For each discrete feature $d$, we calculate the probability $p\_{d}$ that it overlaps with common polymorphisms. We then calculate the information content to denote its weighted value $w\_{d}=1+p\_{d}\*log\_{2}p\_{d}+(1-p\_{d}) \*log\_{2}(1-p\_{d})$.

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in natural polymorphisms. Thus, one weight cannot suffice for varied feature values. For a continuous feature $c$, which is associated with a score $v\_{c}$, we will calculate feature weights for each$ v\_{c}$. In particular, we discretize at each value and compute $w\_{c}^{v\_{c}}$. Then we fit a smooth curve for all $v\_{c}$ to obtain continuous $w\_{c}^{v\_{c}}=1+p\_{c}^{\geq v\_{c}}\*log\_{2}p\_{c}^{\geq v\_{c}}+(1-p\_{c}^{\geq v\_{c}}) \*log\_{2}(1-p\_{c}^{\geq v\_{c}})$. When we evaluate the continuous featurefor a particular variant, we calculate its weighted value using the fitted function.

We score each variant by summing up the weighted values of all its features$ s=\sum\_{d}^{}w\_{d}+\sum\_{c}^{}w\_{c}^{v\_{c}}$. We will also consider the feature dependency structure when calculating the scores (e.g., removing redundant features or performing dimension reduction techniques).

**D-2-b Research plan for Aim 2**

**D-2-b-i Statistical framework for parameter tuning using Bayesian updates**

The initial feature weights $W$ ($w\_{1},w\_{2},…,w\_{m}$) (given $m$ number of features) assigned in D-1-b-v will be further optimized with newly available “gold standard” datasets. We plan to tune these parameters using an incremental Bayesian learning strategy. For a variant $v$, given feature values $F\_{v} (f\_{v, 1}, f\_{v, 2},... ,f\_{v, m})$, $W$can be rewritten as ($t\_{1}\left(f\_{v, 1}\right), t\_{2}\left(f\_{v, 2}\right),... ,t\_{m}(f\_{v, m})$), with functions $T$depicts the relationship between $W$and $F$. All parameters in$T$are the same for all variants. Given the eleVAR score $s$ (equation 3 in D-1-b-v), the probability that $v$ is functional ($y\_{v}=1$ designates a positive result, whereas $y\_{v}=0$ denotes a negative result) follows a logistic function $P(y\_{v}=1|s)= \frac{1}{1 + exp(-k \* (s-a))}$ ($k,a$ are scaling parameters). To update $W$(more specifically, the parameters in functions $T$) with training data $Y$, we implement Bayes’ rule:$ P(T|Y,F\_{V}) ∝ P(Y|T,F\_{V})P(T)$. The probability of observing $T$ (given $Y$ andfeature values $F\_{V} $corresponding to variants in $Y$) is proportional to the probability of observing $Y$given $T$ and $F\_{V}$, multiplied by the prior probability of $T$. Assuming independency between data points in $Y$, which can be achieved by proper training data construction, $P(Y|T,F\_{V})P(T)=\prod\_{i=1}^{n}P(y\_{i}|t\_{1},t\_{2},...,t\_{m}, f\_{i,1},f\_{i,2},...,f\_{i,m})P(t\_{1},t\_{2},...,t\_{m}) $, given $n$ observations in $Y$.

Using the training data, we will maximize this function to find the most probable functions $T$, and these will be used as our updated parameters. The updated $T$ will then be used as tuned parameters in eleVAR to prioritize variants. The procedure will be iterated in several rounds. In the first round of tuning, feature weights obtained in D-1-b-v will be used to construct priors $P(T)$. In subsequent rounds, the updated weights will be set as new priors.