

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ
ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ

"МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ИНСТИТУТ)"

ФИЗТЕХ-ШКОЛА БИОЛОГИЧЕСКОЙ И МЕДИЦИНСКОЙ
ФИЗИКИ

КАФЕДРА БИОИНФОРМАТИКИ И СИСТЕМНОЙ БИОЛОГИИ

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА ПО
НАПРАВЛЕНИЮ 03.03.01

ПРИКЛАДНАЯ МАТЕМАТИКА И ФИЗИКА

НА ТЕМУ:

**Разработка метода для предсказания синергетических
комбинаций малых молекул на основе данных RNA-seq**

Студент _____ Зиганшина Д. О.

Научный руководитель _____ Муртазалиева Х. А.

Зав. кафедрой, регент-профессор, PhD, _____ Бородовский М. Ю.

МОСКВА, 2020

1 Аннотация

Синергизм – тип взаимодействия между двумя или более химическими агентами, который характеризуется тем, что общий эффект лекарств превышает сумму индивидуальных эффектов каждого лекарства. Задача вычислительного предсказания синергетических эффектов остается нерешенной и актуальной во многих областях биомедицины (например, для подбора комбинаций противоопухолевых лекарств или предсказания комбинаций для клеточного перепрограммирования). В работе представлен новый подход для предсказания синергетических комбинаций на основе экспрессионных сигнатур и генных сетей.

Содержание

1	Аннотация	1
2	Обозначения, сокращения, основные определения	3
3	Введение в Хемоинформатику	4
3.1	Представление	4
3.1.1	Внутреннее представление	5
3.1.2	Внешнее представление	9
3.2	Молекулярные дескрипторы	11
3.3	Молекулярные отпечатки	11
3.4	Молекулярное подобие	12
4	Синергия	13
4.1	Введение	13
4.2	Область применения синергии	14
4.3	Предсказание синергетических пар соединений	15
4.3.1	Методы предсказания синергии, разработанные на основе наборов данных консорциума DREAM Challenges	24
4.3.2	Методы с использованием других датасетов	38
5	Материалы и методы	47
6	Полученные результаты	48
7	Заключение. План дальнейших исследований	49
8	Благодарности	50
9	Список литературы	51

2 Обозначения, сокращения, основные определения

Определение 2.1 (*Синергизм*).

Синергизм – тип взаимодействия между двумя или более химическими агентами, который характеризуется тем, что общий эффект лекарств превышает сумму индивидуальных эффектов каждого лекарства.

Определение 2.2 (*Хемоинформатика*).

Хемоинформатика это научная дисциплина на пересечении химии, информатики и математики. Основные задачи в области связаны с построением вычислительных методов для хранения и обработки химической информации и дизайна малых молекул с заданными свойствами.

Определение 2.3 (*Дифференциальная экспрессия генов*).

Явление дифференциальной экспрессии генов состоит в том, что экспрессия генов в клетках одного типа отличается от их экспрессии в клетках другого типа. Регуляция экспрессии генов может происходить на разных уровнях: репликации, транскрипции, трансляции, а также в процессе созревания иРНК и полипептидных цепей, образующихся в результате трансляции.

Определение 2.4 (*Quantitative Structure-Activity Relationship, (Q)SAR*).

Quantitative Structure-Activity Relationship - это одно из интенсивно развивающихся направлений использования математических методов в химии. Под QSAR подразумевается поиск зависимостей между структурой химических соединений и их свойствами. Также часто аббревиатуру QSAR используют для обозначения моделей, в основе которых лежит концепция связи "структура-свойство".

3 Введение в Хемоинформатику

Хемоинформатика, согласно определению данному Й. Гастайгером[1], - это применение методов информатики для решения химических задач. Одна из основополагающих задач в химии это создание соединений с заданными свойствами. Моделирование (количественных) соотношений "структура-активность" ((Q)SAR, Quantitative Structure-Activity Relationships) позволяет выявить взаимосвязь между структурой химических соединений и их активностью, чаще всего биологической. Модели SAR широко используются для виртуального скрининга при разработке лекарств с целью сокращения количества экспериментальных испытаний. Становление химии как науки привело к накоплению огромного количества данных, поэтому возникла необходимость ими оперировать: хранить информацию о миллионах химических соединениях и осуществлять быстрый поиск в этой информации. Более того, количество потенциальных химических соединений почти бесконечно, например, существует более 10^{29} возможных производных n-гексана со 150 заместителями[2]. Для работы с такими объемами информации требовалось создать машиночитаемое представление химических структур. Оно должно быть уникальным и однозначно интерпретируемым.

3.1 Представление

Представление структур химических соединений делится на 2 типа:

- внутреннее
- внешнее

3.1.1 Внутреннее представление

Когда говорят о внутреннем представлении структур химических соединений подразумевается машинное представление. Для внутреннего представления обычно используются молекулярные графы. Молекулярный граф — связный граф, находящийся во взаимно-однозначном соответствии со структурной формулой химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а рёбрам графа — химические связи между этими атомами. Такое представление не хранит информацию о трехмерной структуре молекулы. Обычно граф неориентированный (связи в молекуле не имеют направления) и его вершины помечены (символами атомов)[1]. Если вершины графа непомечены, то он будет отражать только структуру, а не состав молекулы. Две вершины графа могут соединяться несколькими ребрами, так как связь может быть одинарной или кратной.

Граф может быть представлен в виде матрицы разными способами, например матрицей смежности, расстояний, инцидентности.

Матрица смежности для молекулы, состоящей из n атомов, это квадратная матрица размером $n \times n$, содержащая информацию о всех связях в молекуле. Если на пересечении i -ой строки и j -ого столбца стоит 1, то между соответствующими атомами есть связь. Если нет связи между рассматриваемыми атомами, то на соответствующей позиции матрицы стоит 0. То есть матрица смежности является Булевой матрицей. Все диагональные элементы матрицы равны 0 и она является симметричной. Такая матрица является избыточной. Ее можно упростить, убрав дублирование половины матрицы, то есть приведя к верхней треугольной матрице. Также для ясности можно упустить нули и удалить информацию об атомах водорода[1].

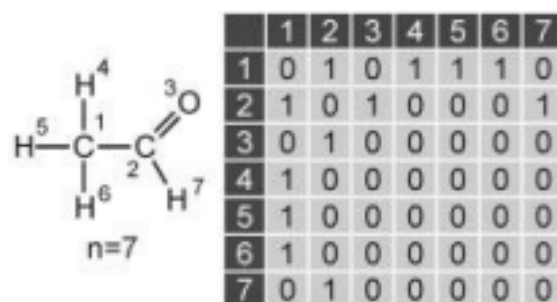


Рис. 3.1: Матрица смежности этанола[1]

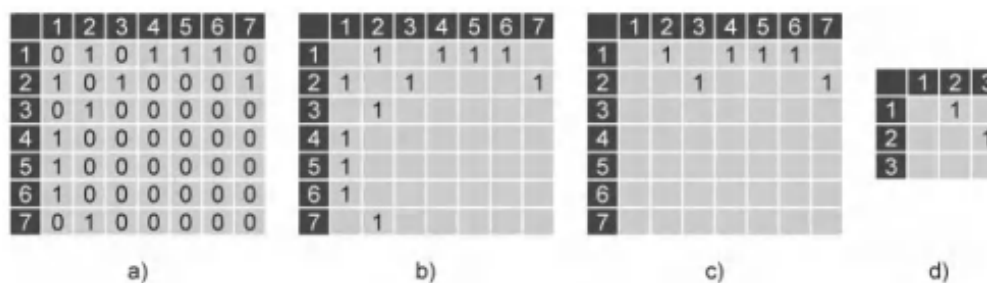
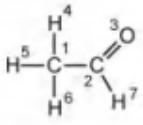


Рис. 3.2: а) избыточная матрица смежности этанола; б) матрица после опускания нулей; в) упрощенная до верхней треугольной матрицы; г) после удаления атомов водорода[1]

Для примера рассмотрим матрицу расстояний, у которой ее элемент представляет кратчайшее расстояние между соответствующими атомами. Расстояние может выражаться в геометрическом расстоянии (ангстремах) или топологическом расстоянии (числе связей)[1].



a)

	C1	C2	O3	H4	H5	H6	H7
C1	0	1.400	2.190	1.022	1.023	1.022	2.106
C2	1.400	0	1.123	1.999	1.982	1.999	1.022
O3	2.190	1.123	0	2.349	2.708	2.995	1.859
H4	1.022	1.999	2.349	0	1.668	1.661	2.895
H5	1.023	1.982	2.708	1.668	0	1.668	2.562
H6	1.022	1.999	2.955	1.661	1.668	0	2.336
H7	2.106	1.022	1.859	2.895	2.566	2.336	0

b)

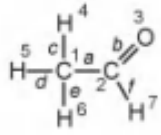
	C1	C2	O3	H4	H5	H6	H7
C1	0	1	2	1	1	1	2
C2	1	0	1	2	2	2	1
O3	2	1	0	3	3	3	2
H4	1	2	3	0	2	2	3
H5	1	2	3	2	0	2	3
H6	1	2	3	2	2	0	3
H7	2	1	2	3	3	3	0

Рис. 3.3: матрица расстояний этанола с а) геометрическим расстоянием; б) топологическим расстоянием[1]

Еще одним представлением является матрица инцидентности. Это матрица размером $n \times m$, где n -число вершин(атомов), m - число ребер (связей). Если на пересечении i -ой строки и j -ого столбца стоит значение 1, то рассматриваемые вершина и ребро инцидентные [1].

	C1	C2	O3	H4	H5	H6	H7
a	1	1	0	0	0	0	0
b	0	1	1	0	0	0	0
c	1	0	0	1	0	0	0
d	1	0	0	0	1	0	0
e	1	0	0	0	0	1	0
f	0	1	0	0	0	0	1

a)



$n=7; m=6$

	C1	C2	O3	H4	H5	H6	H7
a	1	1					
b		1	1				
c	1			1			
d	1				1		
e	1					1	
f		1					1

b)

	C1	C2	O3
a	1	1	
b		1	1

c)

Рис. 3.4: а) избыточная матрица инцидентности этанола; б) матрица после опускания нулей; с) после опускания атомов водорода[1]

Все описанные матрицы не несут информации о типе и порядке связей в молекуле.

Одним из недостатков матрицы смежности является, что число ее элементов равно квадрату числа атомов, а для представления молекулярного графа необходимо, чтобы число элементов в представлении линейно зависело от числа атомов в молекуле. Это достигается с помощью представления таблицей связности, в которой дается список атомов и список связей. Существует много вариантов матриц связности. Например, атомы произвольно нумеруются и в соответствии с индексом заносятся в список атомов. Информация о связях хранится во второй таблице, где для каждой связи записываются индексы атомов, которые она соединяет и ее кратность[1].

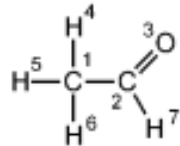
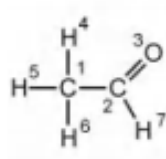
	Atom list		Bond list		
	1 st atom	2 nd atom	bond order		
	1	C	1	2	1
	2	C	2	3	2
	3	O	2	7	1
	4	H	1	4	1
	5	H	1	5	1
	6	H	1	6	1
	7	H			

Рис. 3.5: таблица связности этанола

Другой формой представления является таблица, в которой первые две колонки содержат информацию о индексах и символах атомов. Далее таблица дополняется колонками, в которых указаны индексы соседних атомов и кратность связи. Такая форма представления избыточна, в ней каждая связь записана дважды. Ее можно упростить, убрав повторение связей и опустив атомы водорода.



atom index	element	1 st index of atom	bond order	2 nd index of atom	bond order	3 rd index of atom	bond order	4 th index of atom	bond order
1	C	2	1	4	1	5	1	6	1
2	C	1	1	3	2	7	1		
3	O	2	2						
4	H	1	1						
5	H	1	1						
6	H	1	1						
7	H	2	1						

Рис. 3.6: таблица связности этанола[1]

Таблицы связности могут быть дополнены другими списками, например содержащими информацию о свободных электронах или заряде атомов, что является еще одним преимуществом по сравнению с матрицами смежности[1].

Методы теории графов находят широкое применение в хемоинформатике.

3.1.2 Внешнее представление

Внешнее представление химических соединений используется в случае длительного хранения химической информации и обмена ею между приложениями.

Простейшим типом внешнего представления структур химических соединений являются линейные нотации в виде строки символов. Линейные нотации позволяют свободно обмениваться информацией о химических соединениях без необходимости использования специального программного обеспечения. Наиболее популярные типы линейных нотаций:

- SMILES (Simplified Molecular Input Line Entry System);
- SMARTS (расширение SMILES)
- InChI (IUPAC International Chemical Identifier)

В настоящее время наиболее распространённым видом линейных нотаций являются строки SMILES. Преобразование структуры соединения в строку символов SMILES определяется 6 правилами [1] :

- Атомы представлены их атомными символами.
- Атомы водорода опущены
- Соседние атомы расположены рядом друг с другом.
- Двойные и тройные связи характеризуются «=» и «>», соответственно.
- Разветвления молекулы представлены скобках.
- Циклы описаны путем выделения цифры для двух "соединяющих" атомов в цикле.

Синтаксис SMILES позволяет описывать структурные изомеры. SMARTS является расширением SMILES для поисковых запросов в химических базах данных.

Для кодировки химических структур IUPAC предложил универсальную линейную нотацию InChI. То есть это цифровой эквивалент названию соединения по номенклатуре IUPAC. Он содержит следующие уровни информации: связанность, таутомеризм, изотопы, стереохимия, заряд. Этот формат более сложен для интерпретации пользователем, но сохраняет такую возможность.

InChIKey – хешированная версия InChI, созданная для быстрого поиска.

Второй тип внешнего представления структур химических соединений основан на непосредственном кодировании таблицы связности молекулярного графа. Такие распространённые форматы как MOL, SDF и RDF,

которые в настоящее время являются стандартными для обмена химической информацией, можно считать способами представления в виде текстового файла матрицы смежности молекулярного графа.

3.2 Молекулярные дескрипторы

Традиционный подход к обработке химической информации состоит в отображении химического пространства на дескрипторное пространство, образуемое вычисляемыми для каждого химического объекта векторами молекулярных дескрипторов — числовых характеристик, описывающих химические объекты. Это дает возможность применять методы математической статистики и машинного обучения для работы с химическими объектами. В принципе дескриптором может являться любое число, которое можно рассчитать из структурной формулы химического соединения. Молекулярные дескрипторы можно классифицировать по "размерности": 0D, 1D, 2D, 3D и 4D. 0D дескрипторы описывают совокупную информацию, такую как количество атомов, количество связей, молекулярный вес. 1D дескрипторы описывают число фрагментов в молекуле. В качестве примера можно привести число гидроксильных групп, нитрогрупп[1]. 2D дескрипторы описывают свойства, которые могут быть вычислены из двумерного представления молекул (например, индексы связности) и 3D-дескрипторы зависят от конформации молекул (например, доступная растворителю площадь поверхности).[3]

3.3 Молекулярные отпечатки

Молекулярные «отпечатки» (molecular fingerprints) содержат информацию о присутствии или отсутствии определенных признаков в химическом соединении, например, фрагментов. Могут быть организованы 2

основными способами:

- бинарной строки
- хеш-таблицы

Рассмотрим подробнее молекулярные отпечатки в виде бинарной строки. Каждая подструктура или фрагмент активирует определенное количество позиций (битов) в молекулярном отпечатке. Иногда подструктуре может соответствовать 1 или несколько битов. Алгоритм определяет какие биты были активированы подструктурой. Одна и та же подструктура всегда активизирует одинаковые биты. Если фрагменту соответствует 1 бит, то в случае равенства единице, то фрагмент присутствует в молекуле, иначе его значение равно нулю. Обычно длина бинарной строки 150-2500 битов. Алгоритм работает таким образом, что всегда возможно ассоциировать биты с конкретной подструктурой. Однозначное представление химической структуры строкой позволяет проводить эффективный поиск схожих молекул.

Для хэшированных молекулярных отпечатков нет возможности определить, какие конкретные элементы присутствуют в молекуле.

3.4 Молекулярное подобие

В основе принципа молекулярного подобия лежит идея, что структурно схожие молекулы предположительно обладают сходной биологической активностью. Однако это предположение не всегда может быть верным. Например, "activity cliffs" в которых незначительная модификация функциональных групп вызывает резкое изменение активности [4]. Поиск структурного сходства молекул основан на доле различных фрагментов, которые присутствуют одновременно в обеих молекулах. Поиск

молекул по такому критерию называется поиском по молекулярному подобию (Similarity Search). В качестве количественной меры молекулярного подобия часто рассматривается величина, возрастающая с уменьшением расстояния между химическими соединениями в дескрипторном пространстве.

Структурное сходство двух молекул чаще всего оценивается путем вычисления коэффициента Танимото (Tc). Tc, также известный как индекс Джаккарда, описывает степень схожести двух множеств. Для парного сравнения используются молекулярные отпечатки молекул. Коэффициент Танимото определяется как :

$$Tc = \frac{bc}{b1 + b2 - bc}, \quad (3.1)$$

где b1 - число битов набора первой молекулы, b2 - число битов набора второй молекулы, bc - число битов общих для обеих молекул. Значения коэффициента Танимото лежит в пределах от 0 до 1[3]. Высокие значения Tc указывают на то, что два соединения похожи, но не дает информации о масштабах сходства, например о том, какие конкретные химические группы они разделяют.

4 Синергия

4.1 Введение

Под комбинированной терапией подразумевают одновременное применение нескольких препаратов.

Существует три типа эффектов от комбинации препаратов:

- аддитивный, когда комбинированный эффект эквивалентен сумме независимых эффектов

- синергический, когда комбинированный эффект больше аддитивного
- антагонистический, когда комбинированный эффект меньше аддитивного

Целью комбинированной терапии является достижение синергического или, по крайней мере, аддитивного, но комплементарного эффекта [5, 6]

4.2 Область применения синергии

Большинство заболеваний вызвано сложными биологическими процессами. Широко известным примером являются онкологические заболевания. На данный момент разработана таргетная терапия рака, которая блокирует рост раковых клеток с помощью вмешательства в механизм действия конкретных целевых (таргетных) молекул, необходимых для канцерогенеза, то есть ингибирует критические сигнальные пути рака. Как и ожидается, такая терапия будет более эффективной, чем прежние виды лечения, и менее вредной для нормальных клеток. Однако за резким начальным положительным ответом многих таргетных методов лечения рака часто следует развитие лекарственной устойчивости, приводящей к рецидиву заболевания[7]. Существует множество механизмов, которые могут привести к лекарственной устойчивости, которые включают генетическую и негенетическую гетерогенность, присущую распространенным видам рака, в сочетании со сложными механизмами обратной связи и регуляции, а также динамическими взаимодействиями между опухолевыми клетками и их микроокружением. Любая монотерапия может быть ограничена по своей эффективности, но комбинации лекарств потенциально могут преодолеть эти ограничения[8]. Они состоят из нескольких агентов, каждый из которых обычно используется в кли-

нике как один эффективный препарат. Поскольку агенты в лекарственных комбинациях могут модулировать активность отдельных белков, лекарственные комбинации могут помочь повысить терапевтическую эффективность, преодолев избыточность, лежащую в основе лекарственной устойчивости. Соответственно приводят к более длительным реакциям у пациентов[9].

Кроме того, токсичность и неблагоприятные побочные эффекты, вероятно, снижаются, поскольку дозы комбинаций лекарств обычно ниже, чем дозы отдельных агентов. В настоящее время медикаментозная комбинаторная терапия становится перспективной стратегией лечения многофакторных сложных заболеваний[10].

Также необходимо упомянуть еще одну возможную область применения синергии - химическое перепрограммирование клеток малыми молекулами. Точнее, эта область нуждается в использовании комбинаций малых молекул, так как при переходе из одного клеточного состояния в другое происходят большие изменения в фенотипе и обработки клетки одним соединением недостаточно.

4.3 Предсказание синергетических пар соединений

Изначально эффективные комбинации препаратов предлагались на основе клинического опыта и большинство подходов к выявлению синергетических пар соединений часто носит экспериментальный характер. В исследованиях рака анализ синергизма обычно проводится путем обработки клеточных линий *in vitro* всеми возможными комбинациями соединений. Однако такой подход требует много усилий и затрат. Более того, экспериментальные скрининги накладывают серьезные ограничения на практический размер библиотек и их разнообразие. Вычислительные методы прогнозирования синергии соединений потенциально могут позво-

литель исследователям отбирать наиболее перспективные пары для экспериментального скрининга, и, вследствие сокращения количества изучаемых комбинаций, сократить затраты ресурсов [5].

Большинство существующих методов скрининга нацелены на прогнозирование синергических эффектов двух препаратов, поскольку комбинаторный эффект трех и более препаратов технически сложнее предсказать и практически отсутствуют экспериментальные данные для последующей валидации метода[6]. Также есть наблюдения, что наиболее значительное улучшение достигается при добавлении только одного дополнительного препарата. При дальнейшем добавлении лекарств эффективность постепенно уменьшается и в конечном счете достигает плато[11].

На молекулярном уровне синергетические взаимодействия могут быть реализованы несколькими различными механизмами. Например, соединение может сенситизировать клетки к другому соединению, регулируя его поглощение и распределение, модулируя ростовые свойства клетки, ингибируя деградацию соединения, ингибируя пути, индуцирующие резистентность или снижая токсичность другого соединения[5].

Механизмы синергических эффектов не являются универсальными среди различных лекарств или раковых заболеваний[6].

Можно выделить две гипотезы, лежащие в предсказании синергии препаратов[5]:

- гипотеза сходства: соединения, более схожие по вызываемым транскрипционным изменениям, с большей вероятностью будут синергическими
- гипотеза комплиментарности: соединения, которые вызывают наиболее различные, но взаимодополняющие транскрипционные изменения, с большей вероятностью будут синергическими (так как мы

ищем синергетические пары среди препаратов с определенным необходимым эффектом, то можно сказать, что по гипотезе несходства синергетические соединения комплементарно дополняют друг друга в достижении желаемого эффекта)

На данный момент отсутствуют стандартные подходы к прогнозированию активности пар соединений на основе транскриптомных данных[5].

Поиск синергетических пар можно представить в виде задачи регрессии, когда необходимо предсказать уровень синергии пары препаратов. Эта же задача может быть представлена в виде ранжирования списка пар препаратов по степени синергии[5]. Но также можно рассматривать более простую задачу бинарной классификации, когда необходимо сказать, является пара препаратов синергетической или нет[8].

Данные, используемые для предсказания синергии, можно поделить на несколько типов:

- глубокая молекулярная характеристика клеточных линий, которая включает соматические мутации, изменения числа копий, метилирование ДНК и профили экспрессии генов
- фармакологические признаки препаратов, включая предполагаемые лекарственные мишени и химические свойства препаратов, представленные дескрипторами и молекулярными отпечатками, области медицинских показаний, побочные эффекты и токсикофторы, сигнальные пути
- данные монотерапии, включающие зависимость жизнеспособности клеток от дозы препарата при обработке одиночным препаратом, транскрипционные данные (профиль экспрессии генов клеток после обработки препаратом)

Безусловно, признаки типа клеточной линии существенны, так как они определяют молекулярный контекст. [8]. Признак "сигнальный путь" слабо предсказательный, возможно, потому, что простая ассоциация между лекарствами и путями через целевые белки недостаточно отражает физиологический контекст, в котором работают лекарства [9]. Также плохой предсказательной способностью обладает признак "побочные эффекты" потому что он сильно зашумлен, так как существуют некоторые общие побочные эффекты, связанные с большинством лекарств. Эффективность признака побочных эффектов может быть улучшена, если рассмотреть только тяжелые побочные эффекты, связанные с лекарствами [9]. Монотерапия дает существенную информацию о прямом лечебном эффекте на линию раковых клеток, что является основой лекарственного синергизма. Например, если препарат вообще не действует на определенную клеточную линию, он вряд ли будет синергировать с другими препаратами [6]. Однако такой признак, как зависимость жизнеспособности клеток от дозы препарата, в основном используется для поиска пар противоопухолевых препаратов и не подходит для химического перепрограммирования. Также наблюдалась информативность транскрипционных данных, поскольку профили геномной экспрессии очень динамичны и зависят от контекста [12].

Однако различные типы данных дополняют друг друга при прогнозировании комбинаций лекарственных средств [9]. То есть использование ансамбля различных наборов признаков улучшает качество прогноза [8].

Кроме того, по предоставленному набору данных задачу поиска синергетических пар можно реализовывать в двух сценариях [13]:

- для предсказания уровня синергии пар препаратов предоставляются только данные по используемым клеточным линиям, фармакологические признаки препаратов, данные монотерапии

- для предсказания уровня синергии пар препаратов помимо выше перечисленных данных предоставляется обучающий набор данных, который, например, может состоять из других пар препаратов, для которых будут доступны данные по клеточным линиям, фармакологические признаки препаратов, данные монотерапии и также известно, какие пары препаратов являются синергетическими, какие нет. В данном случае возможно применять методы машинного обучения.

В реальности предоставляемые данные бывают очень разнообразными, поэтому может применяться и смешанный сценарий, то есть использовать обучение с частичным привлечением учителя[12].

Если в работе используется сценарий обучения с учителем, то либо известны синергетические комбинации, либо есть экспериментальные данные жизнеспособности клеток после обработки комбинациями препаратов в зависимости от дозы, по которым рассчитывается синергетический эффект.

Среди подходов к моделированию лекарственной синергии популярны методы машинного обучения. Однако труднодоступность обучающих данных препятствует широкому использованию методов машинного обучения. Используются следующие подходы: регрессия, деревья решений, случайные леса, Гауссовские процессы, SVM, нейронные сети. При сравнении моделей был сделан вывод, что класс алгоритмов показал слабую связь с производительностью метода[8].

Так как предсказание синергии можно отнести к задаче бинарной классификации, то популярны метрики ROC-AUC, accuracy, precision, recall, F1[9, 12]. Рассмотрим их подробнее. Наши данные поделены на два класса. Их метки принято обозначать как «положительные» и «отрицательные». При классификации мы можем верно определить класс (истинно) или допустить ошибку, то есть отнести к ложному классу, поэтому воз-

можны следующие исходы:

- истинно положительные (TP)
- истинно отрицательные (TN)
- ложно положительные (FP)
- ложно отрицательные (FN)

		Prediction outcome		
		positive	negative	
Actual value	positive	TP	FN	TP + FN
	negative	FP	TN	FP + TN
		TP + FP	FN + TN	

Рис. 4.1: Матрица сопряженности возможных результатов бинарной классификации

Одной из наиболее простых метрик является точность (ассигасу). Она показывает количество верно классифицированных объектов (истинно положительных и истинно отрицательных) относительно общего количества всех объектов и считается следующим образом:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4.1)$$

Метрика ассигасу имеет недостаток: она не подходит для несбалансированных классов, где может быть много экземпляров одного класса и мало другого.

Для оценки качества работы алгоритма на каждом из классов по отдельности вводятся метрики precision (точность) и recall (полнота).

Метрика *precision* показывает сколько из всех объектов, которые классифицируются как положительные, действительно являются положительными, относительно общего количества полученных от модели позитивных меток.

$$precision = \frac{TP}{TP + FP} \quad (4.2)$$

Важность этой метрики определяется тем, насколько высока для рассматриваемой задачи «цена» ложно положительного результата.

Метрика *recall* показывает, сколько объектов модель смогла правильно классифицировать с позитивной меткой из всего множества позитивных. Она вычисляется по следующей формуле:

$$recall = \frac{TP}{TP + FN} \quad (4.3)$$

Необходимо уделить особое внимание этой оценке, когда в поставленной задаче ошибка нераспознавания положительного класса высока.

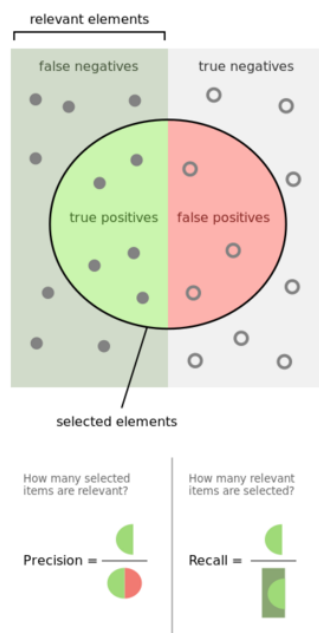


Рис. 4.2: Схематичное представление *precision* и *recall*

Precision и *recall* не зависят, в отличие от *accuracy*, от соотношения классов и потому применимы в условиях несбалансированных выборок. Если

Precision и Recall являются одинаково значимыми, то можно использовать их среднее гармоническое для получения оценки результатов:

$$F1 - score = \frac{2 \cdot recall \cdot precision}{recall + precision} \quad (4.4)$$

Для определения бинарной метки (0 или 1) по какому-нибудь вещественному ответу алгоритма (как правило, вероятности принадлежности к классу) необходимо выбрать порог, до которого ставится метка "0" после "1". Порог, равный 0.5 кажется естественным, но он не всегда оказывается оптимальным, например, при отсутствии баланса классов.

Одним из способов оценить модель в целом, не привязываясь к конкретному порогу, является метрика ROC AUC — площадь (Area Under Curve) под кривой ошибок (Receiver Operating Characteristic curve). Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR).

True Positive Rate (Recall) показывает долю верно классифицируемых объектов положительного класса и считается по формуле:

$$TPR = \frac{TP}{TP + FN} \quad (4.5)$$

False Positive Rate показывает, какую долю из объектов отрицательного класса алгоритм предсказал неверно, и определяется как:

$$FPR = \frac{FP}{FP + TN} \quad (4.6)$$

Когда классификатор не делает ошибок ($FPR = 0$, $TPR = 1$) мы получим площадь под кривой, равную единице; в противном случае, когда классификатор случайно выдает вероятности классов, AUC-ROC будет стремиться к 0.5, так как классификатор будет выдавать одинаковое количество TP и FP. Каждая точка на графике соответствует выбору некоторого порога. Площадь под кривой в данном случае показывает каче-

ство алгоритма (больше — лучше), кроме этого, важной является крутизна самой кривой — надо максимизировать TPR, минимизируя FPR, а значит, кривая в идеале должна стремиться к точке (0,1).

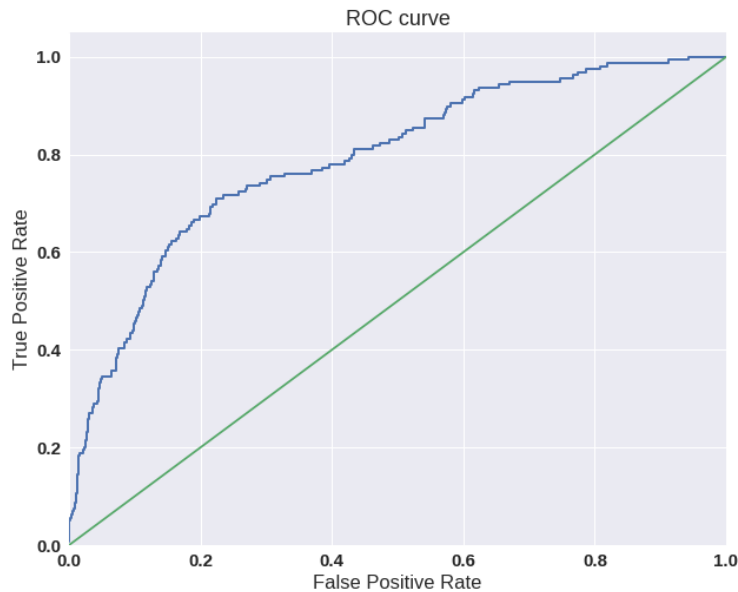


Рис. 4.3: ROC кривая

Для задачи регрессии используется MSE, RMSE, коэффициент корреляции Пирсона[10]. Рассмотрим их подробнее.

Метрика Mean Squared Error (MSE). Измеряет среднюю сумму квадратной разности между фактическим значением и прогнозируемым значением для всех объектов выборки. Возведение во вторую степень необходимо, чтобы отрицательные значения не компенсировались положительными. Чем больше разность, тем больше ее вес в этой метрике. Ниже приведена ее формула.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^{\text{pred}})^2 \quad (4.7)$$

Метрика Root Mean Squared Error (RMSE) - это корень от квадрата

ошибки. Формула приведена ниже.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^{\text{pred}})^2} \quad (4.8)$$

Также в соревновании консорциума DREAM использовалась РС-метрика, которая будет рассмотрена ниже [5], [8], [6], [12].

Для проверки предсказанных синергетических пар используются экспериментальные данные и литература [12].

Ниже описаны некоторые методы предсказания синергии, начиная с первого сценария, когда недоступен обучающий набор данных. В качестве примера необходимо упомянуть соревнования, проводимые консорциумом DREAM Challenges.

4.3.1 Методы предсказания синергии, разработанные на основе наборов данных консорциума DREAM Challenges

В 2012 году был проведен NCI-DREAM Drug Synergy Prediction Challenge: перед участниками была поставлена задача предсказания синергетической и антагонистической активности пар соединений [5]. Перед участниками стояла задача отранжировать 91 пару соединений (все парные комбинации препаратов из 14 соединений) от наиболее синергичных до наиболее антагонистичных при воздействии на клеточную линию OCI-LY3 диффузной крупноклеточной лимфомы (DLBCL). Для этого им были предоставлены следующие данные :

- зависимость жизнеспособности клеток OCI-LY3 от дозы препарата после обработки им (для каждого препарата из набора и включая ДМСО в качестве контрольной среды)
- профили экспрессии генов для 3 биологических копий необработанных клеток и через 6 ч, 12 ч и 24 ч после обработки каждым соединением (для каждого препарата из набора)

- базовый генетический профиль клеточной линии OCI-LY3

Любые дополнительные данные из литературы или экспериментов считались допустимыми, но прямое измерение синергизма соединений было категорически запрещено. В соревновании было предложено 31 метод от участников и метод SynGen от одного из организаторов (этот подход оценивался отдельно от методов, предложенных участниками). Среди методов наблюдалось большое разнообразие, что хорошо показывает отсутствие стандартных подходов к прогнозированию синергетической активности пар соединений на основе транскриптомных данных. Также отсутствие обучающих данных препятствовало использованию методов машинного обучения. Среди 31 команды 10 основывали свои прогнозы на гипотезе о том, что соединения с более высоким сходством транскрипционного профиля с большей вероятностью будут синергетическими (гипотеза сходства), восемь команд предположили обратное (гипотеза несходства). Остальные команды либо использовали комбинацию гипотез сходства и несходства (комбинированная гипотеза, $n = 4$), либо использовали более сложные гипотезы ($n = 9$).

Для проверки участников организаторами был создан валидационный набор данных, в котором экспериментально оценивали синергизм пар соединений по жизнеспособности клеток OCI-LY3.

Для экспериментальной оценки пользовались моделью excess over Bliss (EOB), которая определяет, является ли комбинированное действие двух соединений значительно большим или меньшим, чем независимое сочетание их индивидуальных эффектов. Также используется модель Bliss additivism, которая считает, что соединения D_x и D_y с экспериментально определенными долями ингибирования (доля клеток, погибших после обработки) f_x и f_y имеют аддитивный эффект, если ожидаемая доля ин-

гибирования f_{xy} , индуцируемая их комбинацией определяется как:

$$f_{xy} = 1 - (1 - f_x)(1 - f_y) = f_x + f_y - f_x \cdot f_y \quad (4.9)$$

Excess over Bliss считается как разница между долей ингибирования комбинации f_z и ожидаемой долей ингибирования f_{xy} при аддитивном эффекте:

$$eob = f_z - f_{xy} \quad (4.10)$$

Если $eob \approx 0$, то пара соединений имеет аддитивный эффект. Затем если $eob > 0$ ($eob < 0$), то пара имеет синергетический (антагонистический) эффект. Оценка активности eob была использована для ранжирования всех пар от наиболее синергетических до наиболее антагонистических.

Для оценки предсказаний команд, использовался модифицированный индекс соответствия, который называли вероятностным индексом соответствия. Эта метрика количественно определяет соответствие между ранжированием пар соединений в валидационной выборке и ранжированием, предсказанным командой.

Для подробного разбора оценки предсказаний необходимо рассмотреть индекс соответствия.

Проранжируем список из 91 соединения по экспериментально определенному и усредненному по всем репликатам ЕОВ, от наиболее синергетических до наиболее антагонистических пар. Обозначим ранг пары i ($1 \leq i \leq 91$) как o_i . Аналогично для ранжированного списка пар, предсказанного командой, обозначим ранг пары p_i . Заметим, что если $i \neq j$, то $o_i \neq o_j$, $p_i \neq p_j$. Поэтому можем определить s_{ij} :

$$s_{ij} = \begin{cases} 1 & \text{if } (o_i > o_j \text{ } p_i > p_j \text{ или } o_i < o_j \text{ } p_i < p_j) \\ 0 & \text{if } (o_i > o_j \text{ } p_i < p_j \text{ или } o_i < o_j \text{ } p_i > p_j) \end{cases}$$

Индекс соответствия определяется как :

$$c - index = \frac{2}{91 \cdot 90} \sum_{i=1..90, j=i+1..91} s_{ij} \quad (4.11)$$

Для учета шума при экспериментальном ранжировании вводится вероятностный индекс соответствия. Для $i \neq j$ он вычисляется как :

$$sp_{ij} = \begin{cases} \frac{1}{2} + \frac{1}{2} \text{err}\left(\frac{EOB_i - EOB_j}{\sqrt{sem_{EOB_i}^2 + sem_{EOB_j}^2}}\right) & \text{if } p_i < p_j \\ \frac{1}{2} - \frac{1}{2} \text{err}\left(\frac{EOB_i - EOB_j}{\sqrt{sem_{EOB_i}^2 + sem_{EOB_j}^2}}\right) & \text{if } p_i > p_j \end{cases}$$

где erf - функция ошибки, EOB_i экспериментальное ЕОВ, усредненное по всем репликатам для i -ой пары, sem_{EOB_i} - среднеквадратическое отклонение ЕОВ для i -ой пары. Предположим, что пара i более синергетическая в среднем, чем j , тогда $EOB_i > EOB_j$ и аргумент функции ошибок положительный, то есть самая функция ошибки будет принимать положительные значения. Если будет предсказано, что пара i более синергетическая чем j ($p_i < p_j$), то $sp_{ij} > \frac{1}{2}$. Однако, если предсказание неверно, то есть $p_i > p_j$, то $sp_{ij} < \frac{1}{2}$. Аналогично в случае $EOB_i < EOB_j$. Если предсказание верно, то sp_{ij} лежит от 0,5 до 1, иначе от 0 до 0,5.

Вероятностный индекс соответствия определяется как :

$$PC - index = \frac{2}{91 \cdot 90} \sum_{i=1..90, j=i+1..91} sp_{ij} \quad (4.12)$$

Максимальный PC-index (PC_{max}) был равен 0,9. Минимальный PC-index (PC_{min}) определялся для предсказания с ранжируемым пар противоположным (в обратном порядке) экспериментальному ранжированию пар, он был равен 0,1.

Нормализованный вероятностный индекс соответствия определяется как :

$$PC - index_{norm} = \frac{PC - index - PC_{min}}{PC_{max} - PC_{min}} \quad (4.13)$$

Для проверки результатов использовали вторую метрику - (пересчитанную корреляцию Спирмена). Результаты оценки методов по 2 метрикам были схожи, наблюдались малые отличия для нескольких команд, работавших хуже, чем случайная модель.

Среди 31 методы только три метода были статистически значимы ($FDR = 0,05$) : DIGRE, IUPUI_CCBB and DPST.

DIGRE

Самым лучшим методом был DIGRE (drug-induced genomic residual effect). Он основывается на гипотезе, что при последовательной обработке клеток двумя соединениями транскрипционные изменения, индуцируемые первым препаратом, способствуют эффекту второго. То есть, что синергия обусловлена транскриптомными остаточными эффектами, которые представляют собой транскрипционные изменения, индуцированные первым соединением. Данная гипотеза согласуется с наблюдениями, что последовательность введение лекарств имеет влияние на результат. Несмотря на то, что соединения вводились одновременно в экспериментах, алгоритм моделирует синергию последовательно. В алгоритме можно выделить 3 шага:

- оценка сходства r между 2 соединениями в паре на основе сравнения транскрипционных изменений после обработки одним соединением. Смотрят на перекрытие дифференциально экспрессируемых генов относящихся к восьми сигнальным путям KEGG, связанными с ростом клеток (сфокусированный взгляд) и генов с повышенной экспрессией относящихся к 32 раковым сигнальным путям KEGG (глобальный взгляд)
- аппроксимируют долю выживших клеток после обработки вторым

препаратом пары, учитывая влияние транскрипционных изменений, индуцированных первым препаратом. Для этого используют оценку сходства r , рассчитанную на предыдущем шаге.

$$1 - f_{B+A'} = (1 - rf_{2B})(1 - (1 - r)f_B), \quad (4.14)$$

где $f_{B+A'}$ - доля погибших клеток после обработки соединениями В (предполагается, что в клетке уже были индуцированы транскрипционные изменения соединением А), f_B , f_{2B} - доля погибших клеток после обработки соединением В после однократной и двойной дозы соответственно, которые определяются по предоставленной участникам зависимости жизнеспособности клеток от дозы препарата

- определяется доля погибших клеток, после обработки парой препаратов:

$$Z_{B+A'} = 1 - (1 - f_A)(1 - f_{B+A'}), \quad (4.15)$$

где f_A - доля погибших клеток после обработки препаратом А.

Затем оценка синергии определяется как среднее для 2 долей погибших клеток при обработке парой препаратов в разной предполагаемой последовательности ($Z_{B+A'}$, $Z_{A+B'}$)

Для этого метода PC index = 0.61.

Алгоритм SynGen

В основе алгоритма лежит предположение, что активность главных регуляторов (Master Regulators, MR) определенного клеточного фенотипа, которые выводятся Master Regulator Inference algorithm MARINa[14, 15], имеют важное значение для жизнеспособности клеток. MRs определяются как регуляторы, которые необходимы и достаточны для поддержания

специфичной для фенотипа сигнатуры экспрессии генов. Целью обработки является:

- подавление активности MRs клеточного состояния
- активация MRs клеточной смерти

Таким образом, сначала SynGen выводит MRs для гибели клеток OCI-LY3 и состояния клеток, а затем идентифицирует соединения, которые наиболее комплементарны в индуцировании первого и отмене последнего. Для определения MRs используются 2 сигнала:

- сигнатура "клеточной смерти основанной на профиле экспрессии генов после обработки 14 соединениями, которые обладают заметной токсичностью
- сигнатура "клеточной зависимости связанной с активированным В-клеточным подтипом клеток DLBCL (которые включают OCI-LY3) по сравнению с подтипом В-клеток герминативного центра. Она вычисляется с использованием общедоступных профилей экспрессии генов для клеточных линий подтипа В-клеток герминативного центра (OCI-LY1, OCI-LY7, OCI-LY8, OCI-LY18 и SUDHL5) и для активированной линии подтипа В-клеток OCI-LY3.

Затем SynGen предсказывает синергетические комбинации соединений, выбрав составные пары, которые наиболее комплементарны в реализации или отмене этих MRs-паттернов соответственно. SynGen способен предсказывать только синергетические пары, то есть не предназначен для прогнозирования антагонизма соединений.

В 2015-2016 Dialog for Reverse Engineering Assessments and Methods (DREAM) Challenges в партнерстве с AstraZeneca и Институтом Сэн-

гера провели AstraZeneca-Sanger Drug Combination Prediction DREAM Challenge[8]. Основными задачами соревнования были:

- предсказать, будет ли известная (ранее протестированная) комбинация лекарств эффективна для конкретного пациента
- предсказать, какие новые (непроверенные) комбинации лекарств будут синергетическими у пациентов
- определение новых биомаркеров, которые могут выявить основные механизмы, лежащие в основе синергии лекарств.

Данные были собраны на основе 11,576 экспериментов (зависимость жизнеспособности клеток от дозы препаратов в паре, представленная в виде матрицы размером 6×6 , в которой первая строка и первый столбец являются данными монотерапии для 2 препаратов) с использованием 85 клеточных линий рака. Оценка синергии пары препаратов рассчитывалась на основе матриц жизнеспособности клеток от доз препаратов в комбинации. Таким образом, набор данных включал высоковоспроизводимые измерения жизнеспособности клеток от доз препаратов в парах и оценки синергизма для 910 попарных комбинаций из 118 препаратов, а также информацию о лекарствах, включая предполагаемые лекарственные мишени и их химические свойства. Также была предоставлена глубокая молекулярная характеристика клеточных линий, включая соматические мутации, изменения числа копий, метилирование ДНК и профили экспрессии генов, измеренные до обработки клеток препаратами. Оценка синергии пары препаратов не была предоставлена по всем клеточным линиям рака.

Соревнование проводилось по 2 направлениям:

- предсказать синергию пары препаратов для конкретной клеточной

линии, при этом частично известны уровни синергии для этой комбинации в других клеточных линиях (sub-challenge 1, SC1). первое направление было разделено на 2 категории:

- доступны все данные для предсказания синергии (SC1A)
- данные ограничены мутациями и изменением копийности (имитируя текущую возможность клинического анализа), (SC1B).
- предсказать синергию пары препаратов для конкретной клеточной линии, по которым не предоставлены обучающие данные по другим клеточным линиям, то есть используя переносимые признаки, идентифицированные из ранее изученных независимых пар лекарств (sub-challenge 2, SC2).

Участники обоих направлений использовали одни и те же обучающие наборы данных, но методы оценивались на разных тестовых выборках и с использованием разных метрик. Также важно отметить, что для первого направления уровни синергии, предсказанные участниками, должны быть непрерывными, в то время как SC2 требует бинарных предсказаний, указывающих, являются ли два препарата синергичными или нет.

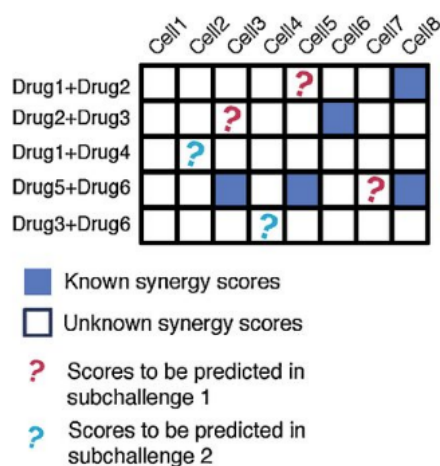


Рис. 4.4: Схематичное представление задач SC1 и SC2

Рассмотрим оценку методов по первому направлению соревнования. Поскольку размер выборки для каждой комбинации препаратов варьируется (например, некоторые комбинации препаратов были протестированы на большем количестве клеточных линий, чем другие), применение корреляции Пирсона непосредственно ко всем показателям синергии потенциально даст больший вес комбинациям лекарств, которые имеют больше экспериментов. Поэтому для придания равных весов всем комбинациям препаратов в качестве метрики была использована средневзвешенная корреляция Пирсона:

$$p_w = \frac{\sum_{i=1}^N \sqrt{n_i - 1} * p_i}{\sum_{i=1}^N \sqrt{n_i - 1}} \quad (4.16)$$

где n_i - число клеточных линий, обработанных данной комбинацией, $N = 167$ (число комбинаций, используемых для теста), p_i - коэффициент корреляции Пирсона, рассчитанный для предсказанных и наблюдаемых уровней синергии в пределах одной комбинации препаратов по возможным клеточным линиям[6].

Для оценки методов по второму направлению использовали многофакторный дисперсионный анализ (ANOVA).

Команды использовали множество различных подходов к предсказанию синергии лекарств, включая регрессию, деревья решений, случайные леса, Гауссовские процессы, SVM, нейронные сети. Однако производительность методов имела слабую связь с выбранным алгоритмом. Самым лучшим методом был Yuanfang Guan с метриками $p_{wSC1A}=0,48$, $p_{wSC1B} = 0,45$ и $ANOVASC2 = 74.89$ для обеих категорий первого направления соревнования и второго направления, соответственно. Основываясь на метрике для второго соревнования, метод Y Guan показал значительно лучшие результаты по сравнению с другими командами.

Также для второго направления конкурса был проверен ансамблевый

метод, основанный на агрегировании всех представленных моделей. Этим подходом добились скромного улучшения производительности по сравнению с лучшим методом. Это явление называется “мудрость толпы”[8].

Метод Yuanfang Guan

Более подробно разберем самую лучшую модель Yuanfang Guan[6]. Команда разработала три модели:

- модель глобальной синергии (GSM) - использует один обучающий набор и делает прогнозы для всех тестовых образцов сразу
- модель локальной синергии (LSM) - строит обучающий набор для каждого неизвестного показателя синергии в тестовом наборе данных и делает прогнозы отдельно. Обучающий набор данных LSM представляет собой подмножество пар лекарств в GSM, включая только те пары, когда появляется любой из препаратов в тестируемой паре лекарств.
- модель единого лекарственного средства (SDM) - SDM аналогичен LSM, за исключением того, что обучающий набор данных генерируется для каждого препарата, а не для комбинации препаратов

Каждая из моделей была разработана с использованием алгоритма случайного леса.

Для второго направления соревнований они использовали ROC-AUC, чтобы оценить эффективность бинарных предсказаний. Обозначили наблюдаемые синергетические оценки > 20 как "1" (наблюдаемая Синергия) и оценки < 20 как "0" (наблюдаемая несинергия). Этот порог использовался DREAM. Но, как было упомянуто ранее, для бинарных предсказаний консорциум DREAM использовал многофакторный дисперсионный анализ (ANOVA) для оценки предсказаний.

Метод **Ranking-system of Anti-Cancer Synergy**

Дополнительно рассмотрим модель, которую называли Ranking-system of Anti-Cancer Synergy (RACS)[12]. Эту модель обучена с частичным привлечением учителя. То есть используется датасет с малым числом комбинаций, которые известны как синергетические. Для создания обучающего набора данных использовали 26 синергетических пар противоопухолевых препаратов из базы данных Drug Combination Database (DCDB) в качестве меченых. В исследовании было охвачено 33 протестированных препарата с лекарственными мишенями и транскриптомными профилями конкретных клеточных линий после однократной обработки одиночным препаратом. На основе комбинирования 33 препаратов составили 502 пары немаркированных образцов (во всем обучающем датасете 528 пар).

На основе 13 соединений, предоставленных NCI-DREAM, было составлено 78 попарных комбинаций, которые были использованы в качестве тестового набора данных для клеточной линии -клеточной лимфомы человека OCI-LY3.

Также был создан еще один набор данных для валидации метода на основе 142 противоопухолевых препаратов, одобренных FDA или проходящих клинические испытания (из баз данных DrugBank, Therapeutic Target Database и PubMed). После удаления соединений без аннотаций gene ontology (GO) или информации о сигнальных путях базы данных KEGG осталось 118 препаратов. При их комбинировании было создано 6 877 немеченых пар в качестве тестовых данных для клеточной линии A549 и MCF7.

Изначально было выбрано четырнадцать признаков, охватывающих химическую структуру, фармакологию, функциональные и сетевые свой-

ства лекарственных мишеней, но только семь признаков были идентифицированы как существенно отличающиеся между синергическими и немечеными парами.

Для предварительного ранжирования был применен полууправляемый метод обучения, включающий многообразный алгоритм ранжирования на основе сходства с мечеными парами в пространстве 7 признаков.

Аналогичным образом были протестированы пять параметров, описывающих корреляции между дифференциально экспрессируемыми генами, и два параметра были значительно различны между синергическими и немечеными образцами. Эти два параметра были использованы в качестве дополнительных транскриптомных фильтров для улучшения предварительного ранжирования. Первый из них *DEG_Overlap* рассчитывается как:

$$DEG_Overlap = \frac{|A \cap B|}{\sqrt{|A| \times |B|}} \quad (4.17)$$

где *A* и *B* представляют собой множества дифференциально экспрессированных генов, возмущенные препаратом *x* и *y* соответственно.

Второй параметр назывался *Pathway_Coverage* и вычислялся по формуле:

$$Pathway_Coverage = \frac{|(A \cup B) \cap N|}{|N|} \quad (4.18)$$

где *A* и *B* представляют собой множества дифференциально экспрессированных генов, возмущенные препаратом *x* и *y* соответственно, *N* множество всех генов, связанных с сигнальными путями рака.

Ниже приведена схема метода RACS с предварительным ранжированием и транскриптомными фильтрами.

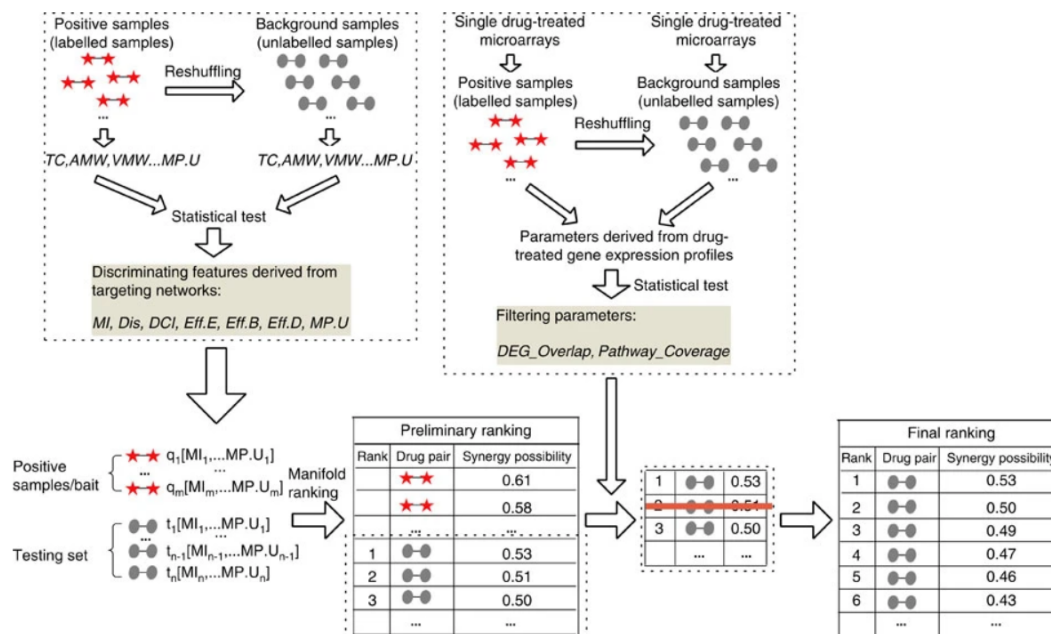


Рис. 4.5: схема RACS

Так как оценка метода была проведена на датасете консорциума DREAM, то метод сравнили с моделями соревнования, о которых было упомянуто выше (среди них выделялись такие методы, как DIGRE, SynGen). В качестве метрик брали ROC-AUC, TPR, PC-index.

У метод RACS PC-индекс равен 0,78, ROC-AUC 0,85, то есть этот метод показал наилучшие результаты. Именно применение транскриптомных фильтров увеличило PC-индекс с 0,69 до 0,78, а значения ROC-AUC с 0,783 до 0,853.

Также метод был проверен на наборе данных для клеточных линий A549 и MCF7, но с применением других метрик.

Стоит отметить, что PC-индекс, предложенный консорциум DREAM, более строгий по сравнению с ROC-AUC. PC-index устойчив к возмущениям данных. Например, значение ROC-AUC метода DIGER снизилось с 65 до 48% после удаления митомицина C из набора из 14 препаратов. Однако PC-индекс метода DIGER оставался почти неизменным. Также

он учитывает шум биологических копий. Однако РС-индекс нельзя было использовать для рака молочной железы или легких, потому что вместо биологических копий использовались множественные комбинации различных концентраций.

4.3.2 Методы с использованием других датасетов

DeepSynergy

Рассмотрим модель DeepSynergy[10], которая является методом глубокого обучения. Модель была разработана для задачи регрессии. DeepSynergy была обучена на наборе данных, который включает 23,062 образца, где каждый образец относится к двум соединениям и клеточной линии. Он охватывает 583 различные комбинации, каждая из которых была протестирована против 39 клеточных линий рака, полученных из 7 различных типов тканей. Пары препаратов были построены из 38 противоопухолевых препаратов (14 экспериментальных и 24 одобренных). Среди 38 соединений данные можно поделить на 2 набора:

- "исчерпывающий состоящий из 22 препаратов, у которых все возможные комбинации были протестированы
- "дополнительный состоящий из 16 препаратов, они были протестированы только в комбинации с препаратами из "исчерпывающего" набора

Для каждого образца была измерена скорость роста клеток при обработке комбинацией препаратов в режиме дозирования 4×4 для 4 биологических копий. Были получены данные монотерапии, то есть зависимость скорости роста клеток при обработке отдельным препаратом для 8 концентраций и 6 биологических копий. На основе этих данных составлялась матрица размером 5×5 , для которой первая строка и первый столбец

являются данными монотерапии. Такие матрицы назывались комбинационной поверхностью.

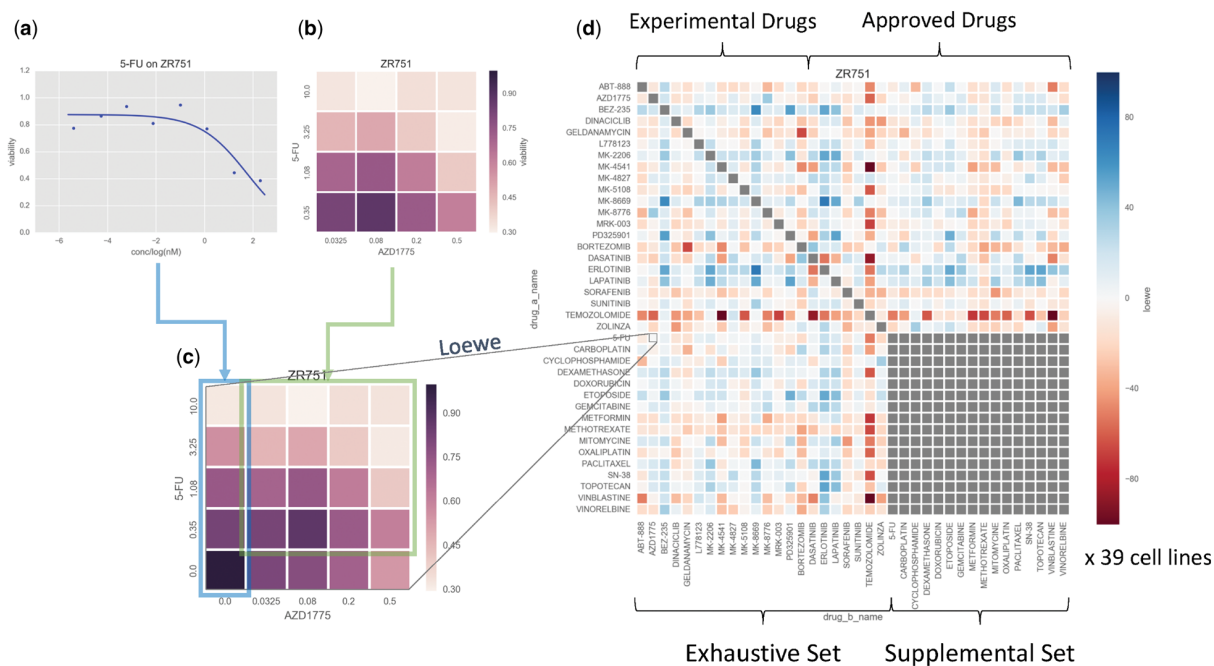


Рис. 4.6: Структура датасета

Уровень синергии комбинации препаратов вычислялся по комбинационным поверхностям с использованием теоретической модели Loewe Additivity[16].

Эффект от препаратов A и B с концентрациями a и b соответственно может быть представлен в виде следующей суммы :

$$E(a,b) = R(a,b) + S(a,b), \quad (4.19)$$

где $E(a,b)$ - наблюдаемый в эксперименте эффект, $R(a,b)$ - аддитивный эффект, определяемый по модели Loewe Additivity, $S(a,b)$ - это величина дополнительного эффекта, он положительный в случае синергетического эффекта и отрицательным в случае антогонистического эффекта пары препаратов. Аддитивный эффект для комбинации препаратов (a, b) вычисляется путем нахождения двух доз a_u и b_u таких, что:

$$E(a_u) = E(b_u) \quad (4.20)$$

Уравнение изоболы (кривой одинакового эффекта):

$$\frac{a}{a_u} + \frac{b}{b_u} = 1 \quad (4.21)$$

Найдя численное решение этих 2 уравнений, считают аддитивный эффект:

$$R_{Loewe}(a,b) = E(a_u) = E(b_u) \quad (4.22)$$

Зная аддитивный и наблюдаемый эффекты, можно посчитать синергетический. Значения синергии варьируются от -326 до 179.

Использовали как химическую информацию о препаратах, так и геномную информацию, отражающую биологию болезни. Авторы вычислили три различных типа химических признаков: extended connectivity fingerprints с радиусом 6, получены с использованием jCompoundMapper, физико-химические свойства с использованием ChemoPy, бинарные признаки основанные на наборе токсикофоров, собранных из литературы. Токсикофоры-это подструктуры, которые токсичны. Пространство химических признаков было уменьшено фильтрацией признаков с нулевой дисперсией. Финальный набор признаков содержал 1309 ECFP_6, 802 физико-химических и 2276 признаков токсикофоров. Клеточные линии были описаны профилем экспрессии генов. Конечный набор был из 3984 геномных признаков.

Модель DeepSynergy - это нейронная сеть с обратной связью, которая принимает входные векторы, представляющие образцы, и выдает одно значение - уровень синергии. Образцы описываются сцепленными векторами, которые включают в себя признаки двух препаратов и одной клеточной линии. То есть нейроны входного слоя получают значения экспрессии генов клеточной линии и химические дескрипторы обоих препаратов в качестве входных данных. Затем информация распространяется по слоям сети DeepSynergy до тех пор, пока выходной блок не выдаст

прогнозируемый показатель синергии.

Поскольку сеть не должна различать комбинацию лекарств АВ, представленную в порядке А-В или В-А, авторы удваивают измерения, представляя каждый образец дважды в обучающем наборе. Один раз свойства препарата используются в порядке А-В и один раз в порядке В-А. Для прогнозирования оба способа представления выборки распространяются по сети и усредняются. Наблюдалось, что DeepSynergy учится предсказывать одно и то же значение для комбинации лекарств АВ в порядке А-В и В-А.

Были рассмотрены различные настройки гиперпараметров, а именно разные стратегии нормализации данных, в сочетании с коническими или прямоугольными слоями с различным количеством нейронов. Использовалось два или три скрытых слоя. Кроме того, исследовались различные скорости обучения, а также методы регуляризации. В итоге DeepSynergy имеет коническую архитектуру с двумя скрытыми слоями, имеющими 8192 нейрона в первом и 4096 во втором скрытом слое.

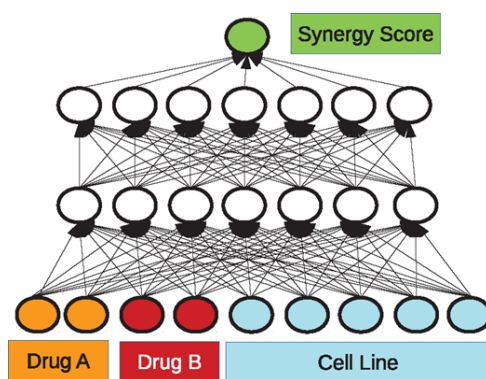


Рис. 4.7: Архитектура DeepSynergy

В качестве основной метрики оценки модели использовалась средне-квадратичная ошибка (MSE), по которой модель оптимизировалась во время обучения. Также использовали корень из среднеквадратичной ошибки (RMSE) и коэффициент корреляции Пирсона. DeepSynergy достигла

тестового MSE: 255, RMSE: 15.91, коэффициент корреляции Пирсона: 0.73. Дополнительно DeepSynergy сравнили со следующими моделями: Median Polish, которая использовалась в качестве базовой случайной модели, Gradient Boosting, Random Forests, Support Vector Machines и Elastic Nets. Ниже приведены результаты сравнения на основе выше перечисленных метрик.

Method	MSE	Confidence Interval	P-value	RMSE	Pearson's r
Deep Neural Networks	255.49	[239.93, 271.06]		15.91 ± 1.56	0.73 ± 0.04
Gradient Boosting Machines	275.39	[258.24, 292.54]	9.6×10^{-17}	16.54 ± 1.37	0.69 ± 0.02
Random Forests	307.56	[286.83, 328.29]	1.2×10^{-73}	17.49 ± 1.63	0.65 ± 0.03
Support Vector Machines	398.39	[371.22, 425.56]	$<10^{-280}$	19.92 ± 1.28	0.50 ± 0.03
Elastic Nets	420.24	[393.11, 447.38]	$<10^{-280}$	20.46 ± 1.29	0.44 ± 0.03
Baseline (Median Polish)	477.77	[448.68, 506.85]	$<10^{-280}$	21.80 ± 1.49	0.43 ± 0.02

Рис. 4.8: Сравнение моделей

DeepSynergy показала MSE, равный 255, в то время как Gradient Boosting, Random Forests достигли 275 и 308 соответственно. Support Vector Machines и Elastic Nets показали аналогичные результаты с MSE, равным 398 и 420 соответственно. Median Polish, которая использовалась в качестве базовой, достигла наихудшего результата с MSE 478. Относительное улучшение DeepSynergy по сравнению с базовой моделью составляет 53%. DeepSynergy значительно превосходит другие методы машинного обучения.

Так как RMSE, MSE зависят от набора данных, их трудно использовать для сравнения между различными наборами данных. Поэтому для получения сопоставимых показателей прогностических характери-

стик DeepSynergy использовались метрики, типичные для задач классификации: ROC AUC, PR AUC, точность (ACC), точность (PREC), чувствительность (TPR), специфичность (TNR). У DeepSynergy ROC AUC равен 0,9.

Метод Zhao H.

Для набора данных использовали 184 пары препаратов (на основе 238 препаратов), одобренные FDA orange book. Была собрана молекулярная и фармакологическая информация, связанная с этими препаратами, включая их мишени и соответствующие нисходящие сигнальные пути, области медицинских показаний, терапевтические эффекты, представленные в анатомической терапевтической и химической классификационной системе (ATX), а также побочные эффекты[9].

Для аннотаций лекарственных мишеней использовали данные о взаимодействии соединений и белков из базы данных STITCH, DrugBank и базы данных терапевтических мишеней TTD. Далее исследовали пути, на которые возможно воздействует препарат через мишень, информация о сигнальных путях была получена из базы данных KEGG. Каждый препарат был связан с путями, в которых состоят его мишени. Пара препаратов может быть представлена в виде вектора, состоящего из пар признаков. Например, в случае признака "мишень"препарат 1 связывает два белка p_1 , p_2 , препарат 2 связывает три белка p_3 , p_4 , p_5 , комбинация препарата 1 и препарата 2 может быть представлена в виде следующих пар признаков: (p_1, p_3) , (p_1, p_4) , (p_1, p_5) , (p_2, p_3) , (p_2, p_4) , (p_2, p_5) , аналогично для других признаков.

Затем был проведен поиск тех пар признаков, которые наиболее часто встречаются в синергетических парах. Для пары препаратов (d_i, d_j) признака F (например, мишень), f_i связан с d_i и f_j с d_j , где f_i и $f_j \subset F$.

Пара препаратов (d_i, d_j) может быть представлена парой признаков (f_i, f_j) . Для каждой пары признаков рассчитывается оценка обогащения s_{ij} следующим образом:

$$s_{ij} = \frac{N_{ij}}{N'_{ij}} \quad (4.23)$$

где N_{ij} - сколько раз пара признаков встречается в синергетических комбинациях лекарств, N'_{ij} - сколько раз пара признаков встречается в фоновом наборе всех возможных попарных комбинаций между лекарственными средствами, участвующих в синергетических комбинациях лекарств. Таким образом, все пары признаков могут быть ранжированы по s_{ij} , то есть по обогащению в синергетических парах.

Затем для выбранного признака F обучали модель предсказывать оценку обогащения для пары признаков комбинации препаратов. Для этого использовали кросс-валидация. В ходе кросс-валидации все комбинации препаратов были случайным образом разделены на пять групп одинакового размера без перекрытия, четыре из которых использовались в качестве обучающего набора и использовались для расчета оценки обогащения для каждой пары признаков, в то время как оставшаяся группа использовалась в качестве валидационного набора, и процедура повторялась в течение пяти раз. Оценка F1 была принята в качестве показателя производительности модели. Порог, выше которого был достигнут самый высокий балл F1 при перекрестной валидации, использовался для предсказания эффективности. То есть пара препаратов считалась синергетической, если оценка обогащения ее пары признаков была выше порогового значения. Таким образом рассмотренная задача относилась к классификации и для сравнения важности признаков использовали метрику ROC-AUC. По ней было отобрано наиболее три важных признака: область терапии, мишень и область показаний. Затем, интегрируя эти признаки, аналогичным образом предсказываем синергетические па-

ры[9].

Среди рассмотренных методов наибольшего результата достигли подходы, для которых был доступен обучающий набор данных. Методы, которые могли использовать только молекулярные данные клеточных линий, фармакологические признаки, данные монотерапии и гипотезы механизма синергии, показывали результаты немного выше случайных моделей[5], что говорит о плохом понимании механизмов, лежащих в основе синергии.

L1000CDS²

Рассмотрим еще один метод предсказания синергетических пар малых молекул, используемый в подходе *L1000CDS²*[17]. В основе работы этого инструмента лежит концепция Connectivity Map[18]. Рассмотрим 2 клеточных состояния, например здоровое и больное, либо 2 типа клеток. На основе данных RNA-seq для этих 2 состояний можно найти сигнатуру, состоящую из генов с повышенной и пониженной экспрессией. *L1000CDS²* производит поиск малых молекул, которые имитируют или обращают сигнатуру, подданую в качестве входных данных. Этот поиск осуществляется на основе сравнения поданной сигнатуры с сигнатурами клеточных линий, химически возмущенных, из базы данных LINCS-1000. Сигнатуры рассчитываются из данных экспрессии с помощью их собственного метода Characteristic Direction (CD). Для приоритизации малых молекул L1000CDS2 вычисляет косинусное расстояние между вектором входной сигнатуры и сигнатурами в LINCS-L1000.

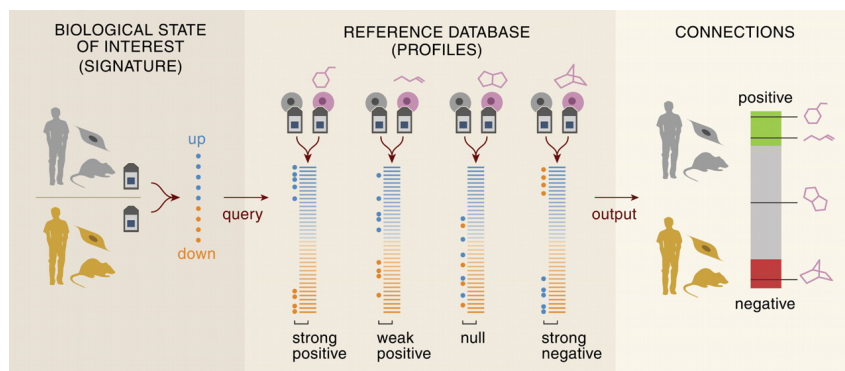


Рис. 4.9: Концепция Connectivity Map[18]

При поиске синергетических комбинаций L1000CDS2 сравнивает каждую возможную пару из 50 лучших подобранных сигнатур и вычисляет уровень синергии каждой пары, как ортогональность между двумя сигнатурами. Идея состоит в том, что если два возмущения ортогональны, то они воздействуют через два независимых пути и вероятно несут комплиментарный синергетический эффект.

5 Материалы и методы

6 Полученные результаты

7 Заключение. План дальнейших исследований

8 Благодарности

9 Список литературы

- [1] Johann Gasteiger и Thomas Engel. *Chemoinformatics: a textbook*. John Wiley & Sons, 2006.
- [2] Christopher Lipinski и Andrew Hopkins. «Navigating chemical space for biology and medicine». В: *Nature* 432.7019 (2004), с. 855—861.
- [3] Ling Xue и Jurgen Bajorath. «Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening». В: *Combinatorial chemistry & high throughput screening* 3.5 (2000), с. 363—372.
- [4] Lo и др. «Machine learning in chemoinformatics and drug discovery». В: *Drug discovery today* 23.8 (2018), с. 1538—1546.
- [5] Mukesh Bansal и др. «A community computational challenge to predict the activity of pairs of compounds». В: *Nature biotechnology* 32.12 (2014), с. 1213—1222.
- [6] Hongyang Li и др. «Network propagation predicts drug synergy in cancers». В: *Cancer research* 78.18 (2018), с. 5446—5457.
- [7] Lei Huang и др. «Driver network as a biomarker: systematic integration and network modeling of multi-omics data to derive driver signaling pathways for drug combination prediction». В: *Bioinformatics* 35.19 (2019), с. 3709—3717.
- [8] Martin Kuiper и др. «Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen». В: (2019).
- [9] Xing-Ming Zhao и др. «Prediction of drug combinations by integrating molecular and pharmacological data». В: *PLoS Comput Biol* 7.12 (2011), e1002323.

-
- [10] Kristina Preuer и др. «DeepSynergy: predicting anti-cancer drug synergy with Deep Learning». в: *Bioinformatics* 34.9 (2018), с. 1538—1546.
- [11] Francesco Napolitano и др. «Automatic identification of small molecules that promote cell conversion and reprogramming». в: (2020).
- [12] Yi Sun и др. «Combining genomic and network characteristics for extended capability in predicting synergistic drugs for cancer». в: *Nature communications* 6.1 (2015), с. 1—10.
- [13] Pingjian Ding и др. «Incorporating Multisource Knowledge To Predict Drug Synergy Based on Graph Co-regularization». в: *Journal of Chemical Information and Modeling* 60.1 (2019), с. 37—46.
- [14] Celine Lefebvre и др. «A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers». в: *Molecular systems biology* 6.1 (2010), с. 377.
- [15] Maria Stella Carro и др. «The transcriptional network for mesenchymal transformation of brain tumours». в: *Nature* 463.7279 (2010), с. 318—325.
- [16] Giovanni Y Di Veroli и др. «Combeneft: an interactive platform for the analysis and visualization of drug combinations». в: *Bioinformatics* 32.18 (2016), с. 2866—2868.
- [17] Qiaonan Duan и др. «L1000CDS 2: LINCS L1000 characteristic direction signatures search engine». в: *NPJ systems biology and applications* 2.1 (2016), с. 1—12.
- [18] Justin Lamb и др. «The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease». в: *science* 313.5795 (2006), с. 1929—1935.