Poisson approximation to PBD

To handle this kind of response variable which comes from 40 experiments, one reasonable way is to return a probability. We firstly denote the probability distribution over possible labels $p(y|x, D)$, x is the input vector of the different feature combinations, D is the training set.

Let dummy variable y_{ij} denote whether the adenosine i on the center of the DRACH was detected to be methylated (1 for methylated, 0 for unmethylated) under experiment j. p_{ij} is the unknown actual probability that the y_{ij} is 1 (methylated). There are a total of 40 CLIP-based experiments (seen as number S of events), we may empirically assume that each experiment j is an independent and repeatable experiment with the same p_{ij} for each specific site i, so it is precisely the binomial distribution $B(n, p_i)$. Since n is large enough and p_i is quite small for majority sites, each binomial distribution can be further approximated by Poisson distribution with expectation $n * p_i$. However, due to different cell lines together with technical artificial variance, it is not reasonable to regard each experiment j for a specific site as i.i.d (or equally likely to succeed), then S_i has the distribution sometimes called Poisson binomial distribution (PBD) where p_{ij} are not necessarily identical for each site i. Fortunately, Hodges and Le Cam provided (Hodges and Lecam, 1960) an approximation theorem that helps to explain how this situation can be well approximated by Poisson distribution (Detailed prove were presented in the Supplementary File S1):
We focus on random site i, let x_{ij} indicate the random variables that have Poisson distribution with $E(x_{ij}) = p_i$, following are the joint distribution of x_{ij} and y_{ij}.

For the additive property of Poisson variables, $T_i = sum(x_{ij})$ has the Poisson distribution. We aim to show that $S_i = sum(y_{ij})$ has a very similar distribution. We let

$$D_i = \sup_u |P(S_i \leq u) - P(T_i \leq u)| \qquad (1)$$

D denotes the maximum absolute difference between the cumulates of S and T, and what we want is to find the condition under which D is small.
It can be further narrowed down to:

$$D_i \leq 2 \sum_{j=1}^{n} {p_{ij}}^2 \qquad (2)$$

Then we denote the random variable $Z_i = X_i - Y_i$, $E(Z_i) = 0$, While

$$Var(Z_i) = E({Z_i}^2) = p_{ij}(1 - e^{-p_i}) + \sum_{k=2}^{\infty} k^2({p_{ij}}^k e^{-p_{ij}})/k! \leq 3{p_{ij}}^2$$

let $\sum Z_i = U_i$, then$E(U) = 0$, $Var(U) \leq 3\mu$
Here we introduce a to be any positive number a. Let $T_i = S_i + U_i \leq \nu - a$, we can further get:

$$D_i = \sup_{v} |P(S_i \le v) - P(T_i \le v)| \le \sup_{v} P(v \le T_i \le v + a) + P(|U_i| \ge a) \qquad (3)$$

By using Chebycheff inequality together with the upper bound of the T_i which is "$(1 + 1/12\lambda)/(2\pi\lambda)^{1/2}$ "(will not go into detail here), the flowing equation can prove:

$$D_i \le (3\mu/a^2) + (a + 1)(1 + 1/12\lambda_i)/(2\pi\lambda_i)^{1/2} \qquad (4)$$

Using (2) and (4) we can get the result：

$$D_i \le 3\sqrt[3]{a_i}$$

More specifically, in our situation, the above approximation theorem implies that maximum absolute difference D_i between the cumulative distributions of S_i and Poisson distribution $\wp(\sum p_{ij})$ tends to 0, as $\alpha_i = max\,\{p_{i1}, \dots, p_{in}\} \to 0$. Moreover, the approximation theorems also suggest that the condition that $\alpha_i \to 0$ is sufficient but not necessary for $D_i \to 0$, in another words, S_i will have approximately a Poisson distribution even if few of p_{ij} are quite large, provided these values contribute only a small part of the total $\sum pi$ which make this model more robust.

Reference:

Hodges, J.L., and Lecam, L. (1960). The Poisson Approximation to the Poisson Binomial-Distribution. *Annals of Mathematical Statistics* 31, 737-740.