**Question 1**

## Report

It is evident from the article that the several observations in regard to the experiment are not deviant with initially approved researches on tumor purity. We initially examined the relationships between the number of mutations detected by four mutation calling techniques as well as the associated tumor purities using gastric cancer as a model. Purities of stomach cancer tumors varied widely, ranging from 5% to 100%. Around 72% of gastric tumors had purity greater than 70%.

The results indicated a substantial positive correlation between the number of mutations detected by MuSE and SomaticSniper algorithms and tumor purity ($p = 1.21e-05$ for MuSE and $p = 1.64e-04$ for SomaticSniper). Similar results were reported for the number of mutations detected by the MuTect2 ($p = 1.01e-02$) and VarScan2 ($p = 6.13e-05$) algorithms, which incorporate tumor purity correction criteria. Notably, the considerably positive connection between the number of mutations and tumor purity was also detected in the other nine cancer types. These findings suggested that tumor purities may have a major effect on mutation discovery.

It is critical to note that when normal samples from multiple tissue types are combined, the diversity within the normal group may rise. This is why, in an earlier version of InfiniumPurify, we identified iDMCs using matched samples in each cancer type. However, rigorous data analysis reveals that mixing normal samples provides outcomes that are equivalent. This, we feel, is a more important tactic that will have a broader application; for instance, purity prediction can be conducted for malignancies not included in the TCGA. For every form of cancer, as long as even the sample size is sufficiently high (e.g. 20,) the iDMCs can be

consistently recognized and the purity determined by correlating the cancer to universal normal controls.

It is clear that tumor purity estimations display inherent features of the source data utilized to assess purity and have only a moderate correlation with other profiles. One possible explanation for these differences is the beginning tissue material used to examine the various locations of the tumor specimen. Estimates based on pathology are considered the top standard.

However, the interpathologist variation seen in this study, as well as previous research, implies that these estimations are likely to contain some mistakes due to their subjectivity. These inconsistencies may possibly be due to the pathologic slide's lack of complete spatial heterogeneity. For clinico-genomic sequencing investigations that require a certain level of purity for inclusion, an alternative to pathology estimations is to deduce purity explicitly from the analyte by doing moderate DNA sequencing to filter out low-purity samples.

Additionally, emphasis is on DM calling methods that require a larger sample size than minfi, limma, or equivalent tools. It is anticipated that InfiniumPurify would be used mostly for population-level research. Control-free DM calling additionally demands that the purities of the samples be sufficiently distributed to allow the statistical test to be done correctly. It is vital using the regulation DM calling method with prudence and using normal controls whenever possible.

**Question 2**

1)

$$\frac{d[E]}{dt} = k_2[ES] + k_3[ES] - k_1[E][S]$$

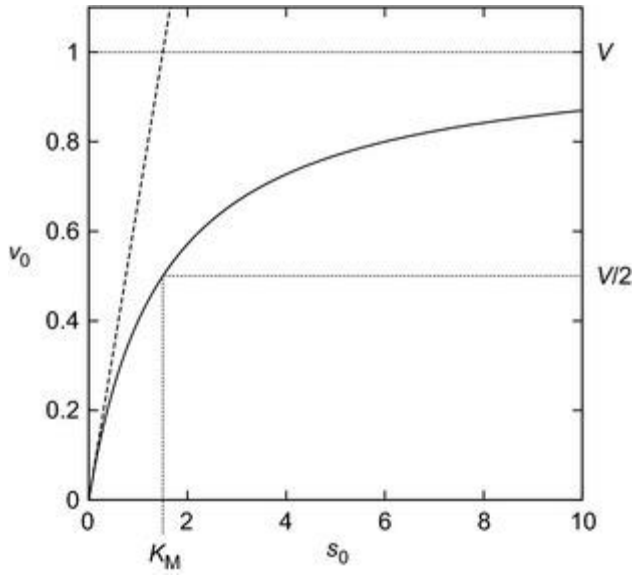
$$\frac{d[S]}{dt} = k_2[ES] - k_1[E][S]$$

$$\frac{d[ES]}{dt} = k_1[E][S] - k_2[ES] - k_3[ES]$$

$$\frac{d[P]}{dt} = k_3[ES]$$

(2) In codes

3)




**$V_{max}$= 0.5**

**Question 3**

## Report

Machine learning (ML) devices are increasingly being used in drug development to expedite target identification. These models are frequently evaluated using cross-validation approaches. The existence of data doppelgängers can impair the effectiveness of such validation approaches as shown from the article. When independently produced data are remarkably similar, designs operate well independently of how well they are trained. Despite their prevalence in genomics data and its inflationary implications, data doppelgängers remain uncharacterized. We demonstrate their frequency in biomedical data, how doppelgängers form, and the confounding consequences of doppelgängers. We advocate detecting data doppelgängers before to the training-validation split to mitigate the doppelgänger impact.

Although deleting data doppelgängers directly from data has proven difficult, we must still caution against doppelgänger consequences. The first piece of advice is to provides more effective cross-checks utilizing meta-data as a guide. We constructed negative and positive examples in this study using meta-data from RCC. This enabled us to forecast PPCC scores limits for situations in which doppelgängers are impossible to exist and leaking is possible . The possible data doppelgängers to be concerned about are those derived from the same class but distinct patients. We are able to identify probable doppelgängers and arrange them into validating sets using this meta-data information, essentially preventing doppelgänger effects and allowing for a more objective evaluation of machine learning performance.

Likewise, technical replicates derived from the same sample should be treated similarly. Rather than assessing performance of the models on the entire set of test data. We can divide data into strata based on their characteristics and data doppelgängers from sources other than the PPCC, and measure model performance separately for each stratum). Assuming that each stratum exists corresponds to a specified fraction of data. Despite the fact that we have a current population, we are nevertheless able to value the performance of in the actual world. When determining the performance of a classifier at a particular stratum, the classifier takes into account the stratum's real-world prevalence.

Significantly, strata with a faulty model performance identify classifier gaps. The non-PPCC doppelgängers in RCC utilized in stratified performance evaluations. The final guideline is to do highly thorough independent validation tests using as many data points as possible. The most often used strategy for removing confounding variables from each input variable prior to machine learning modeling is to regress the confounding variables independently from each input variable. However, we demonstrate that this strategy is insufficient since machine learning techniques can gain information from unregressed data.

Rather than removing confounding effects from individual input variables, we suggest post-hoc confounding control at the threshold of pattern recognition predictions. This enables the splitting of predictive performance into performance justified by mystifies and performance that is unaffected by confounds. This approach is adaptable and allows for modification of confounding variables in both parametric and non-parametric ways. We demonstrate that this strategy correctly accounts for influencing effects in real and well, even when conventional input variable adjustment gives false-positive findings.

Accuracy drops to 0.6 when all doppelgängers are included in the training set, which is the predicted accuracy for a trained model on random signatures. Obviously, the doppelgänger effect is abolished since all PPCC datasets doppelgängers are combined in the training set. This may be a way to circumvent the doppelgänger phenomenon. Confining the PPCC dataset doppelgängers to the restricting or validation sets, on the other hand, are suboptimal options. When the size of the training set is constant the former results in models that may not generalize well due to a lack of knowledge. In the latter case, you may encounter amazing winner-takes-all situations.

ML model results are typically evaluated by comparing the model's accuracy against validation data. This method of model validation is valid only if the validation data are distinct from the training data. However, such assertion is frequently made without prior verification. The study demonstrates that PPCC data doppelgängers serve as functional doppelgängers, resulting in inflationary consequences comparable to data leaking.The commonalities between doppelgänger effects and leakage were demonstrated in our experiment with k-nearest neighbor (kNN) models, where another training validation set with eight considered to be the “ in validation had an identical accuracy distribution to the training validation set with complete leakage . However, not all models are affected equally: kNN and naive bayes models exhibit a more direct linear link between performance hyperinflation and doppelgänger dosage than decision tree and logistic regression models do.

The above is widely held belief may be incorrect in the context of doppelgänger phenomena. We discover that doppelgängers are extremely prevalent in our test data and have a direct impact on prices on machine learning accuracy. As a result, the utility of machine learning for phenotypic analysis and subsequent identification of new pharmacological leads is

diminished. Additionally, we discovered that the magnitude of this inflationary impact varied according to two primary variables: the resemblance of practical doppelgängers and the fraction of functionality doppelgängers in the testing dataset. Regrettably, analytical resolution of doppelgänger effects is not straightforward. To prevent effectiveness inflation, it is critical to identify potential doppelgängers in training and validation data prior to assortment.