Гиперпараметры

- У модели есть параметры и гиперпараметры
- Параметры модели учатся на основе выборки самой моделью (алгоритмом ее обучения)
- Гиперпараметры - это параметры, которые задаем мы и которые влияют на то, как модель учит параметры

Примеры гиперпараметров?

Примеры гиперпараметров?

1. Регуляризация - какая регуляризация и с каким коэффициентом

2. Степень полинома, которым мы аппроксимировали функцию

3. Априор, который мы используем при оценке параметров байесовской модели

4. Параметр C в методе случайных векторов

5. Что-то еще?

Примеры гиперпараметров?

1. Регуляризация - какая регуляризация и с каким коэффициентом

2. Степень полинома, которым мы аппроксимировали функцию

3. Априор, который мы используем при оценке параметров байесовской модели

4. Параметр C в методе случайных векторов

5. Признаки, которые мы даем модели - тоже гиперпараметры!

Train-test split

Обучение (train) **Тест (test)**

Обучаем модель на train, проверяем качество модели на test.

Train-test split?

Обучение (train) **Тест (test)**

Обучаем модель на train, проверяем качество модели на test.

Как подбирать гиперпараметры модели? - Никак

Train-validation-test split!

Обучение (train) **Валидация (validation)** **Тест (test)**

1) **Выбираем некоторые значения гиперпараметров**
2) **Обучаем модель с такими гиперпараметрами на train**
3) **Смотрим качество на validation**
4) **Пробуем таким образом много разных значений гиперпараметров и выбираем то, которое дает наилучшее**

Train-validation-test split!

Обучение (train) **Валидация (validation)** **Тест (test)**

1) **Выбираем некоторые значения гиперпараметров**
2) **Обучаем модель с такими гиперпараметрами на train**
3) **Смотрим качество на validation**
4) **Пробуем таким образом много разных значений гиперпараметров и выбираем то, которое дает наилучшее**

Какие минусы подхода?

Train-validation-test split?

Обучение (train) **Валидация (validation)** **Тест (test)**

Какие минусы подхода?

1) **Существенно уменьшаем объем данных, на которых учится модель**
2) **Большая нестабильность оценки качества при сравнении моделей из-за малого размера выборки**

Кросс-валидация

Обучение (train)

Тест (test)

...

Много разбиений на train и вариацию. На каждом разбиении выбираем лучшие гиперпараметры. Потом смотрим, какие значения гиперпараметров встречаются чаще всего, на основании чего делаем вывод об итоговых значениях гиперпараметров

Что еще можно оценить?

Кросс-валидация

Обучение (train)

Тест (test)

...

Что еще можно оценить?

Для данного набора значений гиперпараметров можем оценить среднее качество модели и дисперсию по разным разбиениям

Кросс-валидация. Как разбивать?

Обучение (train)

Тест (test)

Кросс-валидация. Leave-one-out cross-validation

...

На каждой итерации в валидацию попадает ровно один объект. По остальным учимся

Кросс-валидация. Leave-one-out cross-validation

...

На каждой итерации в валидацию попадает ровно один объект. По остальным учимся

Какие минусы?

Кросс-валидация. Leave-one-out cross-validation

...

Какие минусы?

1) **Невозможно оценить некоторые метрики, подразумевающие, например, что в валидации у нас есть оба класса**
2) **Склонна завышать качество, так как хотя бы один похожий объект в обучении найдется**
3) **Есть формула для оценки качества, получаемого на leave-one-out кросс-валидации**

Кросс-валидация. Монте-карло кросс-валидация

...

На каждой итерации случайно выбираем какой-то процент объектов в валидацию

Кросс-валидация. Монте-карло кросс-валидация

...

На каждой итерации случайно выбираем какой-то процент объектов в валидацию

Какие минусы?

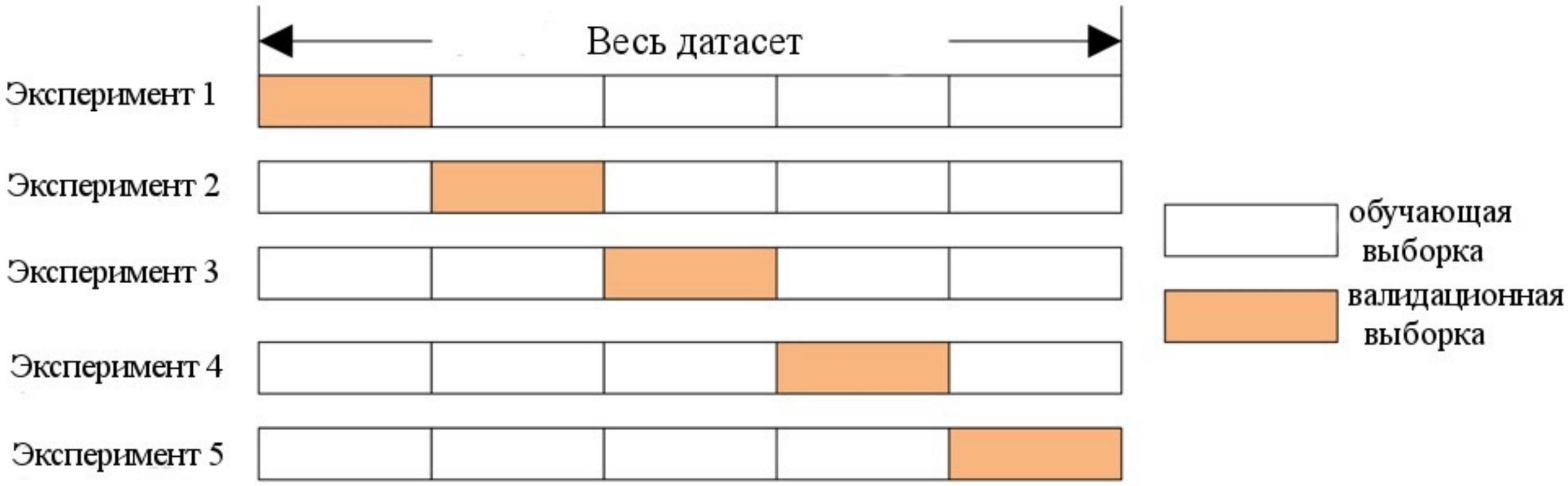
Кросс-валидация. Монте-карло кросс-валидация

Какие минусы?

1) **Нет гарантий, что все объекты побывают и в обучении, и в валидации**

Кросс-валидация. K-fold кросс-валидация

Почему картинка неверна?



Кросс-валидация. K-fold кросс-валидация

Тест отдельно должен быть

Обучающая выборка

Тест (test)

Кросс-валидация. K-fold кросс-валидация

Вся обучающая выборка

Какие минусы?

Тест (test)

Кросс-валидация. K-fold кросс-валидация

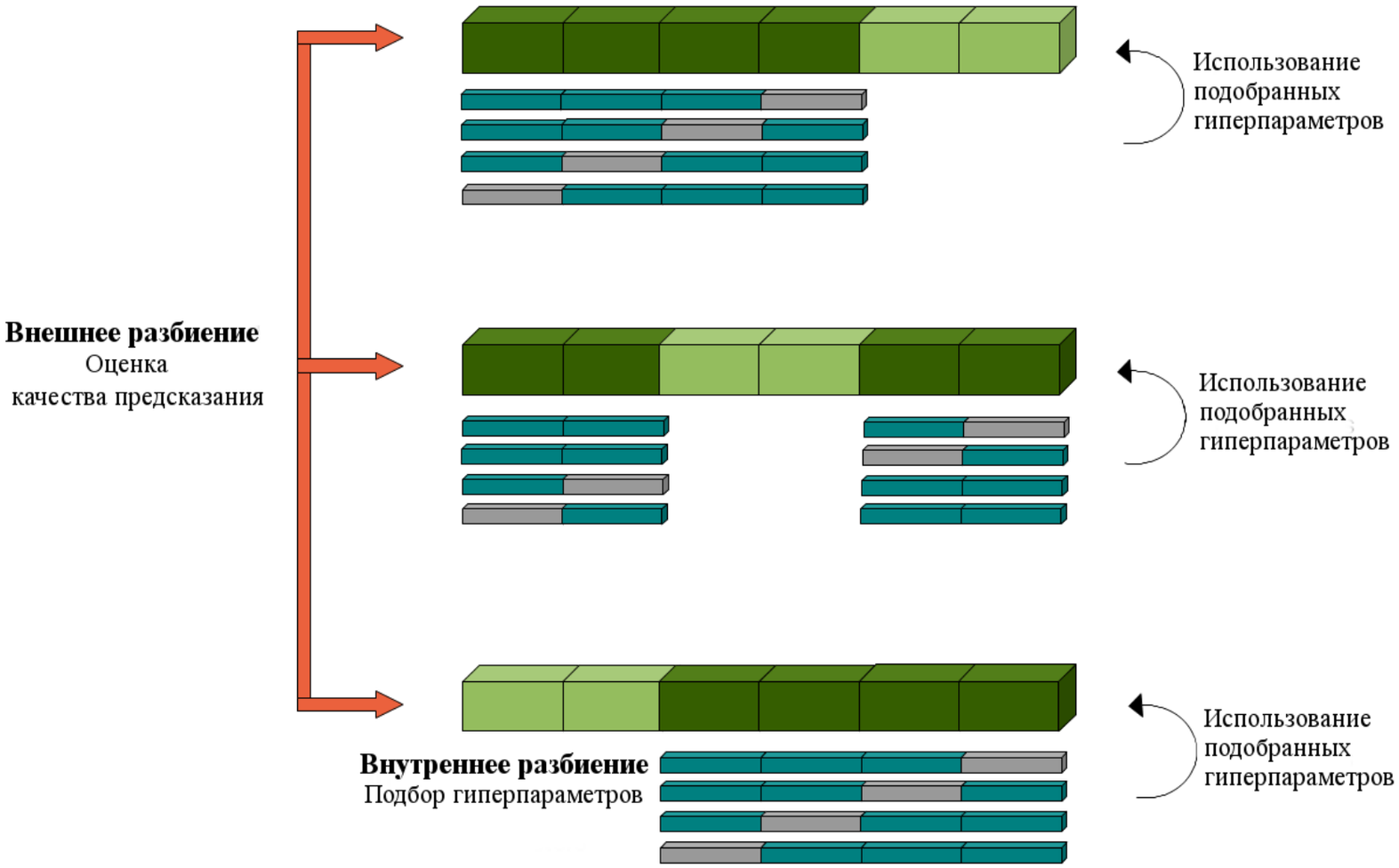
Вся обучающая выборка

Какие минусы?

1) **Не совсем понятно, сколько блоков брать**
2) ***хотелось бы не откусывать тест**

Тест (test)

Вложенная кросс-валидация



Все равно лучше иметь независимые данные для тестирования финальной модели

Как перебирать гиперпараметры?

1. Руками, на основе своих знаний о задаче и используемом алгоритме

2. GridSearch - задаем возможные значения каждого гиперпараметра, а потом проверяем все комбинации параметров

3. Random Search - задаем возможные значения каждого гиперпараметра вместе с вероятностями гиперпараметра принять те или иные значения. Далее много раз сэмплируем значения гиперпараметров и сравниваем модели между собой

4. Байесовская оптимизация -Random Search, но при выборе значений параметров для новой итерации используем знания о том, какие значенияприводили к моделям с бОльшим качеством

Кросс-валидация в биологии

Какие проблемы у любой предложенной валидации?

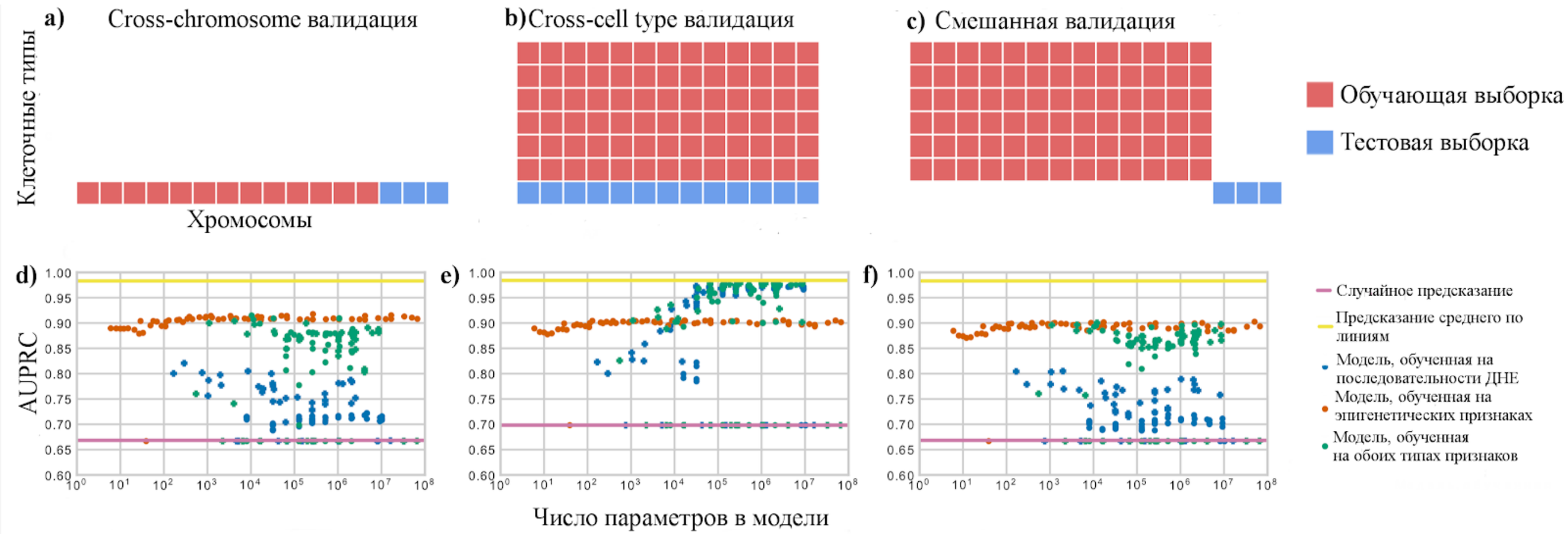
Кросс-валидация в биологии

Какие проблемы у любой предложенной валидации?

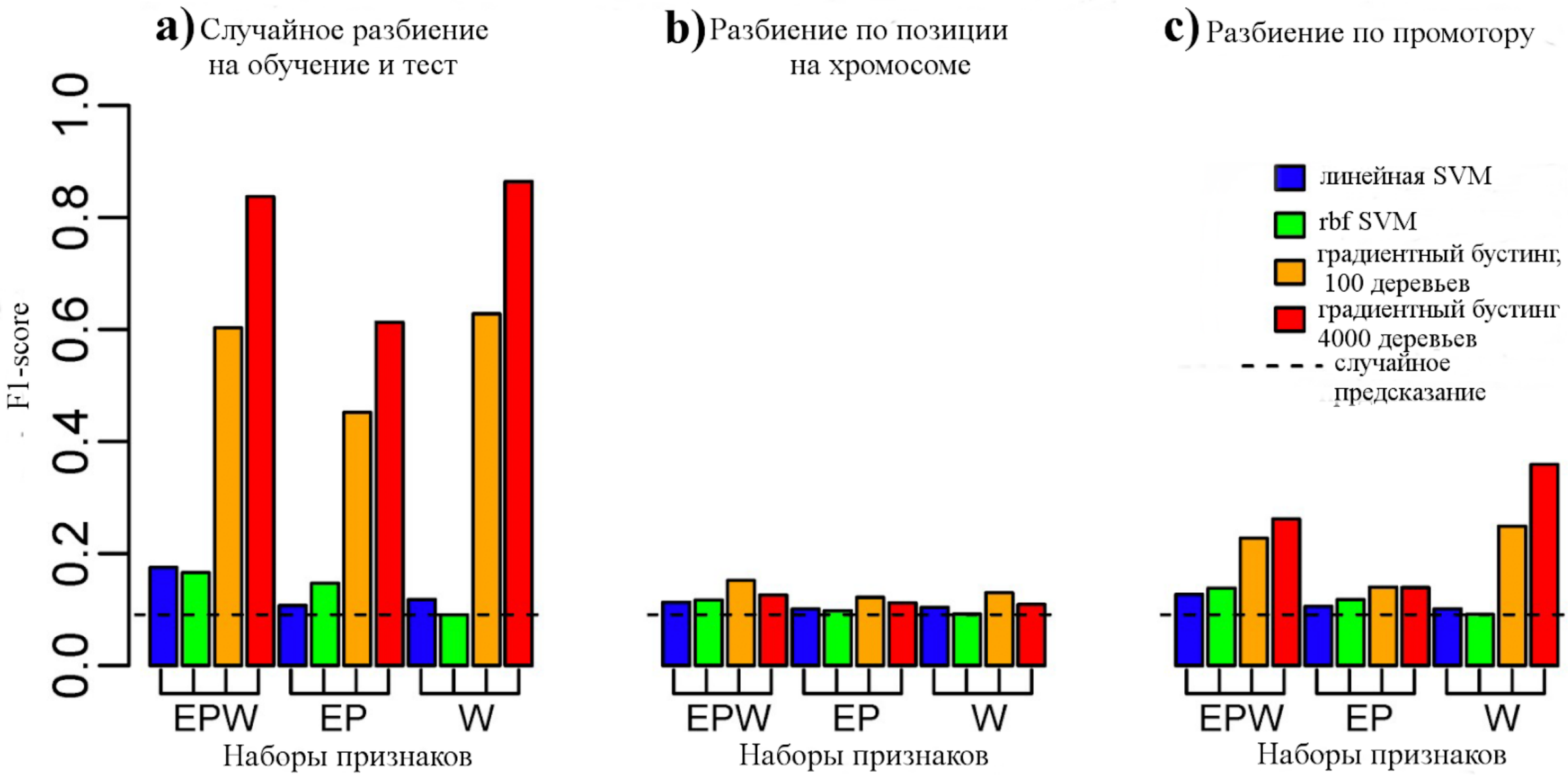
Она не учитывает домена, в котором мы работаем.

Для каждой задачи надо отдельно думать, как правильно сделать валидацию.

Кросс-валидация в биологии



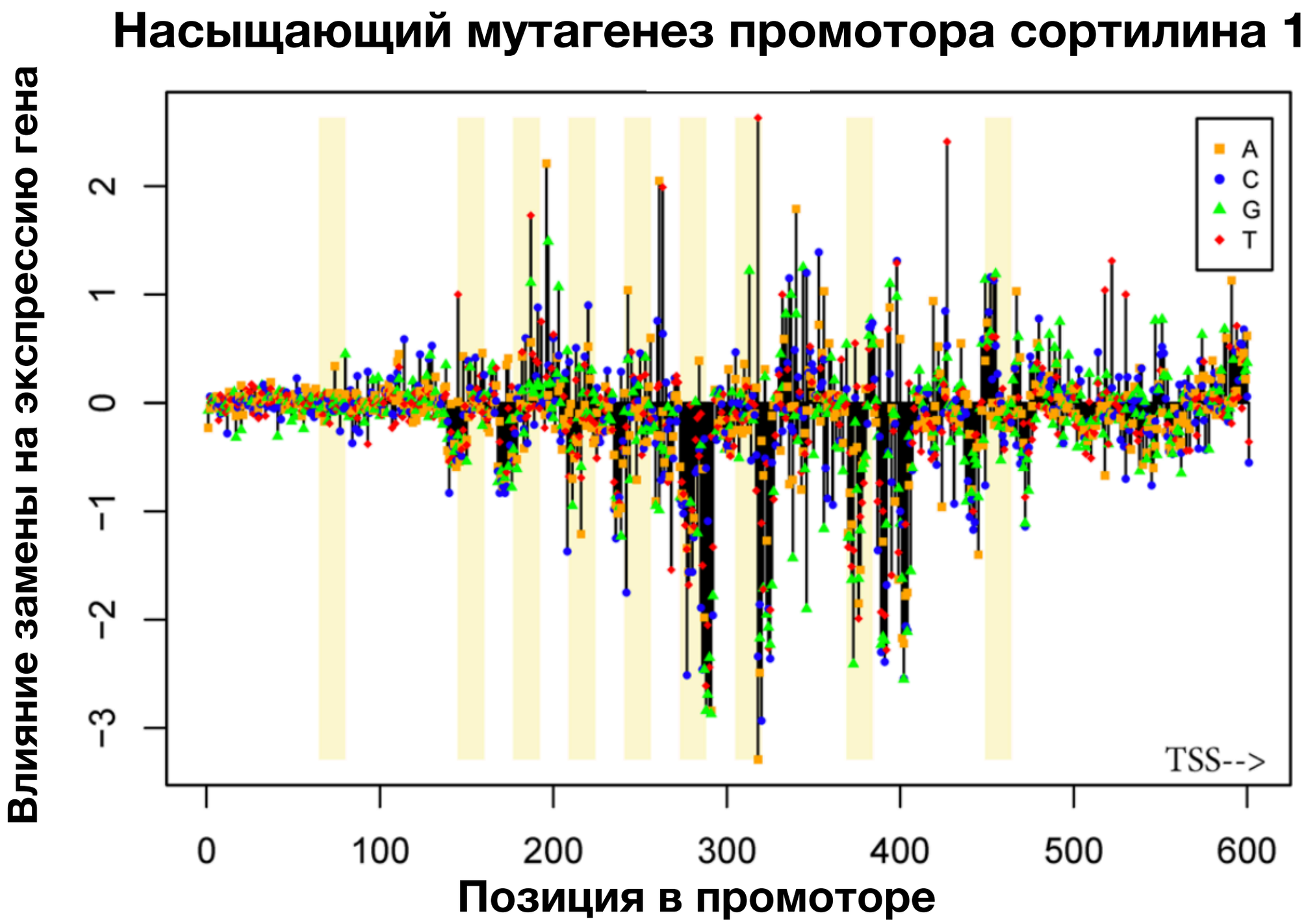
Кросс-валидация в биологии



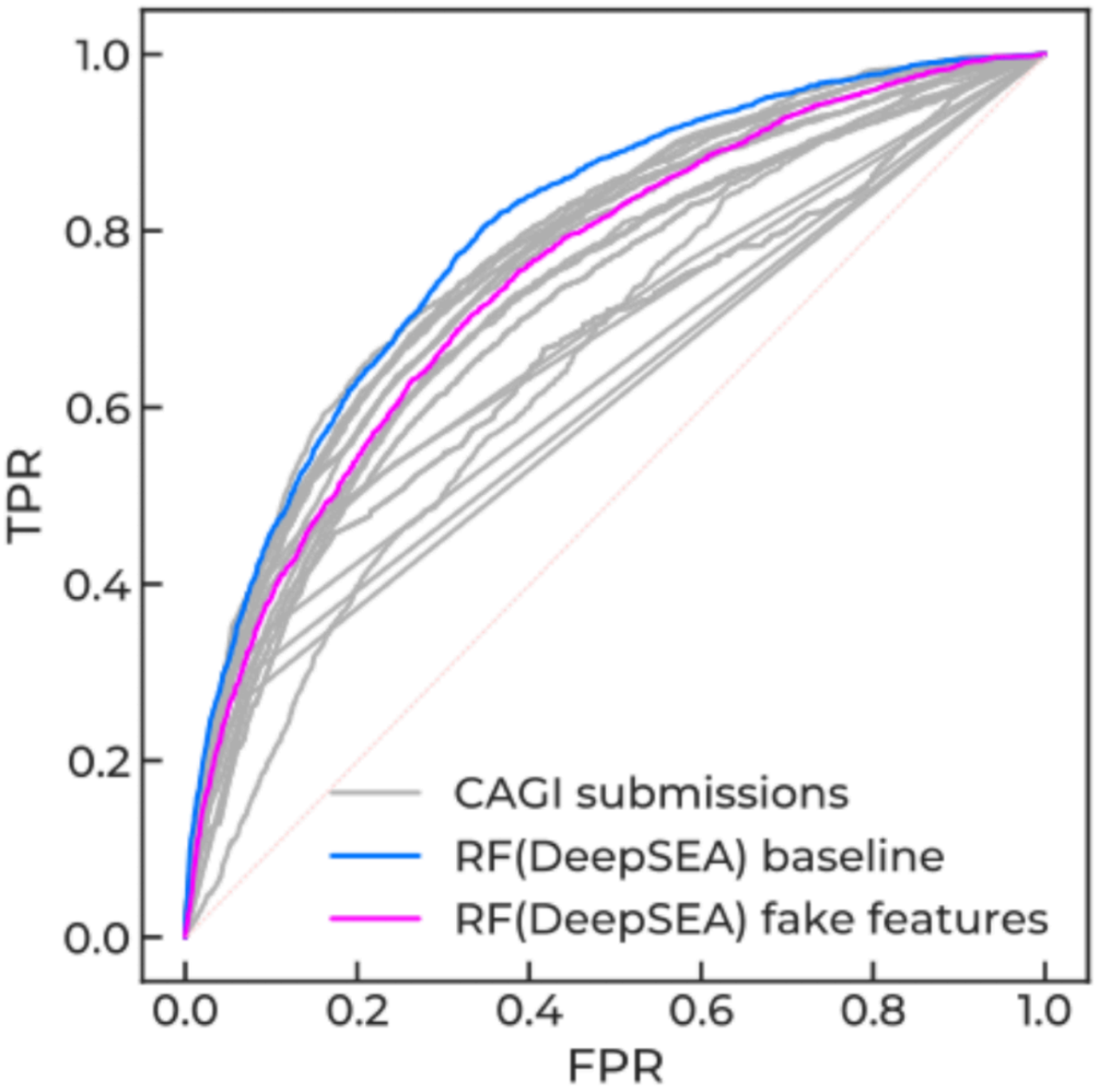
Кросс-валидация в биологии. Предсказание энергии связывания лиганда с белком

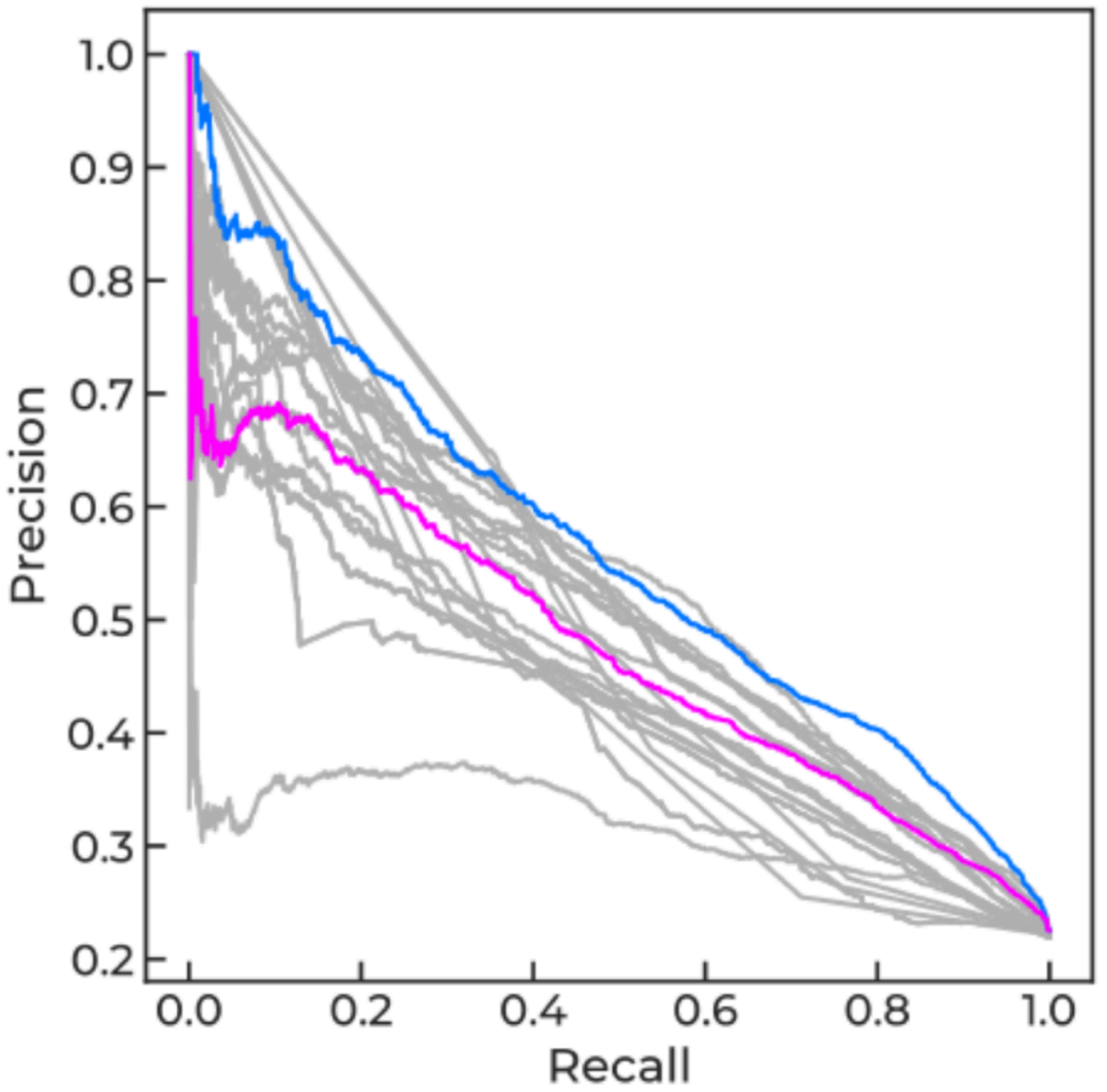
1) случайное;
2) разделение на основании сходства структур лигандов — группа лигандов с похожими структурами либо вся оказывается в обучении, либо вся - в тесте, причем в тесте оказываются группы меньшего размера;
3) разделение на основании энергии связывания — гарантируется, что и в обучении, и
в тесте будут комплексы из всего спектра силы связывания;
4) разделение на основании времени появления комплекс в базе данных PDB -
структуры, опубликованные до определенного года помещаются в обучающую
выборку, а остальные - в тестовую. Это позволяет оценить качество метода в предсказании данных будущих экспериментов;
5) разбиение на основании качества структур комплексов и их сходства

Shigaki et al., Hum Mutat. 2019 Sep;40(9):1280-1291. doi: 10.1002/humu.23797. Epub 2019 Jun 23. PMID: 31106481; PMCID: PMC6879779.

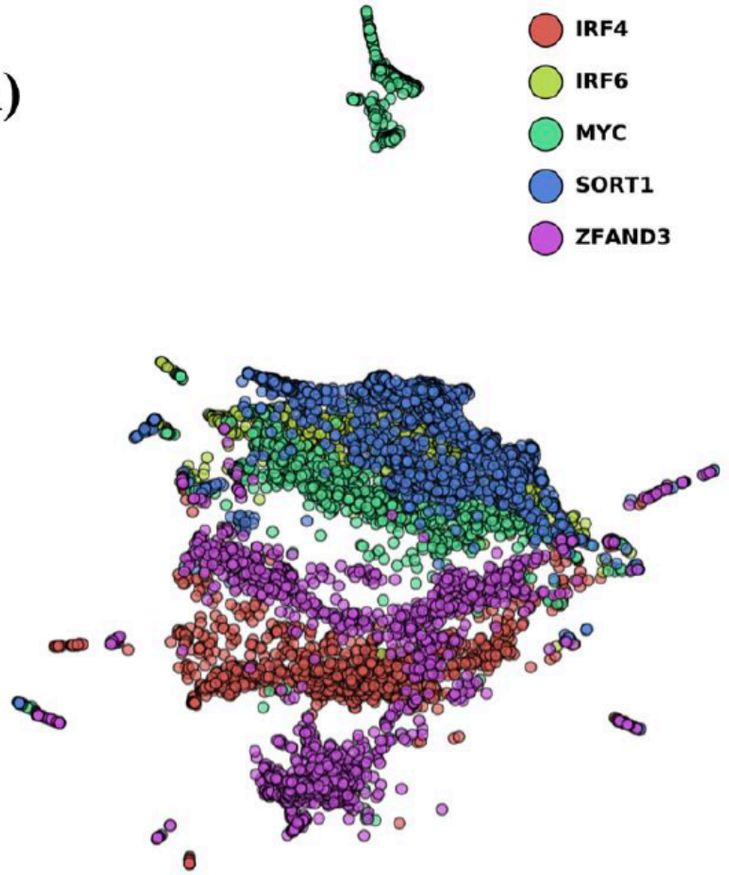


Penzar et al., *What Do Neighbors Tell About You: The Local Context of Cis-Regulatory Modules Complicates Prediction of Regulatory Variants.* Front Genet. 2019 Oct 31;10:1078. doi: 10.3389/fgene.2019.01078. PMID: 31737053; PMCID: PMC6834773.

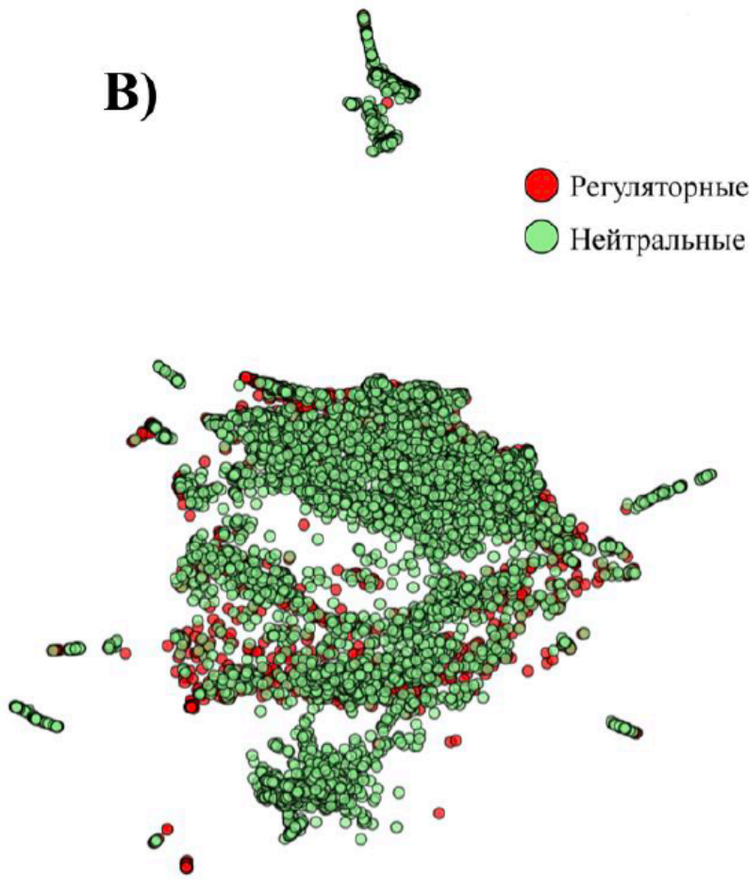




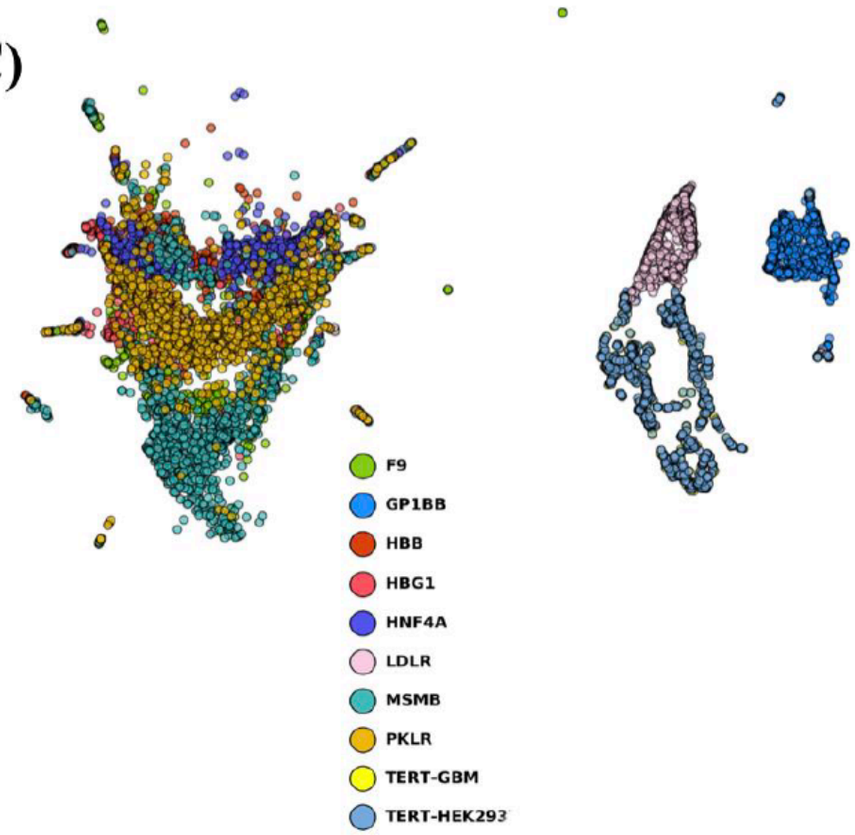
A)


IRF4
IRF6
MYC
SORT1
ZFAND3

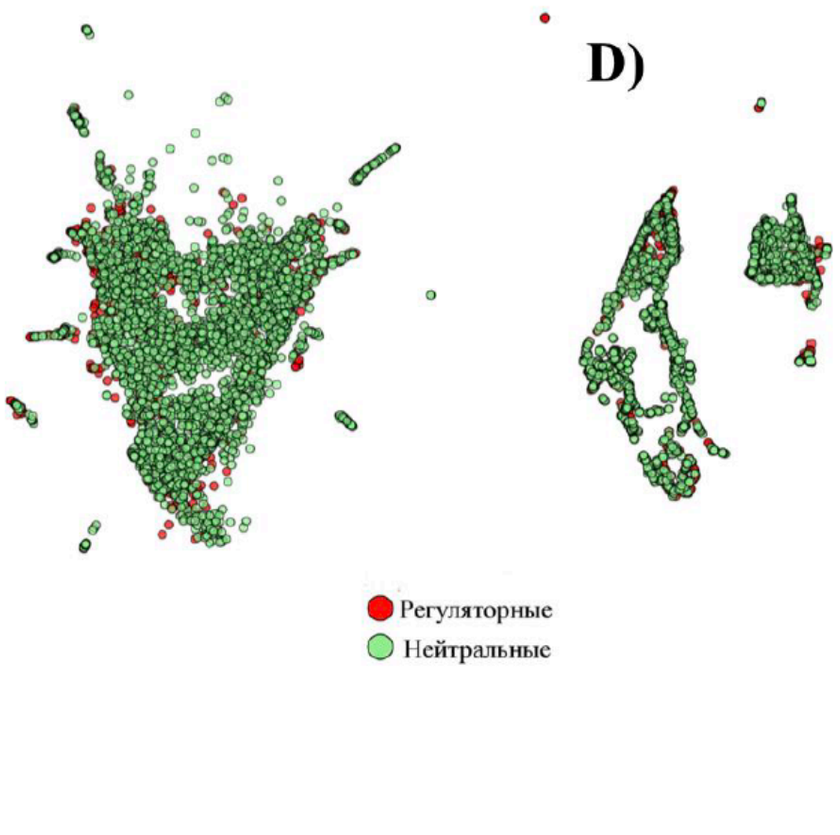
B)


Регуляторные
Нейтральные

C)


F9
GP1BB
HBB
HBG1
HNF4A
LDLR
MSMB
PKLR
TERT-GBM
TERT-HEK293

D)


Регуляторные
Нейтральные

How to win a Kaggle competition?

Anthony Goldbloom

"According to Anthony, in the history of Kaggle competitions, there are only two Machine Learning approaches that win competitions: **Handcrafted** & **Neural Networks**."

Где побеждают ансамбли деревьев решений?

- Recomendation systems (Netflix Prize 2009)
- Learning to rank (Yahoo Learning to rank challenge 2010)
- Crowdflower Search Results Relevance (2015)
- Avito Context Ad Clicks (2015)
- Везде :)

"As long as Kaggle has been around, Anthony says, it has almost always been **ensembles of decision trees that have won competitions.**"