Compositional Deep Probabilistic Models of DNA Encoded Libraries

Abstract

DNA-Encoded Library (DEL) has proven to be a powerful tool that utilizes combinatorially constructed small molecules to facilitate highly efficient screening assays. These selection experiments, involving multiple stages of washing, elution, and identification of potent binders via unique DNA barcodes, often generate complex data. This complexity can potentially mask the underlying signals, necessitating the application of computational tools such as machine learning to uncover valuable insights. We introduce a compositional deep probabilistic model of DEL data, **DEL-Compose**, which decomposes molecular representations into their mono-synthon, di-synthon, and tri-synthon building blocks and capitalizes on the inherent hierarchical structure of these molecules by modeling latent reactions between embedded synthons. Additionally, we investigate methods to improve the observation models for DEL count data such as integrating covariate factors to more effectively account for data noise. Across two popular public benchmark datasets (CA-IX and HRP), our model demonstrates strong performance compared to count baselines, enriches the correct pharmacophores, and offers valuable insights via its intrinsic interpretable structure, thereby providing a robust tool for the analysis of DEL data.

1   Introduction

DNA-Encoded Libraries (DELs) have demonstrated their potency as a robust method for conducting efficient exploration across a vast chemical landscape, and has recently gained significant traction in drug discovery efforts Goodnow Jr et al. [2017], Yuen and Franzini [2017], Neri and Lerner [2018], Madsen et al. [2020], Satz et al. [2022], Peterson and Liu [2023]. These small molecule libraries are synthsized combinatorially by combining diverse building blocks with compatible chemistries. A DNA barcode, which is covalently attached to the molecule, specifies the unique combination of building blocks for each molecule. DELs are then used in selection experiments for proteins of interest, wherein multiple rounds of washing and elution are performed before identification of the surviving library molecules. While proven to be a highly efficient process of exploring chemical space at scale, these selection experiments are noisy and require computation methods with the correct inductive biases to extract useful signals for downstream applications such as hit discovery and lead optimization Gironda-Martínez et al. [2021], Reiher et al. [2021].

Prior work has tackled the analysis of DEL data from various perspectives. Many of these methods are predicated on computing an enrichment score for each molecule as a function of the observed data or molecule structure. Gerry et al. [2019] computes this enrichment score by fitting Poisson distributions to the count data and then computing the ratio of the on-target versus off-target binding affinities derived from the fitted distributions. However, this approach and other similar approaches Kuai et al. [2018], Faver et al. [2019] do not extend to out-of-library predictions as the enrichment is computed from the data itself rather than predicted via molecular structure. To that end, other methods have tackled this problem by utilizing molecule structure via molecular fingerprints and graph neural networks. McCloskey et al. [2020] and Zhang et al. [2023] bin the count data and

construct a classification problem based on their discretizations of the data. In particular, Zhang et al. [2023] proposes exploiting the compositional structure of DELs for the extraction of enrichment signals, but unlike our work, no explicit generative models of this factorized representation or the improved likelihoods to deal with DEL data are built. Other approaches formulate the problem as a latent-variable prediction task, maximizing the probability of observing the count data under some prescribed probability distribution such as the Poisson or Negative Binomial distribution Binder et al. [2022], Lim et al. [2022], Ma et al. [2021]. Shmilovich et al. [2023] extends the representation capabilities of models on DEL data by incorporating 3-D docked poses to enhance the performance of models without requiring additional supervised validation data. However, these prior works do not leverage the inherent hierarchical structure of DEL data.

We introduce an new approach to modeling DEL molecules, which explicitly factorizes the molecular representation in a motivated manner through the construction of a generative model. We propose learning individual synthon representations, and construct the corresponding di-synthon and tri-synthon representation from their respective synthon compositions utilizing learned latent reactions parametrized by neural networks. Signals in DEL selection experiments are obfuscated by the various sources of noise, but given the combinatorial construction of the library, these sources of variation are highly correlated within any particular synthon group. Our method additionally avoids the necessity of enumerating full-molecule structures, which can often be a tedious process prone to errors, and only requires molecular structures of individual synthons.

In addition to our newly proposed paradigm for representing DEL molecules computationally in this factorized fashion, we further investigate the effects of different experimental biases in order to aptly model the count data. In particular, we focus on two prominent sources of noise inherent in DEL data, which stem from pre-selection and replicate-level biases. Modeling these is typically omitted in previous work on this topic. Since DEL molecules are synthesized using a split-and-pool method, the relative abundance of each library member is uncertain in the final mixture. While the library itself is sequenced to obtain a rough estimate of the molecule distribution, this count data is also prone to potential synthesis and sequencing biases. Across different replicates, we also expect to see different experimental or sequencing noise. We propose a structured parametrization in our count likelihoods to account for the effects of these factors in order to better model the observed count data and learn useful latent properties of DEL molecules.

We test our models empirically on DEL selection datasets for two targets: Carbonic Anhyrase IX (CA-IX) and Horseradish peroxidase (HRP) Gerry et al. [2019]. Since these two well-studied targets have known pharmacophores, we demonstrate that our model can effectively pick out the important synthons–even on challenging splits of the data. Furthermore, we demonstrate that our method can obtain competitive performance even without requiring fully enumerated molecule structures. Lastly, we show that our model offer useful insights into the predictions given by the model.

2 Methods

2.1 Molecules As Synthon Compositions

Our model broadly capitalizes on the combinatorial nature of DEL molecules, and creates a composition of representations using the individual building blocks of each molecule. To that end, we describe a generative model of the underlying data-generating process for DEL count data and first introduce some mathematical notation.

Let $\mathcal{X}$ be the set of DEL molecules in our dataset, and $\{\mathcal{S}_A, \mathcal{S}_B, \mathcal{S}_C\}$ be the sets of synthons at the first, second and third positions respectively. Each molecule is denoted by $x_{abc} \in \mathcal{X}$, where the subscript indicates the identity of the synthon at a particular position ($a \in \mathcal{S}_A, b \in \mathcal{S}_B, c \in \mathcal{S}_C$). To simplify (and overload) notation, we omit the subscript for a particular synthon position if it is absent. For instance, x_b denotes the molecule corresponding to the synthon b at the second position, and x_{ab} denotes the molecule corresponding to the combination of synthon a at the first position and b at the second position.

DEL molecules are used in selection experiments wherein molecules undergo multiple rounds of washes to determine the strongest binders. Molecules with strong binding affinity would not be eluted off, but this binding might not be specific to the protein of interest. In order to also account for non-specific binding of the molecules, there are two typically two experimental conditions that

are run, the **target** condition, which describes the data for selection against the protein target of interest, and the **matrix** condition, which describes the data in the absence of the protein target. Once the selection experiments are conducted, the surviving DEL members are sequenced, resulting in DNA read count data which we will denote as $C_t = \{c_t^i | i \in [1, n_t]\}$ and $C_m = \{c_m^j | j \in [1, n_m]\}$ for target and matrix read counts respectively. Here, (n_t, n_m) are the number of count replicates for target and matrix respectively. Moreover, DEL data is usually calibrated with an additional read-out of the library itself, which we denote as c_p (this notation is lowercase, as there is usually only a single read-out of the library). This library read-out is a noisy estimate of the relative abundance of each molecule member.

2.2 DEL-Compose: A Generative Model of DEL Data

Here we introduce **DEL-Compose**, a general paradigm for modeling DEL molecules leveraging their combinatorial nature (Figure 1). Our objective is to maximize the likelihood of observing the count data given an input molecule x_{abc}. Let $\mathcal{Z} = \{z_s \in \mathcal{R}^d | s \in [a, b, c, ab, abc, mol]\}$ be a collection of latent synthon embeddings each of dimension d. z_a denotes an embedding of an individual synthon x_a, while z_{ab} denotes an embedding of a di-synthon x_{ab} which is the reaction output of x_a and x_b. Similarly, z_{abc} denotes an embedding of tri-synthon x_{abc}, which itself is a product of the reaction of x_{ab} and x_c. Finally, z_{mol} is the aggregated embedding of all the above representations. We thus assume embeddings corresponding to all the (partial) products and building blocks of molecule can be viewed as a synthon decomposition and their respective set of chemical reactions.

We utilize this structure to factorize a model of count observations given synthons $p(C_t, C_m | x_a, x_b, x_c)$ into two quantities as shown in Equation 1: a model capturing our beliefs about count observations given a collection of synthon embeddings $\mathcal{Z}$ and a model mapping observed synthons to such embeddings. Note that we do not require access to the full molecule observation x_{abc}, but rather only the individual synthons: x_a, x_b, and x_c in such a model.

$$p(C_t, C_m | x_a, x_b, x_c) = \int p(C_t, C_m | \mathcal{Z}) \cdot p(\mathcal{Z} | x_a, x_b, x_c) d\mathcal{Z}. \tag{1}$$

In practice we will be inferring point estimates of the embeddings corresponding to $p(\mathcal{Z} | x_a, x_b, x_c)$ but present the model in generality. In order to de-noise the contribution of actual molecule binding to the count data read-outs, we explicitly define latent enrichment parameters $\{\lambda_t, \lambda_m\}$, which capture a molecule's affinity for binding in the target and matrix experimental conditions. While there are many auxiliary factors that affect the final read count for DEL experiments, we choose two prominent factors to incorporate in our model, which are pre-selection library read-out, c_p, and replicate-level noise, which we denote as $\{\gamma_t, \gamma_m\}$. The generative model can then be broken down according to Equation 2:

$$p(C_t, C_m, \Phi_{mol}, \mathcal{Z} | x_a, x_b, x_c; \Theta) = p(C_t, C_m | \Phi_{mol}; \Theta_x) p(\Phi_{mol} | \mathcal{Z}; \Theta_o) p(\mathcal{Z} | x_a, x_b, x_c; \Theta_i). \tag{2}$$

Here, Φ_{mol} is the set of variables predicted by the model that parameterizes the output distribution, for instance (λ_t, λ_m). The model is then broken up into three components: (1) Θ_i consists of the model parameters that construct the hierarchical synthon embeddings, (2) Θ_o consists of the model parameters that predict Φ_{mol} from these embeddings, and (3) $\Theta_x = \{c_p, \gamma_t, \gamma_m\}$ consists of the parameters for the observation model that captures experimental noise. We define the joint set of these parameters as $\Theta := \{\Theta_i, \Theta_o, \Theta_x\}$. We summarize the joint model visually in Figure 1 and will develop and define its components in more detail over the following sections.

2.3 Neural Network Representations of synthons and molecules

Here, we delineate our explicit modeling choices with respect to the probabilistic model introduced earlier, which is captured as the top part of the model visualized in Figure 1. Assuming $\mathcal{Z} = \{z_s | s \in [a, b, c, ab, abc, mol]\}$ we unpack $p(\mathcal{Z} | x_a, x_b, x_c; \Theta_i)$ to yield expressions in detail for each latent variable as follows:
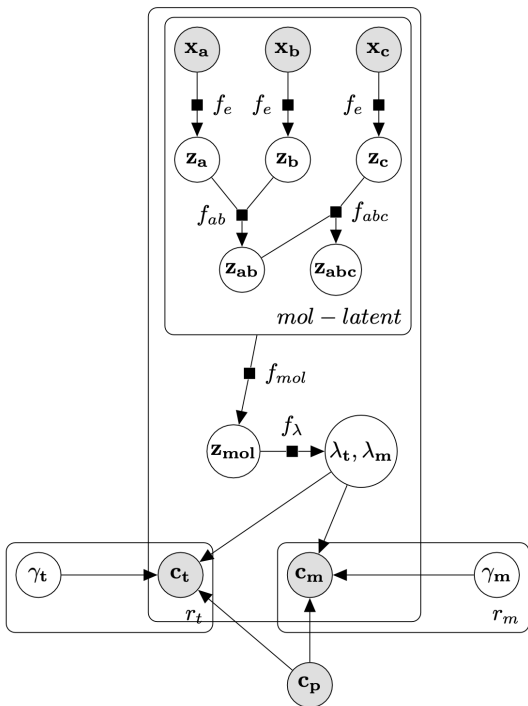
3

Figure 1: Graphical model depicting data-generating process for DEL count data.

$$p(\mathcal{Z}|x_a, x_b, x_c; \Theta_i) = \prod_s^{\{a,b,c\}} \Big[p(z_s|x_s; f_e) \Big] p(z_{ab}|z_a, z_b; f_{ab}) p(z_{abc}|z_{ab}, z_c; f_{abc}) p(z_{mol}|\mathcal{Z}_{\neq mol}; f_{mol}),$$

(3)

where $\{f_s|s \in [a, b, c, ab, abc, mol]\}$ is a set of functions modeling the generative process by which the synthon embedding z_s can be constructed, including (i) transformations from observed structures to embeddings and (ii) latent reactions modeling how synthon embeddings compose in latent space to form embeddings over higher order synthon structures.

Neural networks parametrize synthon embeddings We first identify function f_e as a model parametrized by a multi-layer perceptron with parameters θ_e mapping molecular fingerprints $\phi(x_s)$ of a synthon x_s to an embedding per synthon z_s. Let ϕ be a fingerprint transformation such as Morgan Fingerprints Rogers and Hahn [2010]; we compute the latent mono-synthon representations as $z_s = f_e(\phi(x_s); \theta_e)$. ϕ can also extend to other molecule representations such as graph neural networks (or other higher order functions that act on different data modalities of molecules), but we find fingerprints work well empirically and are fast to compute.

Neural networks parametrize latent reactions between synthon embeddings We then compute di-synthon and tri-synthon embeddings using mono-synthon embeddings, and give some examples as follows. Di-synthon embeddings utilize a latent reaction f_{ab} to compute embedding $z_{ab} = f_{ab}([z_a, z_b]; \theta_{ab})$ and tri-synthon embeddings are analogously given as $z_{abc} = f_{abc}([z_{ab}, z_c]; \theta_{abc})$. Here, (f_{ab}, f_{abc}) are separate neural networks parameterized by $(\theta_{ab}, \theta_{abc})$ respectively. f_{ab} can be interpreted to model latent reactions between mono-synthon embeddings z_a and mono-synthon embeddings z_b to yield a product-synthon-embedding z_{ab}. Likewise, f_{abc} models the function mapping z_{ab} and z_c to the tri-synthon-embedding z_{abc}. There are multiple ways to generate the tri-synthon embedding, for instance instead using all three mono-synthon embeddings, but we choose one formulation that is consistent with the sequential nature of a DEL molecule's synthesis. Although (x_{bc}, x_{ac}) are not actual observed partial products, we can optionally learn their respective embeddings and
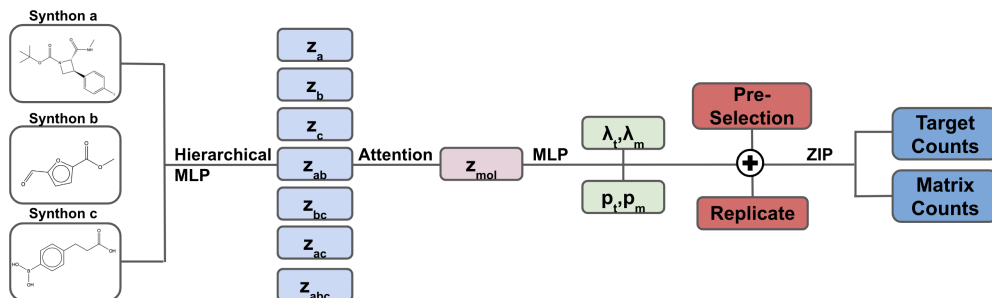
4

Figure 2: Model architecture for our model utilizing synthon-based embeddings. Synthons embedded and composed to mono-, di-, and tri-synthon embeddings, before aggregated into a single representation. Conditioned on the molecule embedding, our model predicts latent properties of the molecules that are used to model the observed count data. While the generative process does not explicitly incorporate x_{bc} and x_{ac}, these are sensible molecular subgraphs that make sense in the representation of the molecule.

incorporate them into the model, as illustrated in Figure 2. Finally, we utilize a multi-head attention layer f_{mol} with parameters θ_{mol} on the different synthon embeddings to construct a single embedding representing the entire molecule, $z_{mol} = \text{Multihead-Attention}([z_a, z_b, z_c, z_{ab}, z_{bc}, z_{ac}, z_{abc}]; \theta_{mol})$. We note that z_{mol} can be different from z_{abc} or trivially equal to it. The model has the freedom to utilize z_{mol} to focus on abstracting partial products if those are more informative about enrichment, while z_{abc} is an explicit representation of the tri-synthon embedding and does not have to be maximally informative about enrichment on its own. We summarize all parameters of these individually specified functions as $\Theta_i = \{\theta_e, \theta_{ab}, \theta_{abc}, \theta_{mol}\}$.

2.4 Probabilistic Models of Enrichment Counts

In this section we will present some choices for the model generating the per-molecule observation model parameters given the embeddings $p(\Phi_{mol}|\mathcal{Z}; \Theta_o)$. We model the observed count as a zero-inflated Poisson (ZIP) distribution (PDF given in Equation 4), the parametrization of which is conditioned on our learned embeddings. A ZIP distribution is a mixture distribution that makes a bi-modal outcome assumption: a zero outcome indicating absence of measurable data, and a measurable enrichment with some appropriate count distribution. Intuitively, because DEL data is highly susceptible to noise of different types, including synthesis, amplification, or sequencing noise, we can think of the zero-probability as a drop-out parameter explaining away the absence of an expected count rather than forcing the model to absorb it by adjusting the expected enrichment.

$$\text{P}(C = c|\lambda, \pi) = \begin{cases} \pi + (1 - \pi)e^{-\lambda} & \text{if } c = 0 \\ (1 - \pi)\frac{\lambda^c e^{-\lambda}}{c!} & \text{if } c > 0 \end{cases} \tag{4}$$

For each of the experimental conditions, the target and matrix, we predict a separate ZIP with correlated parameters. A ZIP is characterized by two parameters, a mean value λ, and a zero probability $p \in [0, 1)$ that describes the rate of dropout. From our learned molecule embedding z_{mol}, we predict two sets of parameters $\{\lambda_t, p_t\}$ and $\{\lambda_m, p_m\}$ using the function f_λ instantiated by an MLP with parameters θ_o (and consequently $\Theta_o := \theta_o$). For this likelihood this means that $\Phi_{mol} = \{\lambda_t, p_t, \lambda_m, p_m\}$, but we note that Φ_{mol} can take different appropriate shapes if other likelihoods are chosen. : $[\lambda_t, p_t, \lambda_m, p_m] = f_\lambda(z_{mol}; \theta_o)$. λ can be thought of as a molecule's intrinsic binding affinity property as learned by the model, while p captures the noise in the data. This also highlights the utility of capturing enrichment-related molecule abstractions using z_{mol}, which can focus on utilizing partial products to predict noisiness and enrichment with more fidelity as our experiments will show. If the data is inherently noisy and experience high dropout (or zero counts), the zero probability parameter should be relatively high.

Since we expect that the binding in the matrix condition should be informative of a molecule's off-target binding, we want to correlate the predicted distribution of the target condition to the matrix. Additionally, we add the pre-selection counts c_p and learned replicate-level effects $\{\gamma_t, \gamma_m\}$

as multiplicative factors in the mean of the predicted distribution. The latter accounts for variance across different replicates of the same experiment, which can result from PCR bias noise. We denote the collection of the relevant parameters for modeling experimental conditions and noise as $\Theta_x = \{c_p, \gamma_t, \gamma_m\}$. Together, we arrive at the following function form:

$$c_t^i \sim \text{ZIPoisson}\left(c_p \cdot \exp\left(\gamma_t^i\right) \cdot \exp\left(\lambda_m + \lambda_t\right), p_t\right), \tag{5}$$

$$c_m^j \sim \text{ZIPoisson}\left(c_p \cdot \exp\left(\gamma_m^j\right) \cdot \exp\left(\lambda_m\right), p_m\right), \tag{6}$$

where i and j are target and matrix replicate indices, respectively. The output and count models introduced in this section are captured in the bottom part of the graphical model depicted in Figure 1.

3   Results and Discussion

Experimental Setup   We conduct experiments on a public dataset of tri-synthon DEL data from Gerry et al. [2019], which includes DEL selection data on two well-studied protein targets: carbonic anhydrase IX (CA-IX) and horseradish peroxidase (HRP). Their data consists of two experimental conditions, one with the protein target, and one matrix condition that is conducted in the absence of the protein as control. For CA-IX this dataset includes 2 replicates of matrix data, and 4 replicates of on-target protein data; while for HRP, this dataset includes 4 replicates of matrix data and 2 replicates of on-target protein data. Additionally, there is data collected on the pre-selection library, which is an indicator of the relative abundance of the different DEL members. Information about important pharmacophores is given in

Metrics   We utilize several well-motivated metrics to evaluate the performance of our model without additional data (ie on-DNA K_D data for DEL molecules). Since we model the observed data through predicting a count distribution, we can measure the performance through the model loss, which is the negative log likelihood (NLL) of the ZIP distribution predicted for molecules in a held-out test set. This is a typical metric to gauge the overall fitness of a probabilistic model. However, as with other applications, likelihood metrics can be complemented with other application-relevant metrics to capture the behavior of the model.

Since we are interested in the quality of the learned model, we want to directly capture its ability to represent signals in the data. We use the expected mean of the predicted distribution as the computed enrichment, or affinity, of a molecule. For our model, this is exactly $\epsilon = (1 - p) \cdot \lambda$, where p is the predicted zero-probability and λ is the predicted latent score for a molecule. While DEL-Compose predicts two count distributions, one for the target and one for the matrix, the latter is mainly used to calibrate a molecule's affinity for the protein target. A molecule with high counts in the matrix but not the target condition should be predicted to have a high matrix enrichment score, but a low target enrichment score. Using these enrichment scores, we can gauge the performance of our model at an synthon-aggregate level, as we know the pharmacophores and their relative levels of activity.

Lastly, we introduce a new metric to evaluate the quality of our model's predictions by the ability of our model to separately out different classes molecules, which we define as having a better predictor. We first introduce some notation: **CA-IX** has three distinct groups, $\{g_{para}, g_{meta}, g_{other}\}$, in order of protein activity for the para-substituted sulfonamides, meta-substituted sulfonamides, and other molecules respectively. **HRP** has four distinct groups, $\{g_{e1}, g_{e2}, g_{e3}, g_{other}\}$, in order of protein activity for the three different Michael acceptor electrophiles (described in Figure 6) and other molecules respectively.

From this, we define a multi-class precision-recall area under the curve (PR-AUC) in order to evaluate the ability of our model to differentiate molecule classes. Let $s(g_a|g_b)$ be the computed PR-AUC using g_a as the positive class and g_b as the negative class. Since we know the expected rankings of these molecule classes (ie $g_{para} > g_{meta} > g_{other}$), we can compute the AUC for each pair and then take an unweighted average over all such pairs. Since the data is heavily skewed towards representation of molecules without appreciable activity towards the protein target, we weigh each molecule class equally. These AUC computations are noted exactly in Equations 7 and 8:

$$\text{PR-AUC}^{\text{CA-IX}} = \frac{1}{3}\Big[s(g_{para}|g_{meta}) + s(g_{para}|g_{other}) + s(g_{meta}|g_{other})\Big], \tag{7}$$
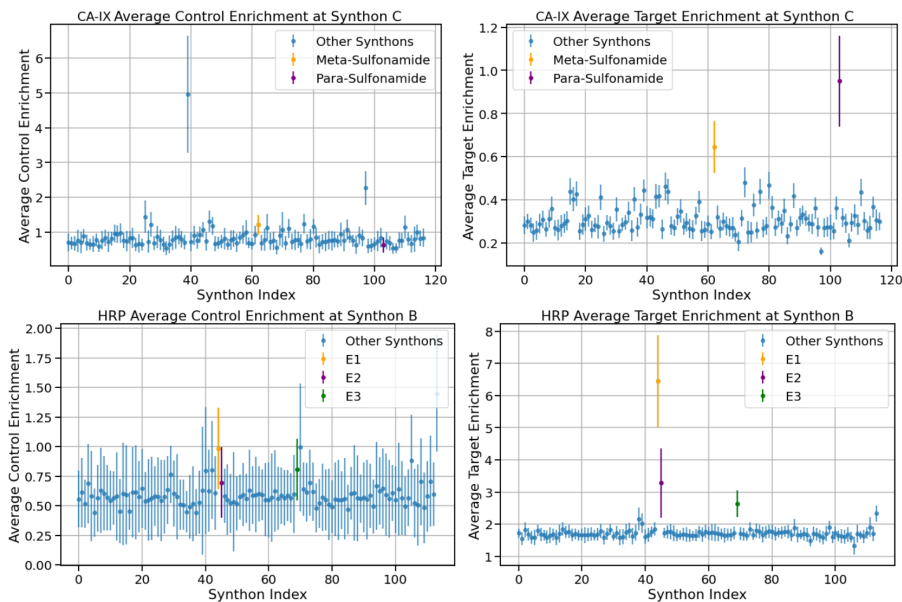
Figure 3: Predicted average marginal enrichment from DEL-Compose of both control and target counts for CA and HRP. Our model can distinguish synthons with high noise in the matrix, but not actual protein binding activity.

$$\text{PR-AUC}^{\text{HRP}} = \frac{1}{6}\Big[s(g_{e1}|g_{e2}) + s(g_{e1}|g_{e3}) + s(g_{e2}|g_{e3}) + s(g_{e1}|g_{other}) + s(g_{e2}|g_{other}) + s(g_{e3}|g_{other})\Big].$$

(8)

3.1 DEL-Compose captures enrichment of important pharmacophores

Our first result is to show that DEL-Compose can correctly determine the important pharmacophores in the dataset. In Figure 3, we plot the average marginal enrichment of a synthon as predicted by our model on the test set. For both protein targets, our model correctly enriches the important synthons, which are the benzene-sulfonamides for CA-IX and the Michael-acceptor electrophiles for HRP. Moreover, our model predicts the correct ranking of these different groups. What is particularly interesting to observe is that our model enriches certain non-sulfonamides synthons in the control experiments of CA-IX, but not in the target. This signifies that our model can correctly distinguish the synthons which might have high noise, or off-target matrix binding.

Next, we compare our deep probabilistic approach to several baselines that compute enrichments based on counts alone. These enrichment baselines try to assess the affinity of a molecule based on some assumed functional form. For instance, the **Diff Enrichment** score makes the assumption that there is a simple additive effect between the matrix and target counts. Poisson enrichment is taken from Gerry et al. [2019], which computes a maximum likelihood Poisson distribution for the target and matrix counts and then computes a ratio of the target at the lower 95% confidence interval (CI) and the matrix at the upper 95% CI. This baseline is perhaps closest to our model, however, our model relates the two distributions directly, whereas this score does not directly correlate the two values.

- **Diff Enrichment**: score $= \frac{1}{n_t}\sum_i c_t^i - \frac{1}{n_m}\sum_j c_m^j$
- **Ratio Enrichment**: score $= \big[\big(\frac{1}{n_t}\sum_i c_t^i\big) + 1\big] / \big[\big(\frac{1}{n_m}\sum_j c_m^j\big) + 1\big]$
- **Poisson Enrichment**: score $= \text{CI}_{\text{lower95}}[\text{Poisson}(\lambda_t)] / \text{CI}_{\text{upper95}}[\text{Poisson}(\lambda_m)]$

Since these baselines are not trained models, but rather explicit functions of the count data, we cannot compare these metrics against our model in terms of predicted likelihood. However, all methods provide a ranking of the test molecules, from which we can compute the aforementioned multi-class PR-AUC. We compare the performance of these models to DEL-Compose on random and synthon splits in Table 1. "Pre" and "Rep" refer to using the pre-selection and replicate factors

	Random Split	Matrix NLL ↓	Target NLL ↓	Sum NLL ↓	PRC-AUC ↑
CA-IX	Diff Enrichment	-	-	-	0.23 ± 0.01
	Ratio Enrichment	-	-	-	0.26 ± 0.02
	Poisson Enrichment	-	-	-	0.25 ± 0.01
	DEL-Compose	3.17 ± 0.03	2.82 ± 0.04	5.99 ± 0.06	0.28 ± 0.33
	DEL-Compose (Pre)	2.97 ± 0.01	2.80 ± 0.02	5.77 ± 0.02	$\mathbf{0.90 \pm 0.03}$
	DEL-Compose (Rep)	3.13 ± 0.03	2.65 ± 0.03	5.78 ± 0.06	0.73 ± 0.34
	DEL-Compose (Pre+Rep)	$\mathbf{2.96 \pm 0.02}$	$\mathbf{2.65 \pm 0.01}$	$\mathbf{5.61 \pm 0.01}$	0.84 ± 0.07
HRP	Diff Enrichment	-	-	-	0.59 ± 0.01
	Ratio Enrichment	-	-	-	0.48 ± 0.01
	Poisson Enrichment	-	-	-	0.57 ± 0.01
	DEL-Compose	6.51 ± 0.11	5.61 ± 0.09	12.12 ± 0.19	0.80 ± 0.12
	DEL-Compose (Pre)	6.30 ± 0.03	5.35 ± 0.02	11.65 ± 0.04	0.78 ± 0.03
	DEL-Compose (Rep)	6.39 ± 0.04	5.53 ± 0.04	11.92 ± 0.08	$\mathbf{0.81 \pm 0.04}$
	DEL-Compose (Pre+Rep)	$\mathbf{6.23 \pm 0.02}$	$\mathbf{5.30 \pm 0.02}$	$\mathbf{11.54 \pm 0.03}$	0.80 ± 0.04

Table 1: Metrics for different variants of DEL-Compose compared to baseline enrichment metrics on **random splits**. Metrics are averaged across the test set over 5 different splits.
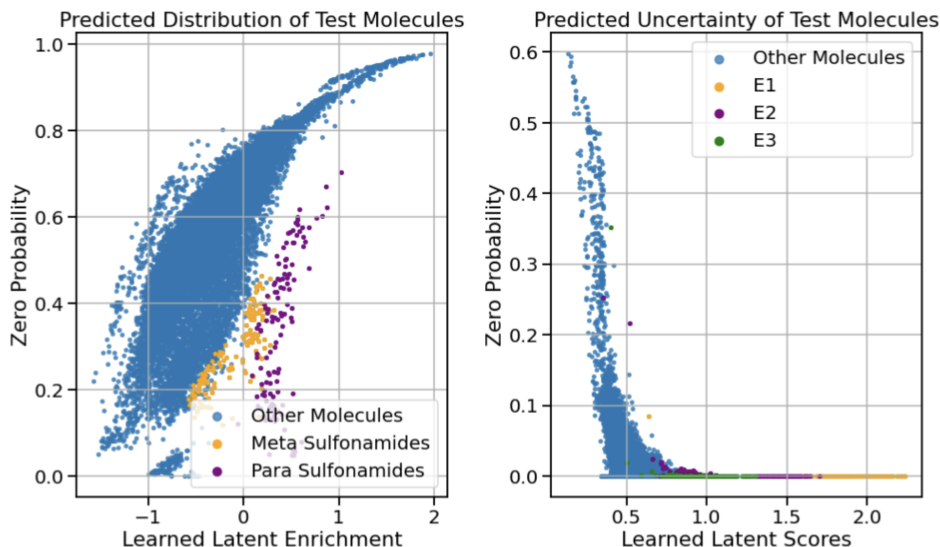

Figure 4: Predicted zero-probability is a good measure of predicted data noise for CA-IX (left) and HRP (right)

in the construction of the output distribution for DEL-Compose. In terms of likelihood, we see that our model that incorporates both pre-selection and replicate factors outperforms all ablations, which validates our intuition that these are important considerations when trying to model DEL data.

Comparing the enrichment baselines to the results of DEL-Compose, we notice that our model outperforms the baselines in terms of multi-class PR-AUC. It is important to note that the baseline metrics do not incorporate the pre-selection data, but even our factorized models without using the pre-selection counts outperform these baselines in most cases. Since the enrichment baselines have oracle access to the actual data, this suggests that DEL-Compose is capturing important aspects of the chemical data. We have not included models learned on top of these computed enrichment scores (which several previous works have proposed), as our model can outperform these oracle metrics already. Additional experiments showing generalizability of our model is shown in Appendix B.

3.2 DEL-Compose facilitates structured interpretation of the data

Due to its modeling structure, DEL-Compose provides good interpretability and insights to the model–which is ultimately useful for the chemist using this model. In Figure 4, we have plotted the
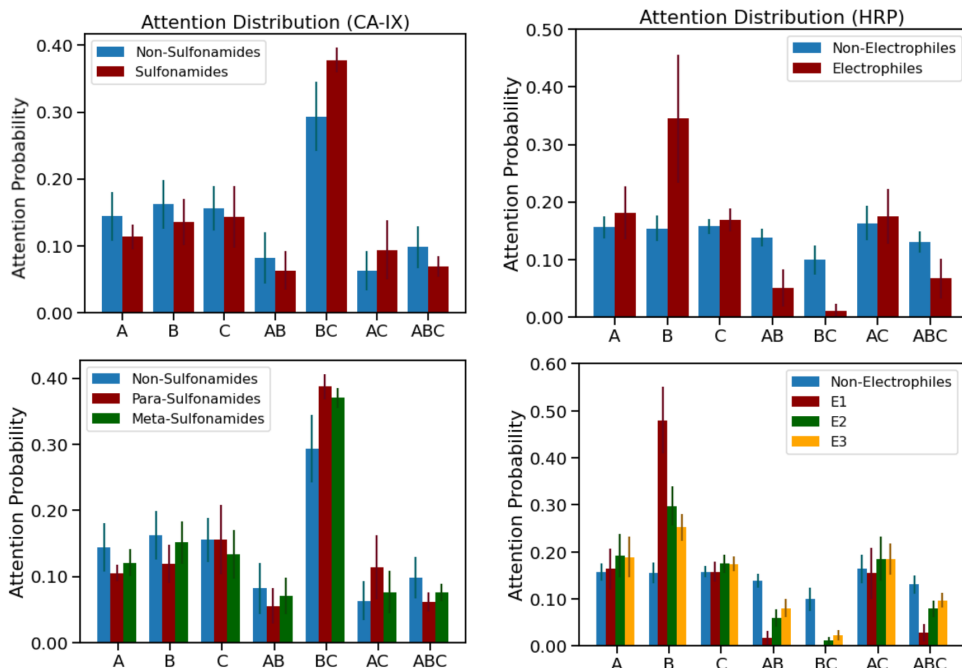
Figure 5: Attention distribution of synthon embeddings. The top set of plots aggregate all the known pharmacophores together, while the bottom set of plots separates out each class of molecules. For CA-IX, the highest attention weights are attributed to the di-synthon x_{bc}, while for HRP, the highest attention weights are attributed to the mono-synthon x_b.

learned latent scores λ as a function of the predicted zero probability p. For HRP, we see that all the molecules with the known pharmacophores have high predicted scores and low zero-probability–the signal is strong and the noise is low. However, when we turn our attention to the plot for CA-IX, we see that there are a number of molecules with high learned scores, but also high zero-probability–this region of the distribution likely contains more noise. Compared to the HRP data, the predicted scores for the benzene-sulfonamide containing molecules for CA-IX have some uncertainty associated, as implicated by their predicted zero-probabilities. Additionally, our model has nice interpretability with respect to the attention module over the synthons, demonstrating that our model correctly picks out the important synthons.

In Figure 5, we plot the attention distribution over the set of synthon representations of DEL-Compose. The top set of plots aggregate known pharmacophores, while the bottom set of plots delineate individual substructure sets. For CA-IX, we see that the attention probabilities are primarily focused on di-synthon x_{bc}, while for HRP, the attention probabilities are primarily distributed on the mono-synthon x_b. The sulfonamides are on the C position for CA-IX, while the electrophiles are on the B position for HRP, indicating that both models chose the highest weight to be placed on synthon embeddings incorporating the correct synthon position. Moreover, the weights on the important mono/di-synthons are higher for the molecules with the known pharmacophores versus other molecules in the library. Interestingly, the model for CA-IX chooses the di-synthon x_{bc}, instead of just the mono-synthon at the C position, which might indicate enrichment of certain di-synthons.

4   Conclusion

In our work, we proposed a novel method for representing DEL molecules leveraging their combinatorial nature. By incorporating the important experimental factors into our probabilistic model, we demonstrate the ability for our model to pick out the substructures important for particular proteins of interest, CA-IX and HRP. While the model learns useful latent variables that correlate to actual binding properties, we show that our model can also provide interpretable insights for the binding problem.

5 Appendix

A Data

We conduct experiments on public DEL data from Gerry et al. [2019], which includes DEL selection data on two well-studied protein targets: carbonic anhydrase IX (CA-IX) and horseradish peroxidase (HRP). This DEL is a tri-synthon library, consisting of 8 synthons at the A position, 114 synthons at the B position, and 118 synthons at the C position (107,616 total molecules) chosen to encourage chemical diversity of the library. Their data consists of two experimental conditions, one with the protein target, and one matrix condition that is conducted in the absence of the protein as control. For CA-IX this dataset includes 2 replicates of matrix data, and 4 replicates of on-target protein data; while for HRP, this dataset includes 4 replicates of matrix data and 2 replicates of on-target protein data. Additionally, there is data collected on the pre-selection library, which is an indicator of the relative abundance of the different DEL members.

Both CA-IX and HRP are proteins with known pharmacophores Gerry et al. [2019]. CA-IX has a well-known binding motif, benzene-sulfonamide Buller et al. [2011]. In this dataset, there are two synthons at the C position that includes benzene-sulfonamides, one that is meta-substituted with respect to the aryl group, and the other which is para-substituted. Studies have shown that the benzene-sulfonamide substituted at the para position is much more highly active, in general, towards CA-IX protein Pagnozzi et al. [2022]. Meanwhile, HRP is a enzyme with high affinity for compounds containing sulfonyl chloride–derived Michael acceptors Gerry et al. [2019], Gan et al. [2013]. In this dataset, there are three such synthons at the B position that shows high activity, and are the three synthons we treat as "gold" labels for HRP. These structures are all visualized in Figure 6 in descending order of affinity.
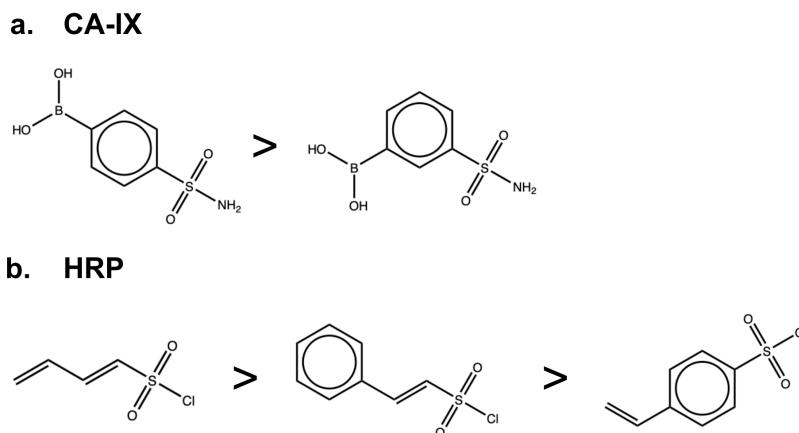

Figure 6: Known pharmacophores for both CA-IX (top) and HRP (bottom) pictured in descending order of affinity. For CA-IX, benzene-sulfonamides are known structures to induce affinity. The substitution of the sulfonamide affects the reactivity of the chemical specie, wherein the para-substituted constituent is found to be much more active. For HRP, electrophilic Michael acceptors are known pharmacophores.

B Generalization Experiments

Random splits are not always ideal for testing molecule datasets **??**. In order to test the generalizability of molecule representations, many approaches attempt to split molecule by molecular scaffolds Wu et al. [2018]. For DELs, rather than generic molecular scaffolding strategies, synthons provide a natural grouping and separation of the chemical space. By using synthons to split the data, we can test the generalizability of the model on unseen chemical structures. In the dataset that we are using, the known pharmacophores are conveniently localized to specific synthons, so we can develop intuitive splitting strategies. Most of the signal is captured by these pharmacophores, so we cannot withhold all of these molecules from training. Instead, we split on the synthon position that does not include

these individual pharmacophores. Specifically, for CA-IX, the benzene-sulfonamides are at the C position, so we create synthon splits by splitting on the B position. For HRP, the electrophilic Michael acceptors are at the B position, so we split the data at the C position. Results are shown in Table 2.

B.1 Setup for data-efficiency generalization

To understand more about our models, we introduce a third setup that tests the ability of the model to adapt under low-resource regimes. Since most of the signal resides in the molecules with known pharmacophores for their respective targets, we investigate the performance of our model when we change the amount of data provided to it. This will allow us to determine the quantity of data required to learn a reasonable model. In particular, these experiments provide a good way to compare different representational modalities, as we expect that our factorized approach should learn faster under resource-limited regimes.

	Synthon Split	Matrix NLL ↓	Target NLL ↓	Sum NLL ↓	PRC-AUC ↑
CA-IX	Diff Enrichment	-	-	-	0.22 ± 0.02
	Ratio Enrichment	-	-	-	0.25 ± 0.02
	Poisson Enrichment	-	-	-	0.24 ± 0.02
	DEL-Compose	3.58 ± 0.34	2.83 ± 0.17	6.41 ± 0.46	0.13 ± 0.01
	DEL-Compose (Pre)	3.13 ± 0.13	2.81 ± 0.18	5.94 ± 0.31	0.63 ± 0.42
	DEL-Compose (Rep)	3.54 ± 0.29	2.64 ± 0.14	6.19 ± 0.39	0.73 ± 0.33
	DEL-Compose (Pre+Rep)	$\mathbf{3.11 \pm 0.11}$	$\mathbf{2.61 \pm 0.14}$	$\mathbf{5.72 \pm 0.25}$	$\mathbf{0.75 \pm 0.33}$
HRP	Diff Enrichment	-	-	-	0.59 ± 0.01
	Ratio Enrichment	-	-	-	0.48 ± 0.01
	Poisson Enrichment	-	-	-	0.56 ± 0.01
	DEL-Compose	8.09 ± 2.04	7.36 ± 1.28	15.45 ± 3.32	0.85 ± 0.16
	DEL-Compose (Pre)	7.44 ± 2.09	6.46 ± 1.35	13.90 ± 3.44	0.85 ± 0.13
	DEL-Compose (Rep)	8.75 ± 3.73	7.36 ± 1.69	16.11 ± 5.41	0.88 ± 0.05
	DEL-Compose (Pre+Rep)	$\mathbf{7.29 \pm 2.11}$	$\mathbf{6.26 \pm 1.26}$	$\mathbf{13.55 \pm 3.36}$	$\mathbf{0.90 \pm 0.06}$

Table 2: Metrics for different variants of DEL-Compose compared to baseline enrichment metrics on **synthon splits**. Compared to random splits, the average NLL loss is higher, which confirms our belief that these are more challenging splits of the data for a model to learn over. Our model still outperforms baselines and ablations even on these more challenging splits.
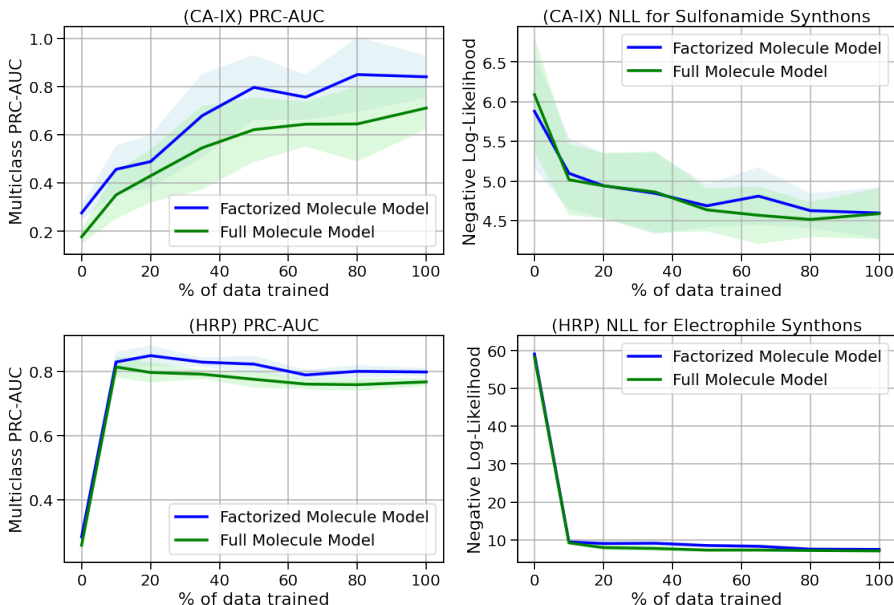

Figure 7: Performance of models using factorized representations vs using full molecule representations. Each model is trained with a different % of the data heldout.

11

One of the main benefits of utilizing a factorized model such as DEL-Compose is that we can avoid building complex enumeration engines for DELs, because we do not require the enumerated full molecule structure. However, while this is beneficial, we want to demonstrate that our factorized models can perform competitively, or even better than models that utilize full molecule representations. To do so, we conduct an in-depth investigation by training both versions of the model under different data-limiting regimes. In Figure 7, we compare the performance of both models as a function of amount of data supplied during training. For both CA-IX and HRP, we notice that the multi-class PR-AUC is superior for the factorized model compared to the full model at each point. Meanwhile, the test likelihoods for the two models are very comparable as a function of the amount of data supplied.

These results support the use of factorized representations as a useful inductive bias to achieve more efficient learning. However, the results are also unsurprising in some sense. The pharmacophores that describe the molecule classes we use are localized within specific synthons, but this property may not hold true for an arbitrary protein. Therefore, we expect that this model might require other improvements or regularizations for more challenging targets and data.

References

[1] Polina Binder, Meghan Lawler, LaShadric Grady, Neil Carlson, Sumudu Leelananda, Svetlana Belyanskaya, Joe Franklin, Nicolas Tilmans, and Henri Palacci. Partial product aware machine learning on dna-encoded libraries. *arXiv preprint arXiv:2205.08020*, 2022.

[2] Fabian Buller, Martina Steiner, Katharina Frey, Dennis Mircsof, Jorg Scheuermann, Markus Kalisch, Peter Buhlmann, Claudiu T Supuran, and Dario Neri. Selection of carbonic anhydrase ix inhibitors from one million dna-encoded compounds. *ACS chemical biology*, 6(4):336–344, 2011.

[3] John C Faver, Kevin Riehle, David R Lancia Jr, Jared BJ Milbank, Christopher S Kollmann, Nicholas Simmons, Zhifeng Yu, and Martin M Matzuk. Quantitative comparison of enrichment from dna-encoded chemical library selections. *ACS combinatorial science*, 21(2):75–82, 2019.

[4] Fei-Fei Gan, Kamila K Kaminska, Hong Yang, Chin-Yee Liew, Pay-Chin Leow, Choon-Leng So, Lan NL Tu, Amrita Roy, Chun-Wei Yap, Tse-Siang Kang, et al. Identification of michael acceptor-centric pharmacophores with substituents that yield strong thioredoxin reductase inhibitory character correlated to antiproliferative activity. *Antioxidants & redox signaling*, 19 (11):1149–1165, 2013.

[5] Christopher J Gerry, Mathias J Wawer, Paul A Clemons, and Stuart L Schreiber. Dna barcoding a complete matrix of stereoisomeric small molecules. *Journal of the American Chemical Society*, 141(26):10225–10235, 2019.

[6] Adrián Gironda-Martínez, Etienne J Donckele, Florent Samain, and Dario Neri. Dna-encoded chemical libraries: a comprehensive review with succeful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279, 2021.

[7] Robert A Goodnow Jr, Christoph E Dumelin, and Anthony D Keefe. Dna-encoded chemistry: enabling the deeper sampling of chemical space. *Nature Reviews Drug Discovery*, 16(2): 131–147, 2017.

[8] Letian Kuai, Thomas O'Keeffe, and Christopher Arico-Muendel. Randomness in dna en-coded library selection data can be modeled for more reliable enrichment calculation. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 23(5):405–416, 2018.

[9] Katherine S Lim, Andrew G Reidenbach, Bruce K Hua, Jeremy W Mason, Christopher J Gerry, Paul A Clemons, and Connor W Coley. Machine learning on dna-encoded library count data using an uncertainty-aware probabilistic loss function. *Journal of Chemical Information and Modeling*, 62(10):2316–2331, 2022.

[10] Ralph Ma, Gabriel HS Dreiman, Fiorella Ruggiu, Adam Joseph Riesselman, Bowen Liu, Keith James, Mohammad Sultan, and Daphne Koller. Regression modeling on dna encoded libraries. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[11] Daniel Madsen, Carlos Azevedo, Iolanda Micco, Lars Kolster Petersen, and Nils Jakob Vest Hansen. An overview of dna-encoded libraries: A versatile tool for drug discovery. *Progress in Medicinal Chemistry*, 59:181–249, 2020.

[12] Kevin McCloskey, Eric A Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A Clark, et al. Machine learning on dna-encoded libraries: a new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866, 2020.

[13] Dario Neri and Richard A Lerner. Dna-encoded chemical libraries: a selection system based on endowing organic compounds with amplifiable information. *Annual review of biochemistry*, 87: 479–502, 2018.

[14] Daniela Pagnozzi, Nicolino Pala, Grazia Biosa, Roberto Dallocchio, Alessandro Dessì, Pankaj Kumar Singh, Dominga Rogolino, Anna Di Fiore, Giuseppina De Simone, Claudiu T Supuran, et al. Interaction studies between carbonic anhydrase and a sulfonamide inhibitor by experimental and theoretical approaches. *ACS Medicinal Chemistry Letters*, 13(2):271–277, 2022.

[15] Alexander A Peterson and David R Liu. Small-molecule discovery through dna-encoded libraries. *Nature Reviews Drug Discovery*, pages 1–24, 2023.

[16] Christopher A Reiher, David P Schuman, Nicholas Simmons, and Scott E Wolkenberg. Trends in hit-to-lead optimization following dna-encoded library screens. *ACS Medicinal Chemistry Letters*, 12(3):343–350, 2021.

[17] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[18] Alexander L Satz, Andreas Brunschweiger, Mark E Flanagan, Andreas Gloger, Nils JV Hansen, Letian Kuai, Verena BK Kunig, Xiaojie Lu, Daniel Madsen, Lisa A Marcaurelle, et al. Dna-encoded chemical libraries. *Nature Reviews Methods Primers*, 2(1):3, 2022.

[19] Kirill Shmilovich, Benson Chen, Theofanis Karaletsos, and Mohammad M Sultan. Del-dock: Molecular docking-enabled modeling of dna-encoded libraries. *Journal of Chemical Information and Modeling*, 63(9):2719–2727, 2023.

[20] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[21] Lik Hang Yuen and Raphael M Franzini. Achievements, challenges, and opportunities in dna-encoded library research: An academic point of view. *ChemBioChem*, 18(9):829–836, 2017.

[22] Chris Zhang, Mary Pitman, Anjali Dixit, Sumudu Leelananda, Henri Palacci, Meghan Lawler, Svetlana Belyanskaya, LaShadric Grady, Joe Franklin, Nicolas Tilmans, et al. Building block-based binding predictions for dna-encoded libraries. 2023.