^\star \mid x_1, y_1 = 0)] \tag{20}$$

$$\geq \mathrm{p}(x_1 \neq x^\star)\mathbb{E}_{\mathrm{p}(x_1|x_1\neq x^\star)}[1 - \frac{1}{d-1}] \tag{21}$$

$$= \frac{d-1}{d}\frac{d-2}{d-1} = \frac{d-2}{d} \tag{22}$$

Where for the second line we have used the fact that the distribution of x_2 just depends on the policy, which just depends on x_1, y_1 and not on the unobserved $x^\star$. The inequality is strict when π does not assign any probability to the previous observation x_1 when $y_1 = 0$.

Repeating this argument for arbitrary t,

$$\mathbb{E}_{x^\star}[\mathbb{E}_\pi[\mathbb{I}[X_{1:t} \neq X^\star]]] = \mathrm{p}(x_{1:t-1} \neq x^\star)\mathbb{E}_{\mathrm{p}(x_{1:t-1}|x_{1:t-1}\neq x^\star)\mathrm{p}(x^\star|x_{1:t-1}\neq x^\star, x_{1:t-1})}[\pi(x_t \neq x^\star \mid x_{1:t-1}, y_{1:t-1} = 0)] \tag{23}$$

$$= \mathrm{p}(x_{1:t-1} \neq x^\star)\mathbb{E}_{\mathrm{p}(x_{1:t-1}|x_{1:t-1}\neq x^\star)} \frac{1}{d-t+1} \sum_{x^\star \neq x_{1:t-1}} \pi(x_t \neq x^\star \mid x_{1:t-1}, y_{1:t-1} = 0) \tag{24}$$

$$\geq \mathrm{p}(x_{1:t-1} \neq x^\star) \frac{d-t}{d-t+1} \tag{25}$$

$$= \frac{d-t}{d} \tag{26}$$

where the final step is from induction over t.

Summing over all t, we obtain the result.

□

C Popular acquisition functions are not useful to optimize mixtures

Most acquisition functions are useless when it comes to designing mixtures. The proof for the Upper Confidence Bound and Expected Improvement acquisitions functions are given below, which hinge on the fact that while $\mathrm{f} : \mathcal{X} \to \mathbb{R}$ is an arbitrary bounded function, in a lifted setting $\langle \mathrm{f}, q \rangle$ is an affine function over the convex set $q \in \mathcal{P}(\mathcal{X})$.

Define the set of delta functions as $\mathcal{S} = \{\delta_x : x \in \mathcal{X}\} \subset \mathcal{P}(\mathcal{X})$

Upper Confidence Bound The UCB acquisition function for a mixture is defined as (Auer et al., 2002; Abbasi-yadkori et al., 2011)

$$\mathrm{UCB}(q) \triangleq \sup_{\mathrm{f} \in \mathcal{F}_{t-1}} \langle \mathrm{f}, q \rangle, \tag{27}$$

where $\mathcal{F}_{t-1}$ is a confidence set of possible f, conditioned on the data at timepoint $t-1$. Note when $q \in \mathcal{S}$ this reduces to $\mathrm{UCB}(x) = \sup_{\mathrm{f} \in \mathcal{F}_{t-1}} \mathrm{f}(x)$. As this is the pointwise supremum over linear functions in q, $\mathrm{UCB(q)}$ is convex in q. The next mixture is chosen by $q_t = \mathrm{argmax}_{q \in \mathcal{P}(\mathcal{X})} \mathrm{UCB}(q)$. Since $\mathcal{S} \subset \mathcal{P}(\mathcal{X})$ we have our first inequality $\max_{q \in \mathcal{S}} \mathrm{UCB}(q) \leq \max_{q \in \mathcal{P}(\mathcal{X})} \mathrm{UCB}(q)$. A second inequality follows by Jensen's inequality:

$$\mathrm{UCB}\left(\int_{\mathcal{X}} q(x) \cdot \delta_x \mathrm{d}x\right) \leq \int_{\mathcal{X}} q(x) \mathrm{UCB}(\delta_x) \mathrm{d}x \tag{28}$$

$$\leq \max_{x \in \mathcal{X}} \mathrm{UCB}(\delta_x) = \max_{q \in \mathcal{S}} \mathrm{UCB}(q) \tag{29}$$

This establishes that $\max_{q \in \mathcal{S}} \mathrm{UCB}(q) = \max_{q \in \mathcal{P}(\mathcal{X})} \mathrm{UCB}(q)$, and as such there must always exist a $q^\star \in \mathcal{S}$ which is a maximizer of $\mathrm{UCB}(q)$

Expected Improvement Let $y = \langle \mathrm{f}, q \rangle$ denote the experimental outcome, and $y^\star = \max_{i<t} y_i$ denote the maximum experimental outcome so far. The improvement is defined as $I(q) \triangleq \max(0, \langle \mathrm{f}, q \rangle - y^\star)$, and the expected improvement is the expectation over the randomness in f,

$$\mathrm{EI}(q) = \mathbb{E}_{\mathrm{p}(\mathrm{f}|\mathcal{D}_{t-1})}[\max(0, \langle \mathrm{f}, q \rangle - y^\star)]. \tag{30}$$

Inside the expectation is the max of two affine functions, $\max(0, \langle \mathrm{f}, q \rangle - y^\star)$. Since the max of two affine functions is a convex function, $\mathrm{EI}(q)$ is an expectation over terms that are convex in q, which

is itself, a convex function. It then follows from the same logic as for the UCB acquisition function, that $\max_{q\in\mathcal{S}} \mathrm{EI}(q) \leq \max_{q\in\mathcal{P}(\mathcal{X})} \mathrm{EI}(q)$, and

$$\mathrm{EI}\left(\int_{\mathcal{X}} q(x)\cdot\delta_x \mathrm{d}x\right) \leq \int_{\mathcal{X}} q(x)\mathrm{EI}(\delta_x)\mathrm{d}x \tag{31}$$

$$\leq \max_{x\in\mathcal{X}} \mathrm{EI}(\delta_x) = \max_{q\in\mathcal{S}} \mathrm{EI}(q), \tag{32}$$

and as such there must always exist a $q^\star \in \mathcal{S}$ which is a maximizer of $\mathrm{EI}(q)$

D Derivation of variational lower bound on EIG

We can construct a lower variational bound on the mutual information, $\mathcal{I}_{BA}$, where r_ϕ denotes a variational distribution with parameters ϕ, and $KL(\mathrm{p}\|r) \geq 0$ denotes the KL divergence between distributions p and r (Barber & Agakov, 2003). To rewrite the expectation over y_t, we use the fact that y_t is a deterministic function of q_t, ϵ_t and θ, using $y_t = \langle \mathrm{f}_\theta, q_t\rangle + \sigma\epsilon_t$ for some noise level σ and $\epsilon_t \sim \mathrm{Normal}(0,1)$.

$$EIG_{x^\star}(q_t) = \mathcal{H}(\mathrm{p}(x^\star_\theta \mid \mathcal{D}_{t-1})) + \mathbb{E}_{\mathrm{p}(\theta|\mathcal{D}_{t-1})\mathrm{p}(y_t|q_t,\theta)}[\log \mathrm{p}(x^\star_\theta \mid \{y_t, q_t\})] \tag{33}$$

$$= \mathcal{H}(\mathrm{p}(x^\star_\theta \mid \mathcal{D}_{t-1})) + \mathbb{E}_{\mathrm{p}(\theta|\mathcal{D}_{t-1})\mathrm{p}(y_t|q_t,\theta)}[\log r_\phi(x^\star_\theta \mid y_t) + KL(p(\cdot \mid \{y_t, q_t\})\|r_\phi(\cdot \mid y_t))] \tag{34}$$

$$\geq \mathcal{H}(\mathrm{p}(x^\star_\theta \mid \mathcal{D}_{t-1})) + \mathbb{E}_{\mathrm{p}(\theta|\mathcal{D}_{t-1})\mathrm{p}(y_t|q_t,\theta)}[\log r_\phi(x^\star_\theta \mid y_t)] \tag{35}$$

$$= \mathcal{H}(\mathrm{p}(x^\star_\theta \mid \mathcal{D}_{t-1})) + \mathbb{E}_{\mathrm{p}(\theta|\mathcal{D}_{t-1})\mathrm{p}(\epsilon_t)}[\log r_\phi(x^\star_\theta \mid \langle \mathrm{f}_\theta, q_t\rangle + \sigma\epsilon_t)] \tag{36}$$

$$\triangleq \mathcal{I}_{BA}(q_t, \phi) \tag{37}$$

E Details on variational distribution r_ϕ

E.1 Architecture

The variational distribution approximates the conditional $\mathrm{p}(x^\star \mid y, q) \approx r_\phi(x^\star \mid y)$. As the design $q = q_\psi$ is fixed for any step of training, we don't include it as an input to the variational distribution. The effect of ψ on the EIG enters through the likelihood $\mathrm{p}(y \mid \langle \mathrm{f}_\theta, q_\psi\rangle)$. For a batch of experimental outcomes $\bar{y} = [y_t, ...y_{t+E}]$ we parameterize the variational distribution $r_\phi : \mathbb{R}^E \to \mathcal{P}(\mathcal{X})$ as a linear regression from $\bar{y}$ to a position-wise independent categorical distribution $\mathrm{Cat}(x \mid \beta) = \prod_{j=1}^L \beta_{x_j}$, by

$$r_{\phi=(A,C)}(x^\star \mid \bar{y}) = \mathrm{Cat}(x^\star \mid \mathrm{softmax}(A\cdot\bar{y} + C)), \tag{38}$$

where $\mathrm{Cat}(x \mid \beta) = \prod_{j=1}^L \beta_{x_j}$ is a position-wise independent categorical distribution, $A \in \mathbb{R}^{L\times B\times E}$ is a weight tensor, and $C \in \mathbb{R}^{L\times B}$ is a bias. The $\mathrm{softmax}$ is applied over the last dimension, which has size B.

E.2 Training

For Gaussian noise, using the reparameterization trick to reduce variance (Foster et al., 2020), we obtain the gradient

$$\nabla_{\psi,\phi}\mathcal{I}_{BA}(x^\star;\psi,\phi) \approx \frac{1}{M}\sum_{j=1}^M \nabla_{\psi,\phi}\log r_\phi(x^\star_{\theta_j} \mid \langle \mathrm{f}_{\theta_j}, q_\psi\rangle + \sigma\epsilon_j) \tag{39}$$

where $\epsilon_{1:M} \overset{iid}{\sim} \mathrm{Normal}(0,1)$.

The variational distribution parameters ϕ, as well as mixture parameters ψ, are optimized jointly using stochastic gradient ascent on $\nabla_{\psi,\phi}\mathcal{I}_{BA}(x^\star;\psi,\phi)$, using the AdamW optimizer with a learning

rate of 0.01, momentum of 0.9, and weight decay of 10^{-5} (Loshchilov & Hutter, 2018). We train all our models on a dataset of 100000 samples of θ, either drawn from the prior $\pi(\theta)$ (see App. F.1), or from the approximate posterior, $\mathrm{p}(\theta \mid \mathcal{D})$ (see App. G). We use 10000 linear warmup steps, and train for either 1000 epochs (for the sparse oracle, Sec. 6.1), or 5000 epoch (for the PLM oracle, Sec. 6.2). Training is done on an A100 Nvidia GPU.

F Details on Bayesian neural network f_θ

We consider a Bayesian neural network $\mathrm{f}_{\theta=(W,V)}(x) = \sum_{n=1}^{N} W_n \exp(V_n \cdot x)$ as a function class for sparse molecule-activity maps.

F.1 Sparsity-inducing prior

We define $w \cdot u_n \triangleq W_n$, and $\log v_n \triangleq V_n$. We restrict the domains of u and v to be $u \in \mathcal{U} = \{u \in [0,1]^N : \|u\|_\infty = 1\}$ and $v \in \mathcal{V} = \{v \in [0,1]^{N \times L \times B} : \|v_{n,l,\cdot}\|_\infty = 1 \; \forall \, n, l\}$. We can then construct a prior, $\pi(\theta) = \pi(w, u, v)$ that induces sparsity in the filters $V_1, ...V_N$ by setting

$$v'_{nlb} \sim \mathrm{Expon}(\frac{1}{2}) \tag{40}$$

$$v_{nlb} = \exp(v'_{nlb} - \max_b v'_{nlb}) \tag{41}$$

and

$$\log u'_n \sim \mathcal{N}(0,1) \tag{42}$$

$$u_n = \exp(\log u'_n - \max_n(\log u'_n)) \tag{43}$$

Where $\mathrm{Expon}(\gamma)$ denotes the exponential distribution with rate parameter γ. This choice is motivated by the use of the Laplace distribution for Bayesian LASSO (Park & Casella, 2008). It carries the belief that only few v_{nlb} are relevant. There exists extensive literature on sparsity-inducing priors (Carvalho et al., 2009; Piironen & Vehtari, 2017; Andersen et al., 2014; Betancourt, 2021). We found the process above to be robust to posterior approximation by NUTS, whereas many other priors led to difficult posterior geometry, a significant amount of divergences, and high Gelman-Rubin statistics. The function $h(x_n) = \exp(x'_n - \max_n x'_n)$ is used to bring u and v to their restricted domains, $u \in \mathcal{U}$ and $v \in \mathcal{V}$. Our prior on w carries our belief about the maximum activity that the assay can record, $\|\mathrm{f}\|_\infty \geq w$, with equality if the filters are disjoint $\prod_l \langle v_{nl}, v_{n'l} \rangle = 0 \forall n \neq n'$. Additionally, when the filters are disjoint, at least one $x^\star$ will have activity $\mathrm{f}_\theta(x^\star) = w$, and all other $x' \neq x^\star$ will have lower activity $\mathrm{f}_\theta(x') \leq w$. This makes it easy to incorporate prior information about the maximum activity, by setting $\pi(\|\mathrm{f}\|_\infty) = \pi(w)$. In the our simulations we use $\mathrm{p}(w) = \delta_{w=1}$ (for Sec. 6.1) and $\mathrm{p}(w) = 1 + \mathrm{Gamma}(1,1)$ (for Sec. 6.2), where $\mathrm{Gamma}(\alpha, \beta)$ is the gamma distribution with shape α and rate β.

F.2 Regularization

We add a regularization term to the joint distribution, to prevent mode collapse of the different bump functions. Additionally, this encourages f_θ for which the argmax approximation is tighter, since it encourages the filters $V_1, ...V_N$ to be disjoint. This term is defined as

$$\log p'(\theta, \mathcal{D}) = \log p(\theta, \mathcal{D}) - \lambda \cdot \frac{1}{B^L} \sum_{n<m} \prod_{l=1}^{L} \langle v_{n,l}, v_{m,l} \rangle, \tag{44}$$

where the normalization is to bring the regularization term to the interval $[0, N_{pairs}]$, and λ is a hyperparameter that controls the strength of the regularization. $N_{pairs} = \frac{N(N-1)}{2}$ is the number of pairs of bump functions.

F.3 INNER PRODUCT $\langle \mathrm{f}_\theta, q_\psi \rangle$

The inner product can be computed by

$$\langle \mathrm{f}_\theta, q_\psi \rangle = c \sum_{s=1}^{S} \alpha_s \sum_{n=1}^{N} W_n \prod_{l=1}^{L} \langle \exp V_{nl}, \beta_{sl} \rangle \tag{45}$$

The synthesis model is defined as $\bar{q}_{\psi=(\alpha,\beta)}(x) = \sum_{s=1}^{S} \alpha_s \prod_{l=1}^{L} \langle \beta_{sl}, x_l \rangle$. To see how the inner product $\langle \mathrm{f}_\theta, \bar{q}_\psi \rangle$ can be computed in closed form, first rewrite the network as

$$\mathrm{f}_\theta(x) = w \cdot \sum_{n=1}^{N} u_n \exp \sum_{l=1}^{L} \sum_{b=1}^{B} \log v_{nlb} \cdot x_{lb} \tag{46}$$

$$= w \cdot \sum_{n=1}^{N} u_n \exp \sum_{l=1}^{L} \log \sum_{b=1}^{B} v_{nlb} \cdot x_{lb} \tag{47}$$

$$= w \cdot \sum_{n=1}^{N} u_n \prod_{l=1}^{L} \sum_{b=1}^{B} v_{nlb} \cdot x_{lb} \tag{48}$$

$$= w \cdot \sum_{n=1}^{N} u_n \prod_{l=1}^{L} \langle v_{nl} \cdot x_l \rangle. \tag{49}$$

The inner product can then be derived as

$$\langle \mathrm{f}_\theta, \bar{q}_\psi \rangle = \sum_{x \in \mathcal{X}} \mathrm{f}_\theta(x) q_\psi(x) \tag{50}$$

$$= \sum_{x \in \mathcal{X}} \left(w \cdot \sum_{n=1}^{N} u_n \prod_{l=1}^{L} \langle v_{nl}, x_l \rangle \right) \cdot \left(\sum_{s=1}^{S} \alpha_s \prod_{l=1}^{L} \langle \beta_{sl}, x_l \rangle \right) \tag{51}$$

$$= \sum_{s=1}^{S} \alpha_s w \sum_{n=1}^{N} u_n \sum_{x \in \mathcal{X}} \prod_{l=1}^{L} \langle \beta_{sl}, x_l \rangle \langle v_{nl}, x_l \rangle \tag{52}$$

$$= \sum_{s=1}^{S} \alpha_s w \sum_{n=1}^{N} u_n \prod_{l=1}^{L} \langle v_{nl}, \beta_{sl} \rangle \tag{53}$$

$$\tag{54}$$

In practice, for numerical stability, we compute the functional inner product in log-space by

$$\log \langle \mathrm{f}_\theta, \bar{q}_\psi \rangle = \operatorname*{LSE}_{s=1}^{S} \left[\log \alpha_s + \log w + \operatorname*{LSE}_{N} \left[\log u_n + \sum_{B} \operatorname*{LSE}_{B} \left[\log v_{nlb} + \log \beta_{slb} \right] \right] \right], \tag{55}$$

where $\mathrm{LSE}_{k=1}^{K}[x_k] = \log \sum_{k=1}^{K} \exp[x_k]$ is the 'logsumexp' operation

F.4 APPROXIMATION OF $x^\star$

This function class has a tractable approximation of $x^\star = \mathrm{argmax}_x \mathrm{f}(x)$, or equivalently, $\mathrm{argmax} x \log \mathrm{f}_\theta(x)$. First, observe that we can write

$$\log \mathrm{f}_\theta(x) = \log \sum_{n} \exp(\log w_n + \sum_{\ell} \log v_{n,\ell}^\top x_\ell) \tag{56}$$

$$= \operatorname*{LSE}_{n=1}^{N} \left[\log w_n + \sum_{\ell} \log v_{n,\ell}^\top x_\ell \right]. \tag{57}$$

The LSE operation can be bounded by

$$\max_n \log w_n + \sum_\ell \log v_{n,\ell}^\top x_\ell \leq \operatorname*{LSE}_{n=1}^N \left[\log w_n + \sum_\ell \log v_{n,\ell}^\top x_\ell \right] \leq \max_n \log w_n + \sum_\ell \log v_{n,\ell}^\top x_\ell + \log N. \tag{58}$$

To approximate the maximum over $\mathrm{f}_\theta(x)$, we can maximize the lower bound. We have

$$\hat{x} = \operatorname*{argmax}_x \max_n \log w_n + \sum_\ell \log v_{n,\ell}^\top x_\ell \tag{59}$$

This is tractable. Define

$$\hat{x}_\ell^n = \operatorname*{argmax}_b v_{n,\ell} \tag{60}$$

$$\hat{n} = \operatorname*{argmax}_n \log w_n + \sum_\ell \log v_{n,\ell}^\top \hat{x}_\ell^n \tag{61}$$

Then $\hat{x}^\star = \hat{x} = \hat{x}^{\hat{n}}$ is the approximate maximizer. This is exact when the filters are disjoint $\prod_l \langle v_{nl}, v_{n'l} \rangle = 0 \forall n \neq n'$.

G Posterior approximation

We implement Eqs. (3) and (4) in the probabilistic programming language NumPyro (Phan et al., 2019). To approximate the posterior distribution $\mathrm{p}(\theta \mid \mathcal{D}_{t-1})$, we primarily use the No-U-Turn-Sampler (NUTS) with a diagonal mass matrix, target acceptance probability of 0.8, and maximum tree depth of 10 (Phan et al., 2019). 8 NUTS chains are run in parallel on a single A100 with 2000 warmup samples, drawing a total of 100000 samples. We have observed that a large amount of samples is highly beneficial to prevent overfitting of our mixtures, q_ψ, especially in the low SNR regime. If we obtain a Gelman-Rubin statistic (r-hat) greater than 2 we take this as an indication that the posterior is highly focused, preventing NUTS from sampling efficiently. In this case the posterior approximation switches to stochastic variational inference (SVI), using a mean-field approximation with a Gaussian distribution with diagonal covariance as the variational family. For SVI we use the AdamW optimizer with a learning rate of 0.001, momentum of 0.9, and adaptive gradient clipping at a ratio of 0.1 (Brock et al., 2021). We use a linear warmup of the learning rate over the first 4000 steps, and use a cosine decay throughout the rest of the training, for a total of 200000 steps. We train using the ELBO with 64 particles to reduce variance of the gradient (Wingate & Weber, 2013; Ranganath et al., 2014).

H Proof of Example 1

Example 4 (No noise, discrete map)**.** *Assume activity is discrete, s.t.* $\mathrm{f} : \mathcal{X} \to \mathcal{Y}$ *for* $\mathcal{Y} = \{0, \ldots |\mathcal{Y}| - 1\}$*. Assume the model is well specified,* $\mathrm{f} \in \{\mathrm{f}_\theta : \theta \in \Theta\}$*, and* $\pi(\mathrm{f}_\theta = \mathrm{f}) > 0$*. Assume we can synthesize arbitrary mixtures* $q \in \mathcal{Q} = \mathcal{P}(\mathcal{X})$*, and our assay has no noise,* $y_t = \langle \mathrm{f}, q \rangle$*.*

Then, by optimizing the EIG (Eq. (5)*), we find* $x^\star$ *after running one experiment.*

Proof. Number the molecules, so $\mathcal{X} = \{1, \ldots, d\}$. Set $q(x) = \frac{1}{Z} \frac{1}{|\mathcal{Y}|^x}$ where $Z = \sum_{x=1}^d \frac{1}{|\mathcal{Y}|^x}$ is the normalizing constant. Now, $y = \langle \mathrm{f}, q \rangle = \frac{1}{Z} \sum_{x=1}^d \frac{\mathrm{f}(x)}{|\mathcal{Y}|^j} \Rightarrow Zy = \sum_{x=1}^d \frac{\mathrm{f}(x)}{|\mathcal{Y}|^j}$. So $\mathrm{f}(x)$ is encoded in the digits of Zy, e.g. for $|\mathcal{Y}| = 10$ we have the decimal expansion $Zy = 0.\mathrm{f}(1)\mathrm{f}(2)\mathrm{f}(3)\ldots$. So for this experiment, the posterior over $x^\star$ is a point mass at the true $x^\star$, and the EIG is maximized. So the posterior entropy can be driven to zero. Optimizing the EIG will find this solution or anything equally as good, which also drives the entropy to zero. □

I Speedup for noiseless 1-sparse molecule-activity maps

Example 5 (No noise, continuous output, 1-sparse)**.** *Assume the molecule activity map is 1-sparse:* $\mathrm{f}(x) = w^\star \mathbb{I}(x = x^\star)$*. Assume our model is well-specified, so* $\mathrm{f}_\theta(x) = w\mathbb{I}(x = x_0)$ *for* $\theta =$

$(w, x_0) \in \mathbb{R}_+ \times \mathcal{X}$ and the prior has positive support $\pi(w^\star, x^\star) > 0$. Assume we can synthesize arbitrary mixtures $q \in \mathcal{Q} = \mathcal{P}(\mathcal{X})$, and our assay has no noise, $y_t = \langle \mathrm{f}, q \rangle$.

Then, by optimizing the EIG (Eq. (5)), we find $x^\star$ after running two experiments in parallel.

Proof. Set $q_1(x) = \frac{1}{d}\vec{1}_d$, where $\vec{1}_d$ is the d-vector of all ones, and set $q_2(x)$ so that no two entries are the same, $q_2(x) \neq q_2(x')$ for any $x \neq x'$. Now, $y_1 = \langle \mathrm{f}, q_1 \rangle = \frac{1}{d} w^\star \Rightarrow w^\star = d y_1$, so $w^\star$ is identified. Moreover, $y_2 = \langle \mathrm{f}, q_2 \rangle = w^\star q_2(x^\star) \Rightarrow q_2(x^\star) = y_2 / w^\star$, so $q_2(x^\star)$ is identified. Since each $q_2(x) : x \in \mathcal{X}$ is different, $x^\star$ is identified. So for this choice of experiments, the posterior over $x^\star$ is a point mass at the true $x^\star$, and the EIG is maximized. □

By contrast, if we tested one molecule at a time, we would need $\mathcal{O}(d)$ experiments instead of $\mathcal{O}(1)$, by the argument in Proposition 1.

J Proof of Example 2

Proof. The EIG executes 2^E-ary search. After testing q_1, the posterior over $x^\star$ is a uniform distribution over the set

$$X_{y_1} \triangleq \begin{cases} \{x : q_1(x) > 0\} \text{ if } y_1 = 1 \\ \{x : q_1(x) = 0\} \text{ otherwise,} \end{cases} \tag{62}$$

where the probability of $y_1 = 1$ is $|X_1|/d$. Testing a batch $q_{1:E}$ and receiving data $y_{1:E}$ we obtain the posterior $\mathrm{p}(x^\star \mid q_{1:E}, y_{1:E}) = \mathrm{Uniform}(X_{y_{1:E}})$ where $X_{y_{1:E}} \triangleq X_{y_1} \cap \ldots \cap X_{y_e}$. The probability of $y_{1:E}$ is $|X_{y_{1:E}}|/d$. The $X_{y_{1:E}}$ are non-overlapping and partition $\mathcal{X}$. So the EIG will be maximized if we set $q_{1:E}$ such that $\lfloor \frac{d}{2^E} \rfloor \leq |X_{y_{1:E}}| \leq \lceil \frac{d}{2^E} \rceil$ for all $y \in \{0,1\}^E$, making all partitions (as close as possible to) equal size.

After testing the first batch, we know $x^\star \in X_{y_{1:E}}$, and our posterior is uniform over this domain. Applying the same logic, the next round will cut the number of possible optima down to $\lceil \frac{d}{(2^E)^2} \rceil$, in the worst case. Iterating k times, we are left with just one molecule when $1 = \lceil \frac{d}{(2^E)^k} \rceil \Rightarrow k = \lceil \frac{\log_2 d}{E} \rceil$. □

K Information gain under a detection threshold

Assays struggle to detect activity when only a very small number of physical molecules in the mixture have activity. Rather than assume we can detect any $\langle \mathrm{f}, q \rangle > 0$, consider a detection threshold τ. Consider an assay $y = \mathbb{I}[\langle \mathrm{f}, q \rangle > \tau]$. The maximum expected information gain from a single experiment is

$$\max_{q \in \mathcal{Q}} \mathcal{H}(\mathrm{Bern}(\mathbb{P}_\theta(\langle \mathrm{f}_\theta, q \rangle > \tau))), \tag{63}$$

where $\mathcal{H}(\mathrm{Bern}(r))$ is the entropy of a Bernoulli distribution with parameter r.

The Bernoulli entropy is maximized at $\mathrm{Bern}(0.5)$. Higher detection thresholds restrict our maximum information gain, making it harder and harder for $\Pr_\theta(\langle \mathrm{f}_\theta, q \rangle > \tau)$ to reach 0.5. For example, assume the maximum activity is $\|\mathrm{f}_\theta\|_\infty$ for all θ. In the extreme where $\tau = \|\mathrm{f}_\theta\|_\infty$, we only have $\Pr_\theta(\langle \mathrm{f}_\theta, q \rangle > \tau) > 0$ if $q(x)$ is a delta mass at a single sequence, i.e. we revert to standard testing of individual molecules. In sum, we can only take advantage of mixtures if we can achieve high enough yield that more than one molecular species in the mixture passes the detection threshold.

Proof. Define the experimental outcome as a random variable by

$$Y(q) = \mathbb{I}[\langle \mathrm{f}_\theta, q \rangle > \tau] \tag{64}$$

$$= \begin{cases} 1, & \text{if} \quad \langle \mathrm{f}_\theta, q \rangle > \tau \\ 0, & \text{otherwise,} \end{cases} \tag{65}$$

where the randomness stems from the randomness in θ. This is a Bernoulli random variable with parameter $\Pr_\theta(\langle \mathrm{f}_\theta, q \rangle > \tau)$. The EIG is given by the mutual information between $Y(q)$ and θ, using

$\text{EIG}_\theta(q) = \mathcal{H}(Y(q)) - \mathcal{H}(Y(q) \mid \theta)$. Since $Y(q)$ is a deterministic function of θ, the conditional entropy is zero, $\mathcal{H}(Y(q) \mid \theta) = 0$. Therefore, the EIG is

$$\text{EIG}_\theta(q) = \mathcal{H}(Y(q)) \tag{66}$$

$$= \mathcal{H}(\text{Bern}(\Pr_\theta(\langle \text{f}_\theta, q\rangle > \tau))) \tag{67}$$

□

L Proof of Proposition 2

The signal of our experiment is $S \triangleq \langle \text{f}_\theta, q\rangle$. This is a random variable with randomness stemming from π, the prior on θ.

Proof. By the chain rule $\text{MI}(\theta, S; Y) = \text{MI}(S; Y) + \text{MI}(\theta; Y \mid S)$. Since $\text{p}(y \mid \theta, S) = \text{p}(y \mid S)$, then $\text{MI}(\theta; Y \mid S) = 0$. So for a fixed design q, the mutual information between Y and S is equal to the EIG, $\text{MI}(\theta; Y) = \text{MI}(S; Y) = \text{EIG}_\theta(q)$. We will now attempt to bound $\text{MI}(S; Y)$. The first bound stems from the bound on the entropy of S, by bounding the mutual information by the prior entropy as

$$\text{MI}(S; Y) = \mathcal{H}(S) - \mathcal{H}(S \mid Y) \tag{68}$$

$$\leq \mathcal{H}(S) = \mathcal{H}(\langle \text{f}_\theta, \bar{q}\rangle), \tag{69}$$

where we use the fact that S comes from a finite set $\mathcal{Z}$. This is an upper bound on the amount of information we can encode into an experiment. The second bound stems from the channel capacity of an additive gaussian noise channel derived by Shannon (Shannon, 1948) by

$$\text{MI}(S; Y) = h(Y) - h(Y \mid S) = h(Y) - \frac{1}{2}\log(2\pi e\sigma^2). \tag{70}$$

We assume that Y has a bounded second moment. The differential entropy, $h(Y)$, can be bounded by its maximum entropy, which is achieved when Y is Gaussian distributed with variance $\text{Var}(Y) = \text{Var}(S + \epsilon\sigma) = \text{Var}(S) + \sigma^2$

$$h(Y) \leq \frac{1}{2}\log(2\pi e(\text{Var}(S) + \sigma^2)) \tag{71}$$

Combining Eq. (70) and Eq. (71) we get

$$\text{MI}(S; Y) \leq \frac{1}{2}\log(2\pi e(\text{Var}(S) + \sigma^2)) - \frac{1}{2}\log(2\pi e\sigma^2) \tag{72}$$

$$\text{MI}(S; Y) \leq \frac{1}{2}\log\left(1 + \frac{\text{Var}(\langle \text{f}_\theta, q\rangle)}{\sigma^2}\right). \tag{73}$$

Defining $SNR(q) \triangleq \frac{c^2}{\sigma^2}\text{Var}(\langle \text{f}_\theta, \bar{q}\rangle)$, and combining the two bounds we get

$$\text{EIG}_\theta(q) \leq \min\left\{\mathcal{H}(\langle \text{f}_\theta, \bar{q}\rangle),\ \frac{1}{2}\log\left(1 + \frac{c^2}{\sigma^2}\text{Var}(\langle \text{f}_\theta, \bar{q}\rangle)\right)\right\}. \tag{74}$$

□

L.1 $\text{SNR}(q)$ is maximized at $\bar{q} = \delta_x$

Proof. To maximize $\text{SNR}(q) = \frac{c^2}{\sigma^2}\text{Var}_\pi(\langle \text{f}_\theta, \bar{q}\rangle)$, we need to maximize $\text{Var}_\pi(\langle \text{f}_\theta, \bar{q}\rangle)$. Define $\tilde{\text{f}}_\theta \triangleq \text{f}_\theta - \mathbb{E}_\pi[\text{f}_\theta]$, the centered f_θ. Then $\text{Var}_\pi(\langle \text{f}_\theta, \bar{q}\rangle) = \mathbb{E}_\pi[(\langle \tilde{\text{f}}_\theta, \bar{q}\rangle)^2]$. For any fixed θ, $\langle \tilde{\text{f}}_\theta, \bar{q}\rangle$ is a linear function in $\bar{q}$. Since x^2 is a convex function, $(\langle \tilde{\text{f}}_\theta, \bar{q}\rangle)^2$ is a convex function for any fixed θ.

So, the variance is an expectation over pointwise convex functions, which means that it is a convex function. We are therefore trying to find the maximum of a convex function $\text{Var}_\pi(\langle \text{f}_\theta, \bar{q}\rangle)$ over a compact convex set $\mathcal{P}(\mathcal{X})$. By Bauer's maximum principle, this attains its maximum at an extreme point (Bauer, 1958). For $\mathcal{P}(\mathcal{X})$, the extreme points are $\{\delta_x : x \in \mathcal{X}\}$, the set of measures concentrated on a single molecule. We can therefore establish

$$\exists \delta_x \in \underset{\bar{q} \in \mathcal{P}(\mathcal{X})}{\text{argmax}} \text{Var}(\langle \text{f}_\theta, \bar{q}\rangle) \tag{75}$$

□

M PROOF OF EXAMPLE 3

EIG bound from Proposition 2

$$\mathrm{EIG}_\theta(q_S) \leq \min\left\{-p_q \log p_q - (1-p_q)\log(1-p_q),\ \frac{1}{2}\log\left(1+\frac{1}{\sigma^2}\left(\frac{1}{d_q d}-\frac{1}{d^2}\right)\right)\right\}, \quad (76)$$

where $p_q \triangleq \frac{d_q}{d}$. The first term is strictly increasing on the interval $d_q \in [0, \frac{d}{2}]$, while the second term is strictly decreasing in d_q.

Additionally, the actual expected information gain can be approximated by

$$\mathrm{EIG}_\theta(q_U) = \left(1-\frac{d_q}{d}\right) D_{\mathrm{KL}}\left(\mathcal{N}\left(0,\sigma^2\right)||p(y)\right) + \frac{d_q}{d} D_{\mathrm{KL}}\left(\mathcal{N}\left(\frac{1}{d_q},\sigma^2\right)||p(y)\right) \quad (77)$$

where $p(y) = (1-\frac{d_q}{d})\cdot\mathcal{N}(y;0,\sigma^2) + \frac{d_q}{d}\cdot\mathcal{N}(y;1/d_q,\sigma^2)$, with $\mathcal{N}(y;\mu,\sigma^2)$ the Gaussian pdf.

Proof. The push-forward for this setup, $S(q)_\# \pi$, is given by

$$S(q_U)_\#\pi = \begin{cases} \frac{d_q}{d}, & \text{if } s = \frac{1}{d_q} \\ 1-\frac{d_q}{d}, & \text{if } s = 0. \end{cases} \quad (78)$$

This is the probability mass function of a Bernoulli random variable, scaled by $\frac{1}{d_q}$, which has entropy $\mathcal{H}(S(q_U)_\#\pi) = -p_q \log p_q - (1-p_q)\log(1-p_q)$, and variance

$$\mathrm{Var}_\pi[S(q_U)] = \frac{1}{d_q^2}\cdot\frac{d_q}{d}(1-\frac{d_q}{d}) \quad (79)$$

$$= \frac{1}{d_d d} - \frac{1}{d^2}. \quad (80)$$

Plugging into Proposition 2 we get

$$\mathrm{EIG}_\theta(q_U) \leq \min\left\{-p_q \log p_q - (1-p_q)\log(1-p_q),\ \frac{1}{2}\log\left(1+\frac{1}{\sigma^2}\left(\frac{1}{d_q d}-\frac{1}{d^2}\right)\right)\right\}. \quad (81)$$

□

For this example, we can even provide good approximations to the true MI.

True MI

Proof. The EIG can be defined as the expected KL-divergence between the prior and the posterior by

$$\mathrm{EIG}_\theta(q_U) = \mathrm{EIG}_S(q_U) = \mathbb{E}_{S(q_U)_\#\pi} D_{\mathrm{KL}}(\mathrm{p}(Y(q_U) \mid S(q_U))\|\mathrm{p}(Y(q_U))), \quad (82)$$

where $Y(q_U) = S(q_U) + \sigma\epsilon$. This is a Bernoulli random varibale plus a Gaussian random variable, with pdf $p(y) = (1-\frac{d_q}{d})\cdot\mathcal{N}(y;0,\sigma^2) + \frac{d_q}{d}\cdot\mathcal{N}(y;1/d_q,\sigma^2)$, where $\mathcal{N}(y;\mu,\sigma^2)$ is the Gaussian pdf. Since $S(q_U)_\#\pi$ is a Bernoulli distribution pdf, the EIG can be simplified to

$$\mathrm{EIG}_S(q_U) = \left(1-\frac{d_q}{d}\right) D_{\mathrm{KL}}(\mathrm{p}(Y(q_U) \mid S(q_U)=0)\|\mathrm{p}(Y(q_U))) \quad (83)$$

$$+ \left(\frac{d_q}{d}\right) D_{\mathrm{KL}}(\mathrm{p}(Y(q_U) \mid S(q_U)=\frac{1}{d_q})\|\mathrm{p}(Y(q_U))) \quad (84)$$

$$= \left(1-\frac{d_q}{d}\right) D_{\mathrm{KL}}(\mathcal{N}\left(0,\sigma^2\right))\|\mathrm{p}(y)) + \left(\frac{d_q}{d}\right) D_{\mathrm{KL}}(\mathcal{N}\left(\frac{1}{d_q},\sigma^2\right))\|\mathrm{p}(y)) \quad (85)$$

□

N PROOF OF PROPOSITION 3

Proof. Let $\mathcal{X} = [d]$ be the search space of size $|\mathcal{X}| = d$. Let $\mathrm{f}_\theta : \mathcal{X} \to \mathbb{R}_+$ be an arbitrary molecule-activity map. We aim to gain information about f_θ, not its parameters.

$q \in \mathcal{M}_c^+(\mathcal{X})$ is a positive measure subject to a 1-norm constraint $\|q\|_1 = c$. Additionally, define the set of delta functions as $\mathcal{S}_c(\mathcal{X}) = \{c\delta_x : x \in \mathcal{X}\} \subset \mathcal{M}_c^+(\mathcal{X})$ subject to the same constraint. Assume Gaussian noise $\mathrm{p}(y \mid \mathrm{f}_\theta, q, \sigma) = \mathcal{N}(\langle \mathrm{f}_\theta, q\rangle, \sigma^2)$. The score function for this likelihood is

$$\nabla_{\mathrm{f}_\theta} \log \mathcal{N}(\langle \mathrm{f}_\theta, q\rangle, \sigma^2) = \nabla_{\mathrm{f}_\theta} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y - \langle \mathrm{f}_\theta, q\rangle)^2}{2\sigma^2} \right) \tag{86}$$

$$= \nabla_{\mathrm{f}_\theta} \left(-\frac{(y - \langle \mathrm{f}_\theta, q\rangle)^2}{2\sigma^2} \right) \tag{87}$$

$$= \frac{y - \langle \mathrm{f}_\theta, q\rangle}{\sigma^2} \nabla_{\mathrm{f}_\theta} \langle \mathrm{f}_\theta, q\rangle \tag{88}$$

$$= \frac{y - \langle \mathrm{f}_\theta, q\rangle}{\sigma^2} q \tag{89}$$

The information matrix is given by the outer product of the score function, evaluated at the true parameter f_{θ_0} by

$$\mathcal{I}_{\mathrm{f}_{\theta_0}}(q) = \mathbb{E}_{\mathrm{p}(y|q,\mathrm{f}_\theta)} \left[\nabla_{\mathrm{f}_\theta} \log \mathrm{p}(y \mid q, \mathrm{f}_\theta) \nabla_{\mathrm{f}_\theta} \log \mathrm{p}(y \mid q, \mathrm{f}_\theta)^\top \Big|_{\mathrm{f}_\theta = \mathrm{f}_{\theta_0}} \right] \tag{90}$$

$$\mathcal{I}_{\mathrm{f}_{\theta_0}}(q) = \mathbb{E}_{\mathrm{p}(y|q,\mathrm{f}_\theta)} \left[\left(\frac{y - \langle \mathrm{f}_\theta, q\rangle}{\sigma^2} q \right) \left(\frac{y - \langle \mathrm{f}_\theta, q\rangle}{\sigma^2} q \right)^\top \Bigg|_{\mathrm{f}_\theta = \mathrm{f}_{\theta_0}} \right] \tag{91}$$

$$= \frac{1}{\sigma^4} \mathbb{E}_{\mathrm{p}(y|q,\mathrm{f}_\theta)} \left[(y - \langle \mathrm{f}_\theta, q\rangle)^2 \, q \otimes q \Big|_{\mathrm{f}_\theta = \mathrm{f}_{\theta_0}} \right] \tag{92}$$

$$= \frac{\mathbb{E}_{\mathrm{p}(y|q,\mathrm{f}_\theta)} \left[(y - \langle \mathrm{f}_\theta, q\rangle)^2 \Big|_{\mathrm{f}_\theta = \mathrm{f}_{\theta_0}} \right]}{\sigma^4} q \otimes q \tag{93}$$

$$= \frac{1}{\sigma^2} q \otimes q \tag{94}$$

The result from Paninski (Paninski, 2005) states that for information optimal sampling, the posterior $\mathrm{p}(\mathrm{f}_\theta \mid \mathcal{D}_t)$ is asymptotically normal, with covariance matrix $\frac{\sigma^2_{info}}{t}$ determined by

$$\sigma^2_{info}(\mathcal{Q}) = \left(\underset{C \in co(\mathcal{I}_{\mathrm{f}_{\theta_0}}(\mathcal{Q}))}{\operatorname{argmax}} \det(C) \right)^{-1}, \tag{95}$$

where $co(\mathcal{I}_{\mathrm{f}_{\theta_0}}(\cdot))$ denotes the convex closure of the set of information matrices, and $\det(A)$ is the determinant of a matrix A. We will now show how restricting $\mathcal{Q}$ to $\mathcal{S}_c(\mathcal{X})$ does not change the asymptotic variance over the case of the full measure $\mathcal{M}_c^+(\mathcal{X})$. First, note that $\mathcal{I}_{\mathrm{f}_{\theta_0}}(q) = \frac{1}{\sigma^2} q \otimes q \in \mathbb{R}_+^{d\times d}$ is positive semidefinite by construction, and that it has a fixed sum $\sum_{i,j} \mathcal{I}_{\mathrm{f}_{\theta_0}}(q)_{i,j} = \frac{1}{\sigma^2} q_i q_j = \frac{c^2}{\sigma^2}$. The convex closure of any set of information matrices inherits these properties, because the set of positive semidefinite matrices is convex, as is the set of matrices with a fixed sum, and the set of positive-valued matrices. We can write this succinctly as $co(\mathcal{I}_{\mathrm{f}_{\theta_0}}(\mathcal{Q})) \subseteq \{M : M_{ij} \geq 0 \ \forall i,j,\ \lambda_1, ...\lambda_d \geq 0,\ \sum_{i,j} M_{i,j} = \frac{c^2}{\sigma^2}\}$, where λ denotes an eigenvalue. We can use $\det(C) = \prod_{i=1}^d \lambda_i$, as well as $\mathrm{Tr}(C) = \sum_{i=1}^d \lambda_i = \sum_{i=1}^d C_{i,i} \leq \sum_{i,j} C_{i,j} = \frac{c^2}{\sigma^2}$ to conclude that the determinant is bounded by $\det(C) \leq \prod_{i=1}^d \frac{c^2}{d\sigma^2}$, as this is the point where all eigenvalues are

equal, and the trace is maximal. Furthermore, since

$$\det(\mathrm{I}_d \cdot \frac{c^2}{d\sigma^2}) = \prod_{i=1}^{d} \frac{c^2}{d\sigma^2}, \tag{96}$$

and

$$\mathrm{I}_d \cdot \frac{c^2}{d\sigma^2} \in co(\mathcal{I}_{\mathrm{f}_{\theta_0}}(\mathcal{S}_c(\mathcal{X}))), \tag{97}$$

this supremum is attained within $\mathcal{S}_c(\mathcal{X})$. To say it another way, the determinant, and hence the asymptotic variance, is maximized at $C^\star = \mathrm{I}_d \cdot \frac{c^2}{d\sigma^2}$. Because $C^\star \in co(\mathcal{I}_{\mathrm{f}_{\theta_0}}(\mathcal{S}_c(\mathcal{X})))$, there can be nothing to gain in the asymptotic sense by lifting to $q \in \mathcal{M}_c^+(\mathcal{X})$, and we may conclude

$$\sigma^2_{info}(\mathcal{S}_c(\mathcal{X})) = \sigma^2_{info}(\mathcal{M}_c^+(\mathcal{X})) \tag{98}$$

□

O Details on sparse synthetic oracle simulations

Our sparse oracle is created by sparsifying draws from our prior $\pi(\theta)$ and setting the maximum to one $\|\mathrm{f}\|_\infty = 1$. The prior is described in App. F.1. We set $N = 1$ and $w = 1$, and then draw 10 independent repeats $\theta_{1:10} \sim \pi(\theta)$. Afterwards, we construct the oracle by $\mathrm{f}(x) = \mathrm{f}_\theta(x) \cdot \mathbb{I}[d_H(x, x^\star_\theta) \leq 1]$, where d_H denotes the Hamming distance.

Such molecule-activity maps $\mathrm{f} : \mathcal{X} \to [0, 1]$ have the following qualities:

1. Exactly one DNA sequence $x^\star$ has activity $\mathrm{f}(x^\star) = 1$.
2. Every single mutation has lower activity. $0 < \mathrm{f}(x) < \mathrm{f}(x^\star) \,\forall\, d_H(x, x^\star) = 1$.
3. Every double mutation abolishes activity. $\mathrm{f}(x) = 0 \forall d_H(x, x^\star) > 1$.

For both LIDS and Thompson, we use the same hyperparameters of the model f_θ. We set the number of filters to $N = 8$, regularization strength $\lambda = 1$, and use $\mathrm{p}(w) = \delta_{w=1}$ as the prior on $\|\mathrm{f}\|_\infty$. Hyperparameters for EIG optimization and posterior approximation are described in Apps. E.2 and G, respectively.

P CI calculation

95% confidence intervals are calculated using a t-critical value and the delta method. Let $\hat{\mathrm{f}}_n \triangleq \mathrm{f}(\hat{x}^\star_n)$ be the activity of the test molecules for the N replicates. First, compute the standard error and mean.

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} \hat{\mathrm{f}}_n \tag{99}$$

$$s^2 = \frac{1}{n-1} \sum_{n=1}^{N} (\hat{\mathrm{f}}_n - \hat{\mu})^2 \tag{100}$$

$$SE = \frac{s}{\sqrt{n}} \tag{101}$$

Then, use $g(x) \triangleq \log(\frac{x}{\mathrm{f}(x^\star) - x})$ and the delta method to calculate a 95% CI.

$$SE_{logit} = g'(\hat{\mu}) \cdot SE \tag{102}$$

$$CI_{logit} = g(\hat{\mu}) \pm t_{n-1}(0.975) \cdot SE_{logit} \tag{103}$$

$$CI = g^{-1}(CI_{logit}) \tag{104}$$

Q Details on PLM oracle simulations

For the hyperparameters of the surrogate f_θ, we set the number of filters to $N = 20$, regularization strength $\lambda = 1$, and use $\mathrm{p}(w) = 1 + \mathrm{Gamma}(1, 1)$ as the prior on $\|\mathrm{f}\|_\infty$. Hyperparameters for training and posterior approximation are described in Apps. E.2 and G.

Q.1 IMPORTANCE SAMPLING $\langle \mathrm{f}, q \rangle$

To simulate an experiment we need to evaluate $\langle \mathrm{f}, q \rangle = \mathbb{E}_{x \sim q(x)}[\mathrm{f}(x)]$. To do so we exploit the fact that $\mathrm{p}(x)$ can be sampled from, as it is given by a PLM, which is a generative model. We set a defensive mixture distribution $s(x) = \frac{1}{2}q(x) + \frac{1}{2}\mathrm{p}(x)$ as the proposal distribution (Hesterberg, 1995) and approximate the expectation by importance sampling.

$$\langle \mathrm{f}, q \rangle \approx \frac{1}{N} \sum_{i=1}^{N} \frac{q(x^i)}{s(x^i)} \mathrm{f}(x^i), \quad x^i \sim s(x) \tag{105}$$

R ABLATION OF POSTERIOR APPROXIMATION

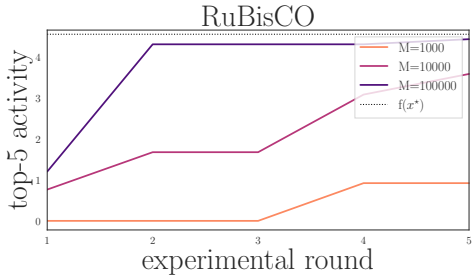


Figure 6: Posterior approximation. LIDS using different number of posterior samples M. The setup is the same as in Sec. 6.2.

S VISUALISATION OF LIDS

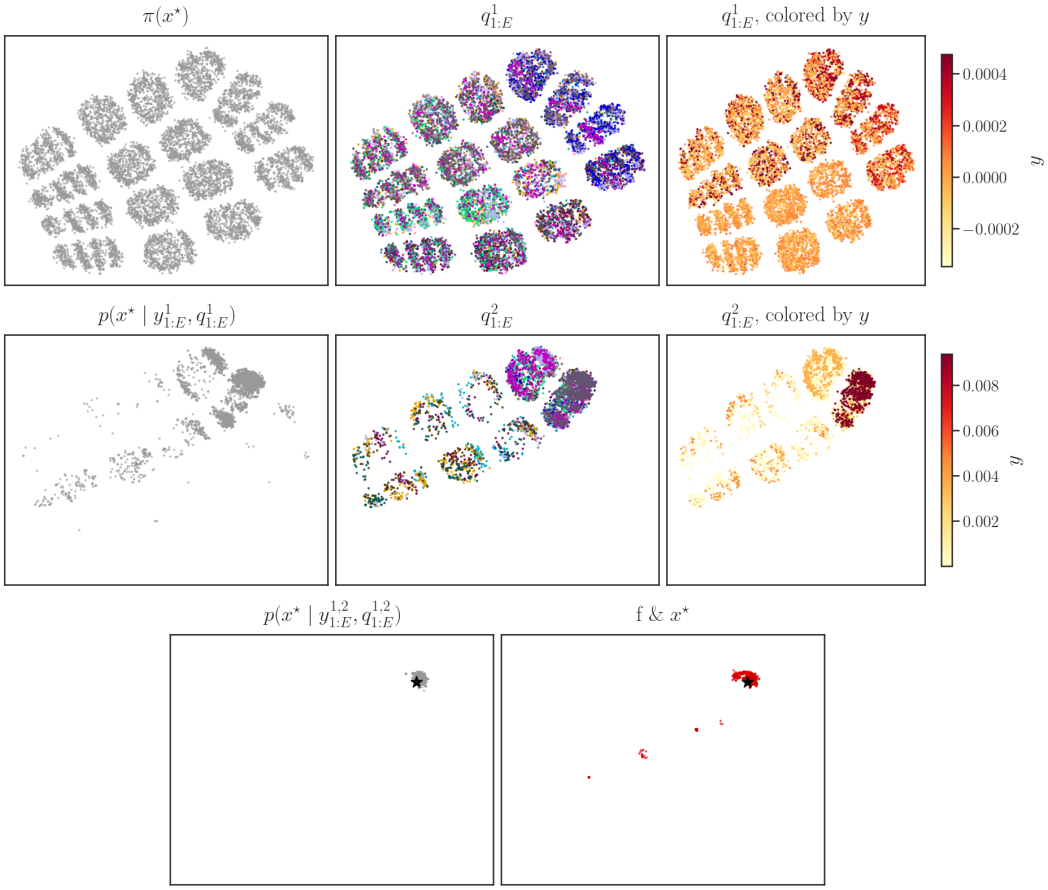


Figure 7: Two rounds of LIDS on a sparse synthetic oracle. Data from one of the duplicates from Sec. 6.1. UMAP projection of samples from different distributions: First row contains the prior $\pi(x^\star)$ (left), the mixtures designed at $t = 1$ by LIDS $q^1_{1:E}$ (middle), and the same mixtures, colored by their activity $y = \langle \mathrm{f}, q \rangle + \epsilon\sigma$ (right). The second row contains the posterior after one experimental round $\mathrm{p}(x^\star \mid \mathcal{D}_1 = (q^1_{1:E}, y^1_{1:E}))$ (left), the mixtures designed at $t = 2$ by LIDS $q^2_{1:E}$ (middle), and the same mixtures, colored by their activity $y = \langle \mathrm{f}, q \rangle + \epsilon\sigma$ (right). The third row contains the posterior after two rounds of experiments $\mathrm{p}(x^\star \mid \mathcal{D}_1 = (q^{1,2}_{1:E}, y^{1,2}_{1:E}))$ (left), and the sparse oracle we are optimizing, normalized to be a distribution $\mathrm{f}(x)/\sum_{x \in \mathcal{X}} \mathrm{f}(x)$ (right). After these two rounds, LIDS has converged on the global optimum.