Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

на правах рукописи

Мошков Никита Евгеньевич

ПРИМЕНЕНИЕ АЛГОРИТМОВ ГЛУБОКОГО ОБУЧЕНИЯ ДЛЯ СЕГМЕНТИРОВАНИЯ ОДИНОЧНЫХ КЛЕТОК И ФЕНОТИПИЧЕСКОГО ПРОФИЛИРОВАНИЯ

Резюме диссертации

на соискание учёной степени кандидата компьютерных наук

Научные руководители:
PhD, Кертес-Фаркаш Аттила,
PhD, Хорват Питер,
PhD, Кайседо Хуан

Москва - 2022

Содержание

1 Введение **4**

1.1 Актуальность исследований 4

1.2 Задачи диссертации 5

1.2.1 Обзор существующих методов сегментирования клеточных ядер . . . 5

1.2.2 Аннотация клеточных ядер с помощью глубинного обучения 5

1.2.3 Оценить подход аугментации в режиме тестирования для сегментирования клеточных ядер 5

1.2.4 Морфологическое профилирование на основе изображений с помощью глубинного обучения 6

1.2.5 Оценка различных источников признаков для скрининга лекарств . . . 6

1.3 Важность представленной работы 6

1.4 Публикации 7

2 Справочная информация **9**

2.1 Нейронные сети для сегментирования клеточных ядер и одиночных клеток . . 9

2.1.1 U-Net 10

2.1.2 Mask R-CNN 10

2.2 Cell Painting и фенотипическое профилирование 11

2.3 Вычислительные методы в химической биологии 12

3 Краткое изложение исследований **14**

3.1 Сегментирование клеточных ядер: на пути к автоматизированным решениям . 14

3.2 AnnotatorJ: плагин ImageJ для упрощения ручного аннотирования клеток и их ядер 16

3.3 Аугментация во время тестирования для сегментирования клеточных ядер методами глубинного обучения в изображениях микроскопии 19

3.3.1 Аугментация во время тестирования 19

3.3.2 Материалы и методы 20

3.3.3 Результаты 22

3.4 Обучение представлениям для профилирования пертурбаций на основе изображений 26

3.4.1 Наборы данных Cell Painting 27

3.4.2 DeepProfiler 28

3.4.3 Экспериментальная установка 29

3.4.4 Рабочий процесс и оценка профилирования 31

3.4.5 Выбор сильных пертурбаций и комбинированный набор данных Cell Painting 33

3.4.6 Причинно-следственные связи в скрининговых экспериментах 34

3.4.7 Результаты и наблюдения 35

3.5 Прогнозирование активности соединений на основе фенотипических профилей и химических структур 39
3.5.1 Материалы и методы 40
3.5.2 Эксперименты и результаты 42

4 Выводы **50**

Список иллюстраций **53**

Список таблиц **58**

Список литературы **59**

1 Введение

1.1 Актуальность исследований

Решающим элементом в решении фундаментальных вопросов биологии и разработке эффективных методов лечения заболеваний является понимание клеточных молекулярных процессов [1]. Анализ отдельной клетки стал одной из важнейших задач в области естественных наук в 21 веке. Принципиально новая идея [2] заключается в том, чтобы рассматривать каждую отдельную клетку в тканях как отдельный строительный блок со своим состоянием и, следовательно, рассматривать ткани как разнообразный набор таких строительных блоков, а не как однородную сущность. Средством для широкого исследования этой идеи стали новые высокопроизводительные технологии секвенирования генома, протеомики, метаболомики и визуализации.

Такие достижения сделали возможным автоматический и объективный анализ в масштабах, достигающих миллионов и миллиардов клеток, что дало нам возможность проводить высокопроизводительные эксперименты с одиночными клетками (визуализация живых клеток [3][4], профилирование экспрессии генов [5] и протеомика [6]), а затем провести анализ с помощью вычислительных методов, применимых к полученному типу данных, и попытаться извлечь из этих данных биологический смысл.

Различные типы данных (или модальности данных) могут позволить нам изучить состояние каждой конкретной клетки с разных точек зрения. Одной из практических задач, где вся возможная информация может быть полезна для принятия решений, является поиск лекарств, особенно в персонализированной медицине. Самой большой проблемой является точное и экономически эффективное сочетание и использование существующих дорогостоящих методов лечения.

Здесь мы сосредоточимся в основном на данных визуализации и одним из первых шагов анализа изображений одиночных клеток является *сегментирование клетки или ядра* – классификация каждого пикселя как фона или предмета переднего плана (семантическое сегментирование), или определение принадлежности пикселя к определенному объекту (сегментирование объекта), примеры приведены на рисунке 1. В последние годы эта область развивается благодаря внедрению и созданию алгоритмов глубинного обучения для этой задачи, что привело к значительным улучшениям точности сегментирования [7].

За сегментированием может последовать идентификация биологических фенотипов через количественную оценку морфологии клеток, изменение которой может показать, например, различия между клетками, прошедшими и не прошедшими через воздействие химических соединений в экспериментах по скринингу лекарств [8]. Фенотипы могут быть описаны векторами признаков, также называемыми *профилями*, а процесс их извлечения называется профилированием, а морфологическое профилирование также может называться *профилированием на основе изображений*. [9] [10].

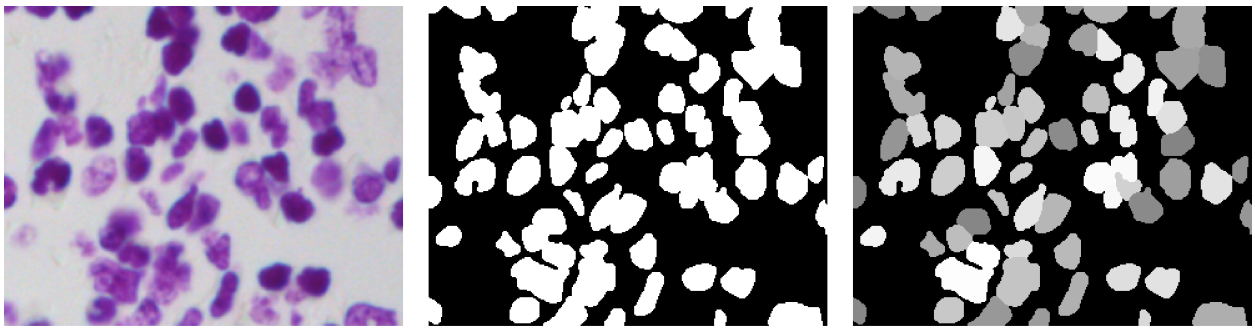


Рис. 1: Примеры сегментирования слева направо: исходное изображение, семантическое сегментирование, сегментирование отдельных объектов. Источник изображения и аннотации сегментирования: набор данных Data Science Bowl 2018 [11].

1.2 Задачи диссертации

1.2.1 Обзор существующих методов сегментирования клеточных ядер

Предварительная обработка изображения и сегментирование клетки (или его ядра) обычно являются первыми этапами анализа изображений одиночных клеток. Точное сегментирование влияет на качество последующего анализа, поэтому этот шаг является чрезвычайно важным.

Автор диссертации участвовал в написании обзорной статьи [7], в которой собраны данные о состоянии области сегментирования клеточных ядер в 2020-2021 годах. Помимо методов сегментирования для 2D и 3D данных, в ней также рассматриваются методы предварительной и последующей обработки, существующие наборы данных и инструменты для аннотирования изображений клеток (или их ядер).

1.2.2 Аннотация клеточных ядер с помощью глубинного обучения

Для обучения моделей сегментирования одиночных клеток (или их ядер) на основе глубинного обучения необходимы аннотированные данные, и, чем больше набор данных, тем более робастной будет модель. Ручное аннотирование является дорогостоящим процессом, поскольку требует значительного количества времени и усилий со стороны экспертов в области биологии. Чтобы сделать процесс аннотирования более быстрым и точным, был разработан плагин AnnotatorJ [12] для ImageJ/FIJI [13] (программное обеспечение для анализа биологических изображений), который объединяет идентификацию одиночных клеток (и их ядер) с глубинным обучением и ручным аннотированием.

1.2.3 Оценить подход аугментации в режиме тестирования для сегментирования клеточных ядер

Аугментация в режиме тестирования (test-time augmentation) была существующим подходом для улучшения классификации изображений [14]. В этой диссертации тестируется аугментация в режиме тестирования для сегментирования клеточных ядер. Обученная модель глубинного обучения для сегментирования обрабатывает исходное входное изображение и

несколько преобразованных вариантов того же изображения. Затем полученные результаты сегментирования объединяются. Основная идея заключается в том, что комбинация результатов сегментирования исходного изображения и его преобразованных вариантов будет работать лучше, чем сегментирование только исходного изображения или, по крайней мере, даст нам подсказки о неопределенностях в сегментировании. Конечным результатом здесь является экспериментальная оценка этого подхода для двух популярных сетей глубинного обучения для сегментирования.

1.2.4 Морфологическое профилирование на основе изображений с помощью глубинного обучения

Исследуется использование моделей глубинного обучения для профилирования на основе изображений (фенотипирование одиночных клеток). Эти модели глубинного обучения могут быть либо предварительно обучены (с помощью набора данных ImageNet [15]), либо обучены (со слабой разметкой) для конкретного набора данных одиночных клеток. Используя эти модели, можно извлечь характеристики (профили) одиночных клеток. Полученные характеристики используются в последующем анализе (например, для предсказания механизмов действия лекарств). Мы исследуем, обеспечивают ли признаки, полученные с помощью сетей глубинного обучения, лучшие результаты при последующем анализе, чем классические морфологические признаки [16], в частности для изображений, полученных с помощью Cell Painting [10] (также см. раздел 2.2).

1.2.5 Оценка различных источников признаков для скрининга лекарств

Относительная предсказательная сила сравнивается для трех источников признаков: представления химических структур [17] веществ, профилей экспрессии генов, полученных с помощью метода L1000 [18] и морфологических профилей на основе изображений, полученных с помощью Cell Painting [10], обработанных с помощью CellProfiler [19] для задачи предсказания активности веществ.

1.3 Важность представленной работы

Обзор [7] (Задача 1.2.1) новейших методов 2D и 3D сегментирования дает практикам представление об использовании и наиболее подходящих методах сегментирования для различных видов микроскопии. Поскольку конечными пользователями методов сегментирования обычно являются биологи, рекомендации по выбору наиболее эффективной и простой в использовании системы могут оказаться полезными для сообщества, так как точное сегментирование чрезвычайно важно для последующих задач.

Использование алгоритмов глубинного обучения невозможно без точных аннотированных наборов данных изображений, а в области сегментирования ядер такие наборы данных обычно создаются экспертами. Мы разработали инструмент [12] (Задача 1.2.2), чтобы сделать создание аннотированных наборов данных быстрее, удобнее и, следовательно, дешевле.

Одним из возможных способов получения лучшего результата сегментирования является применение методов постобработки. Одним из таких потенциальных методов является аугментация в режиме тестирования, которое традиционно используется для классификации изображений. Систематическая оценка [20] (Задача 1.2.3) этого метода в задаче сегментирования клеточных ядер для самых популярных фреймворков глубинного обучения и самого популярного на сегодняшний день набора данных ядер дает представление о его полезности.

Основной целью морфологического профилирования на основе изображений является получение такого представления (вектора признаков), которое точно отражает состояние клетки [21]. Сети глубинного обучения для классификации изображений могут быть способны захватывать такие представления, особенно с шагами постобработки, такими как агрегирование. Морфологическое профилирование изображений на основе глубинного обучения в сочетании с экономически эффективным методом Cell Painting [10] может быть использовано для поиска лекарств и других биологически важных вопросов (Задача 1.2.4).

Помимо морфологии, профили экспрессии генов и представления химических структур [17] могут использоваться для извлечения полезной информации в задаче поиска лекарств. Сравнение (Задача 1.2.5) их предсказательной силы может дать представление и продемонстрировать пригодность моделей машинного обучения для ранних стадий процесса поиска лекарств.

1.4 Публикации

Статьи, связанные с темой исследования:

- **Moshkov N.**, Mathe B., Kertesz-Farkas A., Hollandi R., Horvath P. Test-time augmentation for deep learning-based cell segmentation on microscopy images. Scientific Reports. 2020. Vol. 10, 5068. Q1 journal, IF 3.998 (2020).
  DOI: `https://doi.org/10.1038/s41598-020-61808-3`

- Hollandi R.*, **Moshkov N.***, Paavolainen L., Tasnadi E., Piccinini F., Horvath P. Nucleus segmentation: towards automated solutions. Trends in Cell Biology. 2022. Q1 journal, IF 20.808 (2021). DOI: `https://doi.org/10.1016/j.tcb.2021.12.004`

- Hollandi R., Diosdi A., Hollandi G., **Moshkov N.**, Horvath P. AnnotatorJ: an ImageJ plugin to ease hand-annotation of cellular compartments. Molecular Biology of the Cell. 2020 Vol. 31. № 20. P. 2157-2288. Q1 journal, IF 3.791 (2020). DOI: `https://doi.org/10.1091/mbc.E20-02-0156`

Препринты, связанные с темой исследования:

- **Nikita Moshkov**, Tim Becker, Kevin Yang, Peter Horvath, Vlado Dancik, Bridget K. Wagner, Paul A. Clemons, Shantanu Singh, Anne E. Carpenter, Juan C. Caicedo. Predicting compound activity from phenotypic profiles and chemical structures bioRxiv 2020.12.15.422887, DOI: `https://doi.org/10.1101/2020.12.15.422887`

- **Nikita Moshkov**, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca Senft, Yu Han, Mehrtash Babadi, Peter Horvath, Beth A. Cimini, Anne E. Carpenter, Shantanu Singh, Juan C. Caicedo. Learning representations for image-based profiling of perturbations. bioRxiv 2022.08.12.50378, DOI: https://doi.org/10.1101/2022.08.12.503783

Конференции, связанные с темой исследования:

- HEPTECH AIME19 AI & ML (2019). Test-time augmentation for deep learning-based cell segmentation on microscopy images (постер). Ссылка: https://indico.wigner.hu/event/1058/contributions/2542/

Статьи, не связанные с целью исследования 2017-2022:

- **Moshkov N.***, Smetanin A.*, Tatarinova T. Local ancestry prediction with PyLAE. PeerJ. 2021. Article 12502. Q2 journal, IF 2.816. DOI: https://doi.org/10.7717/peerj.12502

- Piccini F., Balassa T., Carbonaro A., Diosdi A., Toth T., **Moshkov N.**, Tasnadi E. A., Horvath P. Software tools for 3D nuclei segmentation and quantitative analysis in multicellular aggregates. Computational and Structural Biotechnology Journal. 2020. Vol. 18. P. 1287-1300. IF 6.018 (2020), Q1 journal. DOI: https://doi.org/10.1016/j.csbj.2020.05.022

- Grexa I., Diosdi A., Harmati M., Kriston A., **Moshkov N.**, Buzas K., Pietiäinen V., Koos K., Horvath P. SpheroidPicker for automated 3D cell culture manipulation using deep learning. Scientific Reports. 2021. Vol. 11, 14813. Q1 journal, IF 4.379 (2021). DOI: https://doi.org/10.1038/s41598-021-94217-1

- Kornienko I. V., Faleeva T. G., Schurr T. G., Aramova O. Y., Ochir-Goryaeva M. A., Batieva E. F., Vdovchenkov E. V., **N. E. Moshkov**, Kukanova V. V., Ivanov I. N., Sidorenko Y. S., Tatarinova T. V. Y-Chromosome Haplogroup Diversity in Khazar Burials from Southern Russia. Russian Journal of Genetics. 2021. Vol. 57. No. 4. P. 477-488. IF 0.581. DOI: https://doi.org/10.1134/S1022795421040049

Конференции, не связанные с темой исследования:

- HEPTECH AIME ML&VA on Clouds (2018). Image database generation techniques for DIC brain tissue cell segmentation (постер). Ссылка: https://indico.wigner.hu/event/904/contributions/1874/

2 Справочная информация

2.1 Нейронные сети для сегментирования клеточных ядер и одиночных клеток

История автоматизированных подходов к сегментированию клеток и их ядер началась около 60 лет назад, и самые первые методы были основаны исключительно на пороговом определении яркости [22]. В течении очень долгого времени пороговое определение яркости (пример на рисунке 2) было доминирующим подходом, являясь единственной частью систем сегментирования или комбинируясь с другими классическими подходами. Позже, незадолго до эры глубинного обучения, появились другие подходы к сегментированию клеточных ядер, основанные на классическом машинном обучении [23], активных контурах [24] [25] и многослойной модели "gas of circles"[26]. Сложность биологических вопросов вместе с данными, подлежащими анализу (в биологии развития [27], поиске лекарств [28], функциональной геномике [29] и патморфологии [30]) стали требовать более точного сегментирования клеток (и из ядер), и область начала искать более общие решения задачи для этой задачи. Применение сверточных нейронных сетей и доступность вычислительных ресурсов для их обучения позволило нам перейти к таким решениям.

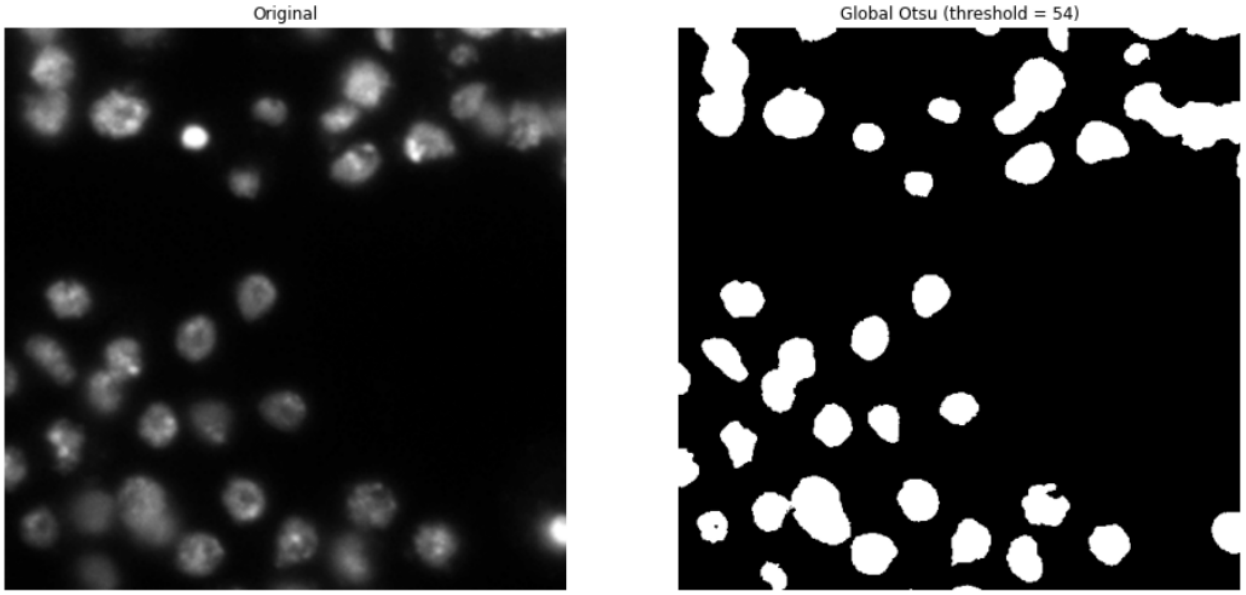


Рис. 2: Микроскопический снимок и маска сегментирования, полученная с помощью пороговой обработки Otsu [31] из пакета Scikit-image [32]. Источник изображения: набор данных Data Science Bowl 2018 [11].

Одним из первых этапов анализа изображений одиночных клеток является обнаружение/сегментирование клеток и/или их ядер. ***Сегментирование одиночных клеток*** (и сегментирование изображений в целом) является чрезвычайно развивающейся областью: с увеличением производительности GPU (графических процессоров) и нейронных сетей глубинного обучения, таких как U-Net [33] (см. также 2.1.1), которая стала прорывом в сегментировании биологических изображений на основе глубинного обучения (и в области сегментиро-

вания на основе глубинного обучения в целом). Этот подход до сих пор служит базовым для задач семантического сегментирования (т.е. классификации пикселей) и используется как часть недавних систем сегментирования клеток и их ядер, таких как CellPose [34], StarDist [35] и их производных. Помимо специализированных методов для сегментирования клеток, методы, изначально разработанные для сегментирования естественных изображений, такие как Mask R-CNN [36] (см. также 2.1.2), также применяются для задач сегментирования одиночных клеток либо с помощью дообучения, либо как часть сложного алгоритма сегментации [37].

Помимо самих сетей глубинного обучения, существуют общие методы регуляризации обучения и, следовательно, для обучения более робастных моделей, такие как аугментация данных для обучения (модификация исходных обучающих данных путем поворота, переворачивания или добавления шума) [38], dropout слои [39], L1 или L2 регуляризация [40]. Задача сегментирования одиночных клеток (ядер) не является исключением для использования этих методов.

2.1.1 U-Net

U-Net [33] (Рисунок 3) - это архитектура на основе глубинного обучения, разработанная в 2015 году в основном для семантического сегментирования биологических изображений (также победитель соревнования ISBI по трекингу клеток). Свое название она получила от архитектуры U-образного кодера-декодера: входные данные сначала сжимаются сверточными слоями, а затем расширяются до исходного размера. U-Net до сих пор широко используется в качестве базового метода к сегментированию клеточных ядер, и существует множество методов, основанных на данной архитектуре для различных наборов данных [7].

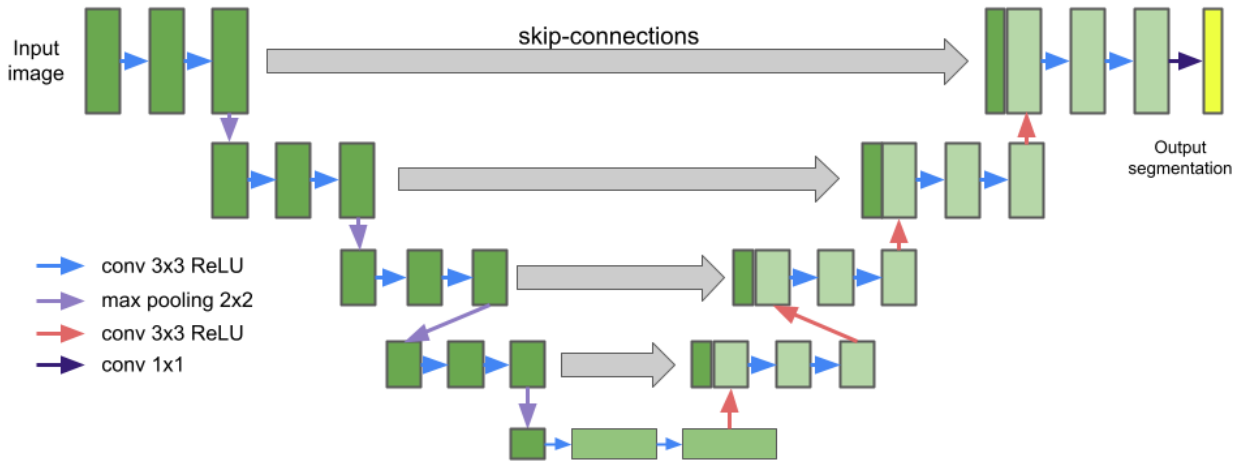


Рис. 3: Стандартная архитектура U-Net.

2.1.2 Mask R-CNN

Mask R-CNN [36] (Рисунок 4) была разработана в 2017 году для сегментирования отдельных объектов (каждый пиксель на изображении присваивается отдельному объекту) в естествен-

ных изображениях. Mask R-CNN использует архитектуру ResNet [41] в качестве базовой (обычно ResNet50 или ResNet101), за которой следует Region Proposal Network (RPN). Это первый этап Mask R-CNN, который завершается набором предложенных регионов с объектами.

RoIAling (RoI - region of interest) является одним из ключевых усовершенствований Mask R-CNN по сравнению с Faster R-CNN [42], которая использует RoI pooling. Обе эти операции извлекают RoIs из карт признаков, но RoIAling более точен. За ним следуют верхние слои: они предсказывают класс, рамки объекта и выходную бинарную маску для каждого RoI. Классы не учитываются при генерации масок. RoIAling и верхние слои являются вторым этапом Mask R-CNN.

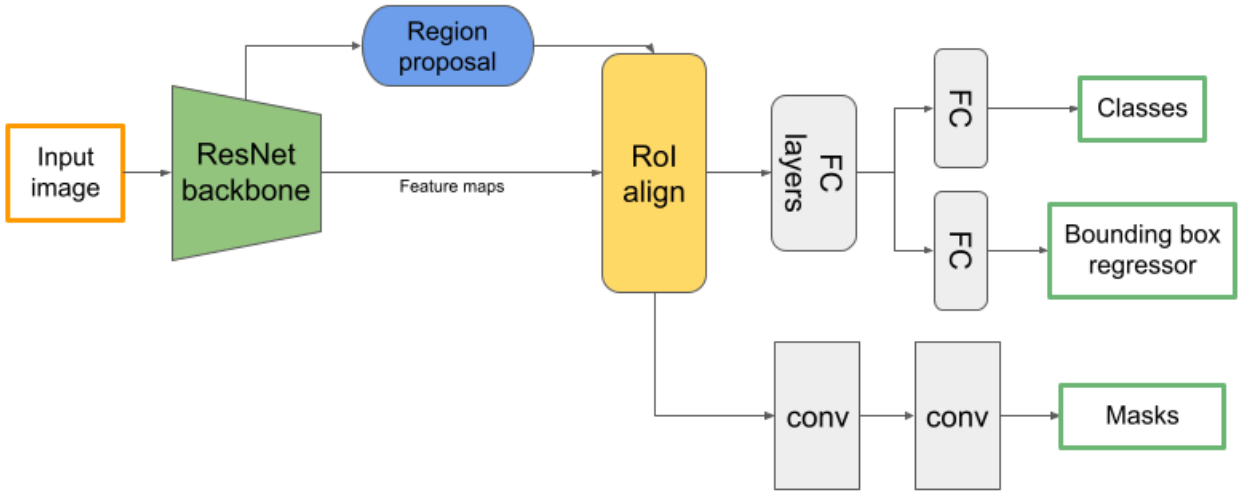


Рис. 4: Стандартная архитектура Mask R-CNN.

2.2 Cell Painting и фенотипическое профилирование

Ранее в открытии лекарств доминировал *таргетный* подход, но сейчас преимущество отдается *фенотипическому* подходу к открытию лекарств [43]. Таргетный подход к открытию лекарств сосредоточен на поиске мишеней - продуктов генов, которые являются отправной точкой для исследования, а затем исследователи вырабатывают представление о том, как на них воздействовать [44]. Фенотипический подход к открытию лекарств является эмпирическим: большое количество веществ тестируется в мишень-независимый анализах и отслеживается фенотипическая вариация [45]. Фенотипическое открытие лекарств расширило пространство поиска лекарств, мишеней и механизмов действия, сделав возможным их открытие [46].

Одним из способов идентификации фенотипической вариации является количественная оценка морфологии клеток, которая может продемонстрировать различия между обработанными и необработанными клетками в экспериментах по скринингу лекарств [8] [9]. Эффективным методом анализа для поиска лекарств на основе фенотипов является метод Cell Painting. Этот метод был разработан, чтобы зафиксировать как можно больше биологически значимых морфологических признаков, при этом этот протокол совместим с существую-

щими системами микроскопии и в то же время относительно не дорог [10]. Получаемые изображения являются пятиканальными и захватывают восемь клеточных органоидов (см. рисунок 5).

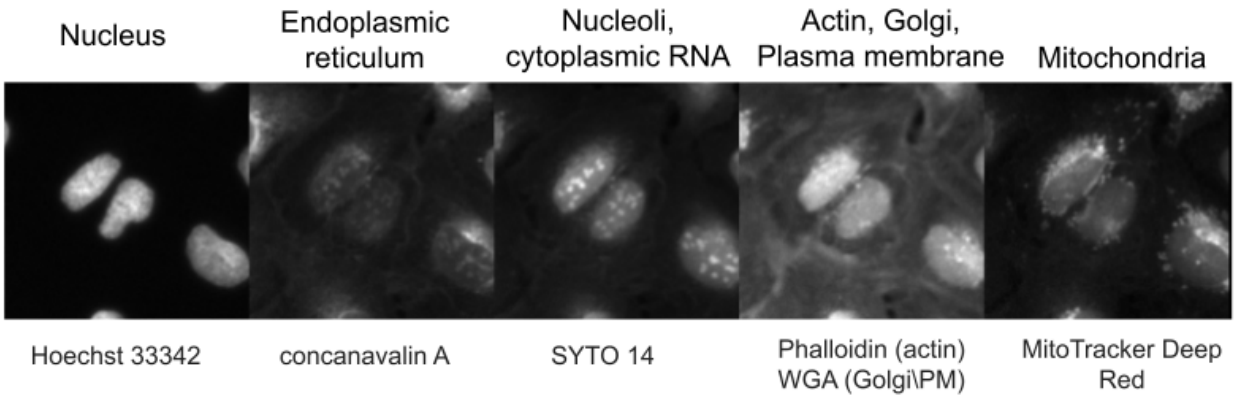


Рис. 5: Пример изображения, полученного методом Cell Painting, органоиды (подписи сверху) и красители (подписи внизу). Изображение из набора данных BBBC022 [47].

Морфология клеток может быть описана вектором признаков - или *профилем* (либо для одиночных клеток, либо в совокупности для популяции клеток), извлекаемым многоэтапным алгоритмом [48]. Эта задача может быть названа ***морфологическим профилированием*** [9] [48] или более широким термином ***фенотипическое профилирование*** [49]. Извлеченные профили обрабатываются в дальнейшем анализе. Наиболее популярным программным обеспечением для получения морфологических профилей одиночных клеток является CellProfiler [19], признаки разрабатываются вручную, хотя признаки, полученные с помощью моделей глубинного обучения, подлежат исследованию [50] [51] [52].

CellProfiler [19] - программное обеспечение с открытым исходным кодом для количественного анализа фенотипов клеток. Оно предназначено для биологов-аналитиков, поэтому не требует специализированного опыта в области информатики, биолог-аналитик разрабатывает только схему анализа с модулями и их настройками, а для определенных типов данных доступны лучшие практики (`https://cellprofiler.org/published-pipelines`). На выходе CellProfiler получается вектор признаков с человекочитаемыми признаками, которые могут быть организованы в группы, такие как яркость, текстура и форма.

2.3 Вычислительные методы в химической биологии

Областью, тесно связанной с поиском лекарств, является химическая биология, которая изучает взаимодействие малых молекул (лекарства обычно являются малыми молекулами) с биологическими системами (например, отдельными клетками, тканями, организмами). Как и любая другая область, химическая биология имеет свой набор вычислительных методов для различных задач [53] [54].

Первая проблема, которую необходимо решить, - это представить химические молекулы в удобном и эффективном для вычислительных методов виде. Одно из самых простых таких

представлений – SMILES (Simplified Molecular Input Line Entry System)[55], эффективное и широко используемое в настоящее время. Пример приведен на рисунке 6.

Другим классом представления молекул являются молекулярные фингерпринты - двоичные или числовые векторы размером от 16 до 16384. Молекулярные фингерпринты могут быть основаны на правилах или получены с помощью методов глубинного обучения, эффективность этих представлений не одинакова [56]. Наиболее часто используемые молекулярные фингерпринты – это фингерпринты Моргана [57], которые представляют собой бинарные векторы.

Другим термином, связанным с представлением соединений, является скаффолд. Скаффолд - это основная структура соединения, которая состоит из всех кольцевых структур и связей между ними, предложенная Бемисом и Мурко [58], пример на рисунке 6.

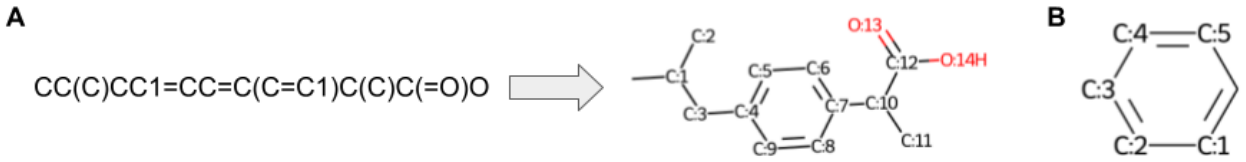


Рис. 6: A. SMILES представление Ибупрофена и его сгенерированное графическое представление. B. Скаффолд Бемиса-Мурко молекулы Ибупрофена. графическое представление и скаффолд были сгенерированы программным пакетом RDKit (`https://www.rdkit.org/`).

За последние несколько лет появились различные подходы глубинного обучения для вычислительной химии, основанные на сверточных или рекуррентных нейронных сетях, автоэнкодерах, графовых сверточных сетях [54].

Одним из заметных недавних методов, основанных на графовых сверточных сетях, является Chemprop (`http://chemprop.csail.mit.edu/`)[17] [59] [60]. Он принимает SMILES-строки в качестве входных данных (можно использовать и другие векторы признаков) и реконструирует молекулярные графы, где атомы являются узлами, а связи - ребрами. Затем применяется серия шагов передачи сообщений для объединения информации от соседних атомов и связей, для более точного представления молекулы.

3 Краткое изложение исследований

Этот раздел содержит краткое описание исследовательских проектов и их результатов. Часть подробностей не приводится, однако их можно найти в соответствующих публикациях.

3.1 Сегментирование клеточных ядер: на пути к автоматизированным решениям

В этом разделе кратко пересказывается [7].

Область сегментирования клеточных ядер развивалась в течение последних нескольких лет с помощью глубинного обучения. Практики начали широко использовать методы сегментирования на основе глубинного обучения, особенно после соревнования DSB 2018 [11], которое ясно показало превосходство методов на основе глубинного обучения над классическими. Кроме того, вычислительные ресурсы теперь более доступны, а методы стали более дружественными к пользователю за счет предоставления руководств по использованию, а иногда и графических пользовательских интерфейсов. Цель данного обзора - представить описания методов и наборов данных, связанных с сегментированием клеточных ядер, и помочь специалистам в этой области.

Поскольку методы глубинного обучения требуют данных для обучения, мы начинаем обзор с описания открытых аннотированных наборов данных клеточных ядер, как в 2D, так и в 3D и для различных видов микроскопии. Аннотации для этих наборов данных представлены в виде масок фона и переднего плана (BG-FG) или в виде масок объектов (когда каждый объект аннотирован отдельно). Первое наблюдение заключается в том, что не так много аннотированных наборов данных находятся в общем доступе, особенно 3D данных. Возможно, причина в том, что лаборатории начали массово переходить на 3D не так давно, к тому же использование 3D вместо 2D не всегда является необходимостью. Примером 3D набора данных, который может быть использован в качестве тестового (и фактически уже используется), является A549-Dataset [61]. Еще одно наблюдение - очень немногие типы микроскопии хорошо представлены даже в случае двухмерных наборов данных. Большинство наборов данных содержат только флуоресцентные, яркопольные или окрашенные гематоксилином и эозином (H&E) изображения. Заметным исключением является крупный набор данных LIVECell [62].

За частью обзора наборов данных следует часть об инструментах аннотации. Большинство из этих инструментов были выпущены недавно. Мы отметили наличие открытых и бесплатных инструментов для аннотирования 2D и 3D данных.

Последняя часть обзора посвящена методам и инструментам сегментирования. Рассмотренные методы сегментирования были классифицированы с использованием значимых критериев для практиков. Во-первых, методы были классифицированы по размерности входного изображения (2D, 3D или как 2D, так и 3D). Затем для каждого метода проверялась доступность кода. Другим важным критерием является наличие расширенных версий руководств пользователя, поскольку пользователи методов сегментирования для биологических задач

не обязательно обладают знаниями в области информатики и нуждаются в четком пошаговом руководстве по использованию этих методов. Последний важный критерий - работает ли инструмент в вычислительном облаке, что стало очень распространенным сценарием для выполнения задач, требующих больших вычислительных затрат.

Еще одним вкладом данного обзора является ассистент для выбора метода сегментирования клеточных ядер (называется *unbiased*), который доступен онлайн на GitHub Pages `https://biomag-lab.github.io/microscopy-tree/`. Он должен помочь в выборе потенциально полезных методов на основе вида микроскопии, размерности изображений и потенциальных сложностях в интересующих данных.

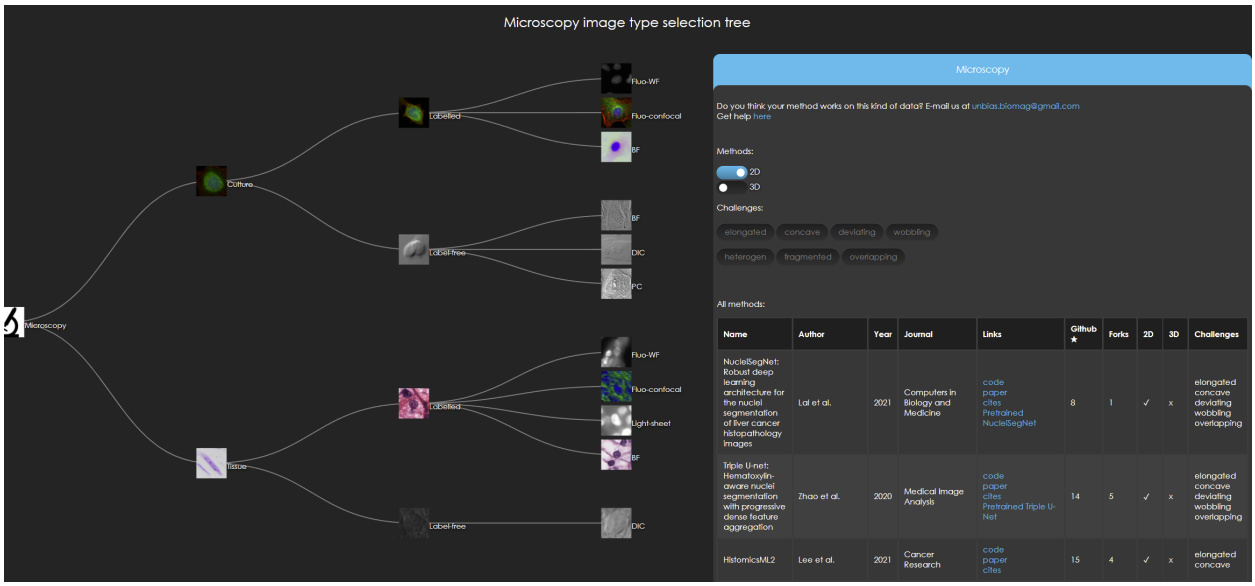


Рис. 7: Интерфейс ассистента для выбора методов сегментирования. Слева расположено дерево типов микроскопии. В правом верхнем углу расположены элементы управления фильтрацией для выбора 2D/3D-методов и конкретных методов для решения сложностей сегментирования. В правом нижнем углу находится список методов сегментирования.

Главным результатом обзора стало выражение сомнений и вопросов относительно текущего состояния этой области. Первая проблема связана с недостаточным разнообразием существующих наборов данных с точки зрения видов микроскопии. Оказалось, что большинство открыто опубликованных аннотированных наборов данных относятся либо к изображениям H&E, либо к флуоресцентным изображениям. Другие способы микроскопии (например, DIC (дифференциально-интерференционный контрастная микроскопия), light-sheet или phase contrast) слабо представлены в общедоступных наборах данных. Кроме того, размер опубликованных наборов данных также имеет значение, большинство наборов данных не содержат большого количества объектов и изображений.

Еще один момент - это призыв к решению общих проблем при сегментировании клеточных ядер, таких как соприкосновение, наложение и неправильная форма ядер [35] [63] [64] [65]. Существующие методы глубинного обучения способны частично решить эти проблемы, но необходим более значительный прогресс. В этом отношении положительное влияние

могут оказать как новые архитектуры моделей, так и большие наборы данных для обучения.

Настоящей проблемой, которая лежит на поверхности, но редко обсуждается, является отсутствие единого подхода к оценке методов сегментирования клеточных ядер. После изучения всех методов, представленных в обзоре, стало ясно, что методы оценки и наборы данных не совпадают. Несмотря на то, что существуют наборы данных, которые должны быть стандартом, в разных статьях используются разные подмножества тестовых наборов данных. Проблема может быть решена путем обсуждения внутри сообщества и обеспечения соблюдения стандартов. Двумя платформами для проведения таких стандартизированных тестов могут быть Kaggle и BIAFLOWS [66].

Заключительный вывод статьи состоит в том, что область может попытаться продвинуться к разработке генерализованных моделей, способных сегментировать ядра в изображениях различных типов микроскопии. Некоторые модели уже способны на это, хотя и с ограниченным количеством поддерживаемых типов микроскопии, например, модели, полученные в ходе соревнования DSB 2018 [11] [67].

3.2 AnnotatorJ: плагин ImageJ для упрощения ручного аннотирования клеток и их ядер

В этом разделе кратко пересказывается [12].

Для обучения моделей для сегментирования одиночных клеток (ядер) на основе глубокого обучения необходимы аннотированные данные. Для обучения более робастных моделей необходимы большие наборы данных, но ручное аннотирование является дорогостоящим процессом, так как требует значительного количества времени и усилий от экспертов в области биологии. Чтобы сделать процесс аннотирования более быстрым и точным, был разработан плагин AnnotatorJ [12] для ImageJ/FIJI [13] (программное обеспечение для анализа биологических изображений), который объединяет идентификацию одиночных клеток с глубоким обучением и ручным аннотированием.

Главной особенностью AnnotatorJ является помощник для построения контуров. Ассистент контуров использует предварительно обученную модель U-Net для прогнозирования области, занимаемой интересующим объектом. После этого пользователь может уточнить контуры объекта, если это необходимо.

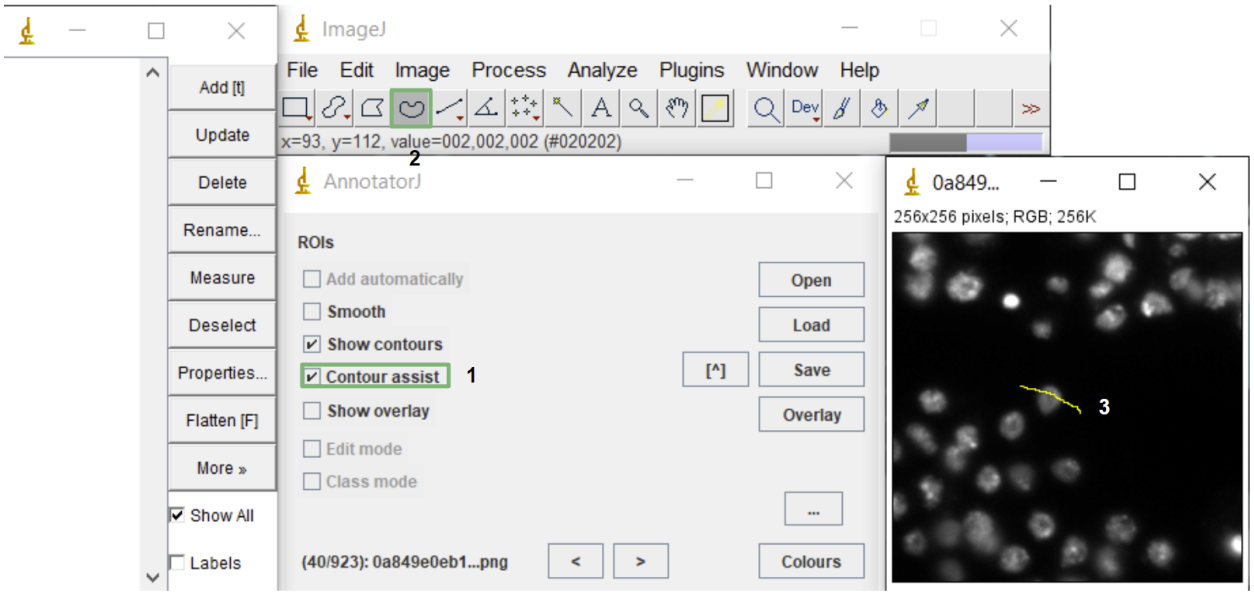


Рис. 8: Первый шаг аннотирования с помощью помощника для построения контуров: инициализация контура путем рисования линии на объекте. Цифрами и зелеными рамками показаны шаги, которые необходимо выполнить в интерфейсе. Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11].

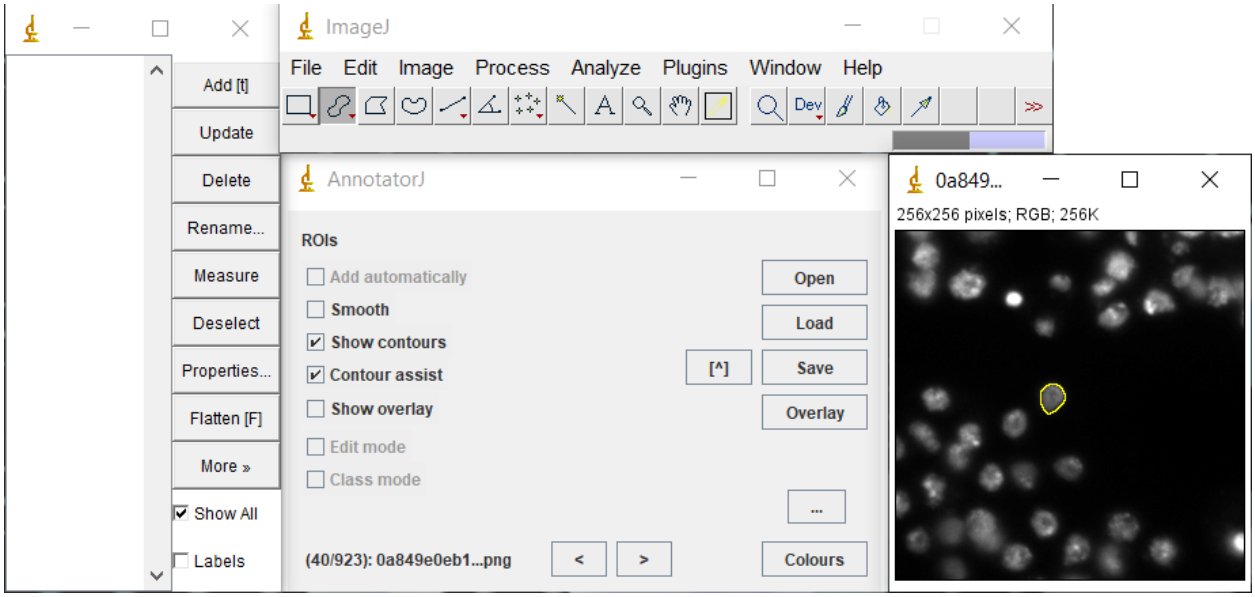


Рис. 9: Инициализированный контур с помощью предварительно обученной модели сегментирования глубокого обучения (справа). Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11].

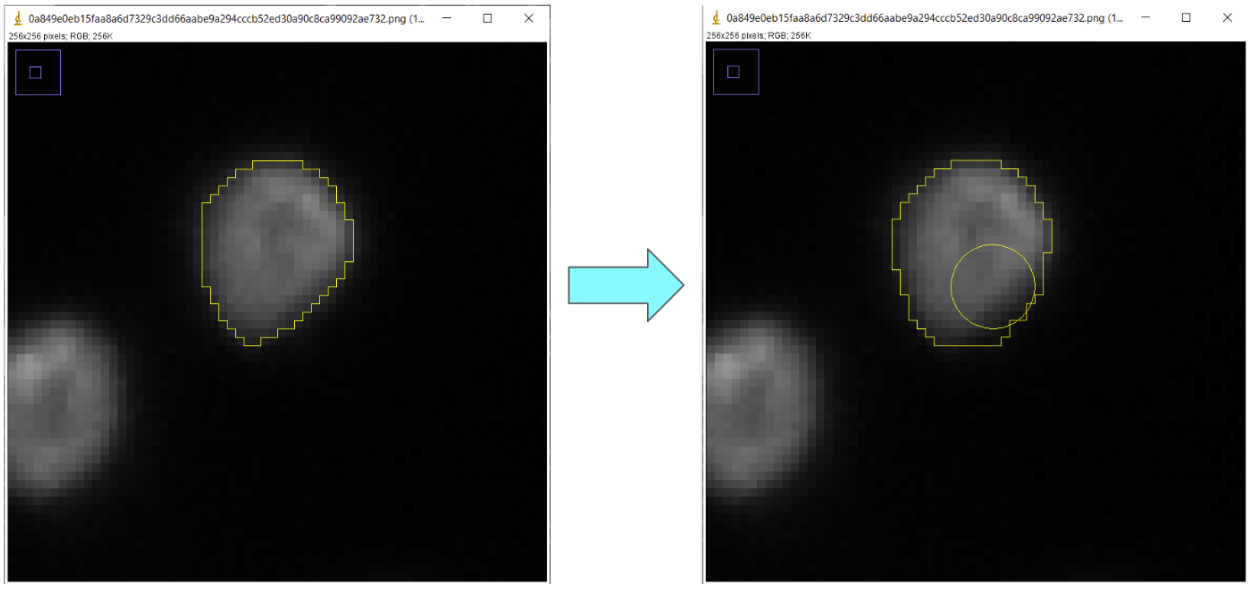


Рис. 10: Ручное уточнение контура объекта. Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11].

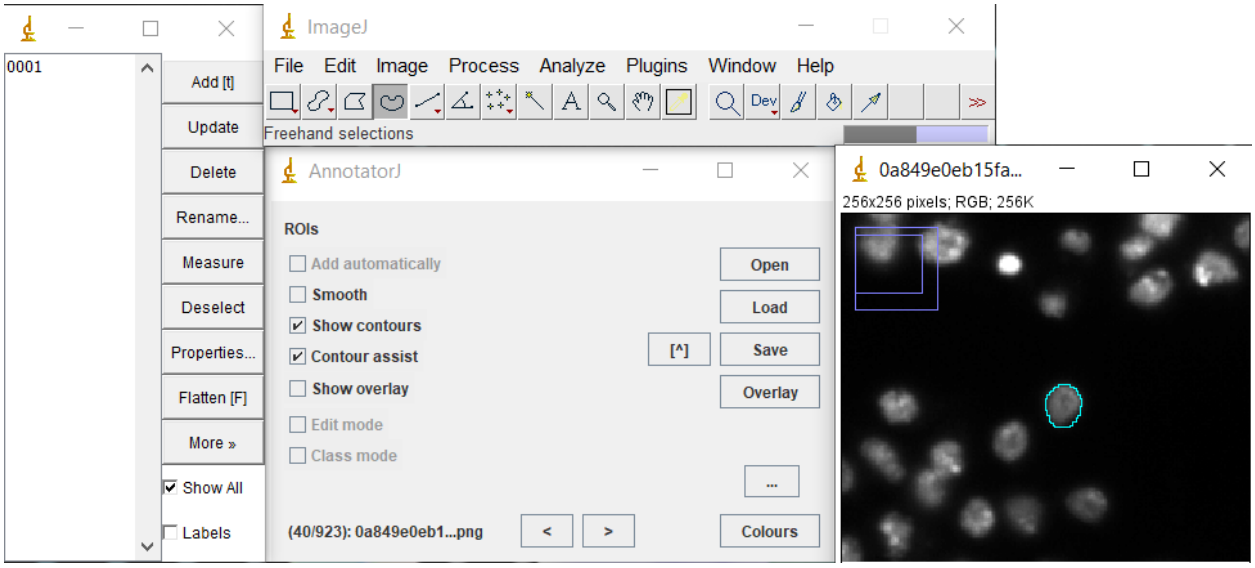


Рис. 11: После уточнения границ, объект добавляется нажатием клавиши ‘Q‘. Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11].

Чтобы сделать обученные модели совместимыми с ImageJ/Fiji, который разработан на Java, мы использовали программные библиотеки DL4J и ND4J (`http://deeplearning4j.org/`). AnnotatorJ в свободном доступе по ссылке `https://github.com/spreka/annotatorj`.

3.3 Аугментация во время тестирования для сегментирования клеточных ядер методами глубинного обучения в изображениях микроскопии

В этом разделе кратко пересказывается [20].

Сегментирование клеточных ядер на основе глубинного обучения в значительной степени опирается на данные, аннотированные вручную, которые в большинстве случаев аннотируются экспертами в данной области. Чтобы увеличить количество обучающих данных и обучить более робастные модели, аугментация данных [38] (см. 2.1) стала общепринятой техникой в глубинном обучении. Аугментация данных часто используется в случае неоднородных или небольших наборов данных, что часто случается в области сегментирования клеток и их ядер.

В то время как обычный подход к аугментации данных выполняется во время обучения, идея другого подхода, аугментации во время тестирования (test-time augmentation - TTA) (рисунок 12), заключается в том, чтобы выполнить предсказания на исходной и аугментированной версиях данных и затем объединить предсказания. Эта техника существует уже некоторое время и была успешно использована в различных задачах анализа изображений [68] [69] [70]. Эксперименты с аугментацией во время тестирования проводились в условиях задачи сегментирования клеточных ядер.

3.3.1 Аугментация во время тестирования

Аугментация во время тестирования выполняется в четыре шага:

1. Аугментация оригинального изображения.
2. Предсказание на оригинальном и аугментированных версиях изображения.
3. Дизагументация: если оригинальное изображение было повернуто или перевернуто, то ориентация предсказания возвращается к оригинальной для дальнейшего объединения предсказаний.
4. Объединение предсказаний: этот шаг различается для Mask R-CNN и U-Net и описан далее.

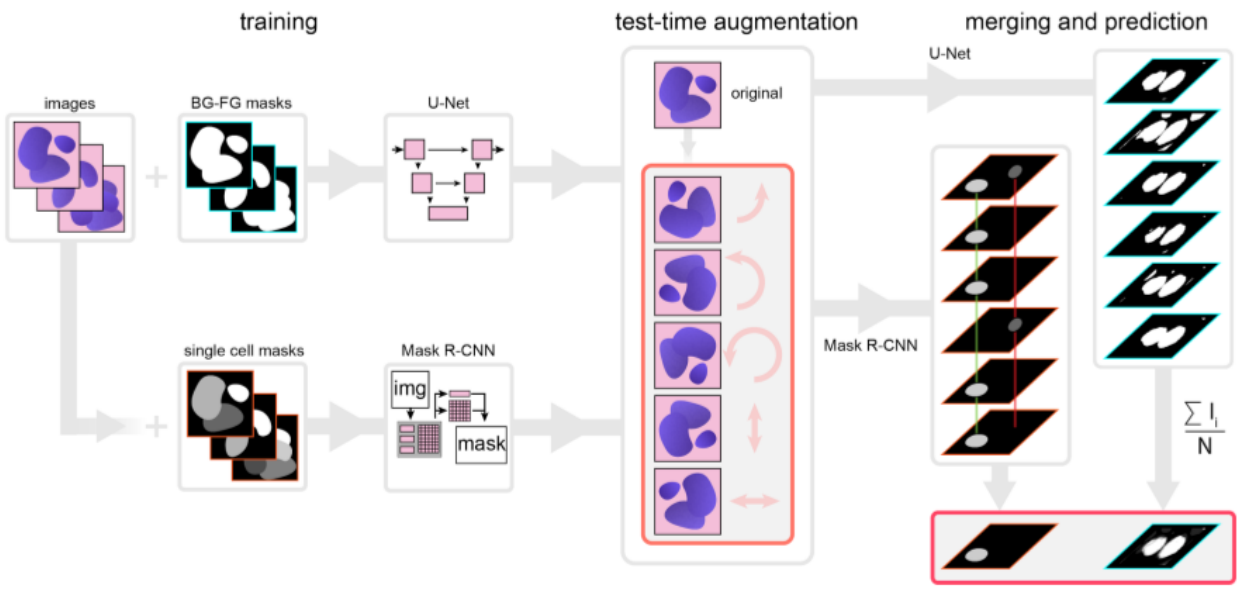


Рис. 12: Принцип работы аугментации во время тестирования. Входные данные: предсказание на нескольких аугментированных экземплярах одних и тех же тестовых изображений с обученными моделями. Для объединения предсказаний использовалось голосование на уровне пикселей (U-Net), или комбинация совмещения объектов и метода голосования (Mask R-CNN). Источник рисунка [20].

Для U-Net шаг (4) прямолинеен, просто суммируем и усредняем дизаугментированные карты вероятностей. Полученная карта вероятностей преобразуется в бинарную маску с помощью порога (0.5), которая в дальнейшем используется для оценки сегментирования (рис. 12, справа).

В случае Mask R-CNN требуется больше пост-процессинга, так как эта модель сегментирует отдельные объекты. Здесь каждый объект обрабатывается отдельно: для каждого обнаруженного объекта проводится голосование по принципу большинства. Перед голосованием необходимо выполнить выравнивание объектов: объекты из предсказаний исходной и аугментированной версий входного изображения проверяются, можно ли считать их одним и тем же объектом. В этом случае два объекта (каждый из которых относится к разным версиям входного изображения) считаются одним и тем же объектом, если индекс Жаккара (Ур. 1) между ними составляет не менее 0.5. Если один и тот же обнаруженный объект присутствует в большинстве сегментаций, то он будет включен в окончательную маску. Маска включенного объекта корректируется большинством голосов на уровне пикселей.

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

3.3.2 Материалы и методы

Для экспериментов были выбраны популярные архитектуры глубинного обучения для сегментирования: U-Net [33] (для семантического сегментирования) и Mask R-CNN[34] (для сегментирования объектов), а большая часть использованных данных была использована из

набора данных соревнования Data Science Bowl 2018 [11], а также дополнительные источники [71] [72] [73] [74] [75] [76] [77]. Изображения были обрезаны до размера 512×512 пикселей. Изображения меньшего размера, чем 512×512 были увеличены. Этот первичный набор данных был разделен на два набора данных: один с флуоресцентными изображениями (далее Fluorescent или Fluo) и изображения тканей (далее Tissue). Для этих двух наборов данных были сделаны следующие разбиения на тренировочные и тестовые данные:

- 95% изображений в тренировочном и 5% в тестовом наборе (далее Fluo_5 или Tissue_5)
- 85% изображений в тренировочном и 15% в тестовом наборе - 6 наборов для кросс-валидации (первое разбиение кросс-валидации далее Fluo_15 или Tissue_15)
- 70% изображений в тренировочном и 30% в тестовом наборе (далее Fluo_30 или Tissue_30)

Для каждого набора данных и их разбиений были обучены отдельные модели. Для обучения использовалась аугментация с использованием горизонтального и вертикального переворота, $90°$, $180°$ и $270°$ повороты. Аугментации были произведены до тренировки, следовательно, размер тренировочной выборки для каждого разбиения был равен $6*$*количество уникальных изображений в тренировочном наборе данных*.

В экспериментах с U-Net (архитектура описана в 2.2.1) использовалась широко распространенная реализация [78] на базе Tensorflow [79] и Keras. Модели обучались в течение 200 эпох с постоянным коэффициентом скорости обучения 3×10^{-4}. Начальные параметры инициализировались случайным образом. Как функция потерь использовалась бинарная кросс-энтропия с оптимизатором ADAM [80]. Размер пакета во время обучения равен 1 (из-за ограничений памяти GPU это был обычный размер пакета для сегментирования). Модели U-Net были обучены в двух вариантах: с использованием и без использования аугментаций в обучении. Для экспериментов с Mask R-CNN использовалась кодовая база Matterport [81], также основанная на Tensorflow и Keras. Скрипты для оценки сегментирования использовались из [37].

Модели обучались в течении трех эпох для различных групп слоев в следующем порядке:

- Инициализировать модель претренированным весами на наборе данных COCO (`https://github.com/matterport/Mask_RCNN/releases/download/v1.0/mask_rcnn_coco.h5`)
- Эпоха 1: все слои сети были обучены с коэффициентом скорости обучения 10^{-3}.
- Эпоха 2: обучение ResNet stage 5 (ResNet состоит из 5 стадий, каждая из которых имеет сверточные блоки и блоки идентификации, включая 3 сверточных слоя на блок) и головных слоев с коэффициентом скорости обучения 5×10^{-4}.
- Эпоха 3: Обучение только головных слоев с коэффициентом скорости обучения 10^{-4}.

Параметры модели были инициализированы весами (`https://github.com/matterport/Mask_RCNN/releases/download/v1.0/mask_rcnn_coco.h5`) претренированными на наборе данных COCO. В качестве функции потерь использовалась бинарная кросс-энтропия с оптимизатором ADAM [80], размер пакета 1. Эта стратегия обучения повторяет стратегию из [37]. Модели Mask R-CNN обучались только с использованием аугментаций в обучении.

mAP_{DSB} для изображения вычисляется следующим образом: посчитать среднюю точность (average precision) для всех тестовых изображений с порогом индекса Жаккара t (индекс Жаккара вычисляется между истинной маской объекта и предсказанной маской объекта) и взять среднее по всем порогам индекса Жаккара T (2). В этом уравнении, $TP(t)$, $FP(t)$ и $FN(t)$ обозначают количество истинно положительных, ложно положительных и ложно отрицательных объектов, соответственно:

$$\begin{aligned} mAP_{DSB} &= \frac{1}{|T|}\sum_{t\in T}\frac{TP(t)}{TP(t)+FP(t)+FN(t)}, \\ T &= \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95\} \end{aligned} \tag{2}$$

Предсказания U-Net были оценены с помощью индекса Жаккара (Ур. 1). Эффективность аугментации во время тестирования было оценено разницей индекса Жаккара с ней ($merged$) и без нее ($original$):

$$delta = merged - original \tag{3}$$

3.3.3 Результаты

Аугментация во время тестирования в среднем улучшила результаты сегментирования для всех разбиений данных, если использовалась вместе с моделями Mask R-CNN. Средний прирост в метрике mAP_{DSB} составляет от 0.01 до 0.02. Хотя на большинстве тестовых изображений mAP_{DSB} улучшилось, есть несколько изображений где качество сегментирования несколько ухудшилось (рис. 13).

Аугментация во время тестирования, используемая вместе с моделями U-Net, улучшила результаты сегментирования по метрике индекс Жаккара. Мы можем наблюдать, что для большинства контрольных точек модели в каждом сценарии обучения, за исключением начала обучения, когда модель недостаточно оптимизирована (Рисунок 14).

В некоторых тестовых примерах, аугментация во время тестирования значительно изменило качество сегментирования. Такие примеры для U-Net и Mask R-CNN показаны на рисунке 15.

Мы применили аугментацию во время тестирования в дополнение к методу [37] (лучший метод для тестового набора соревнования DSB 2018 согласно данным Kaggle на момент публикации статьи) и смогли увеличить производительность на 0.011 в метрике mAP_{DSB}.

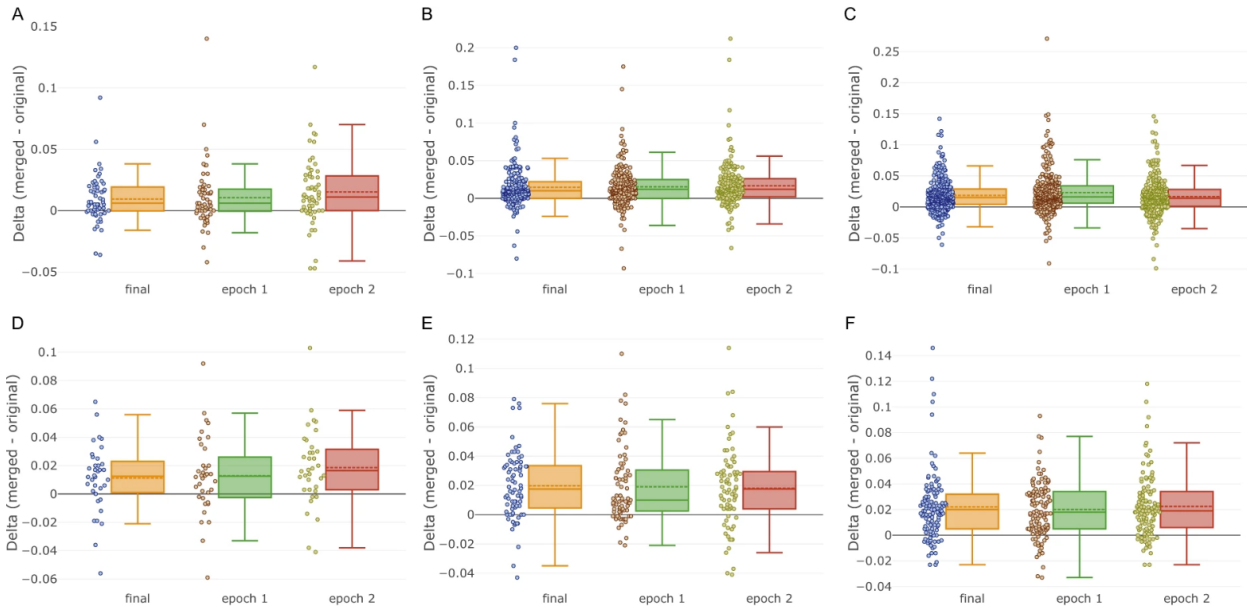


Рис. 13: Результаты аугментации во время тестирования для Mask R-CNN (delta) для разбиений данных на тренировочные и тестовые. Каждая точка - изображение. Столбцы - эпоха оубчения. Штриховая линия - среднее, обычная линия - медиана. A. Fluorescent_5. B. Разбиение Fluorescent_15 (кросс-валидация 1) C. Fluorescent_30. D. Tissue_5. E. Tissue_15 (кросс-валидация 1) F. Разбиение Tissue_30. Источник рисунка [20].

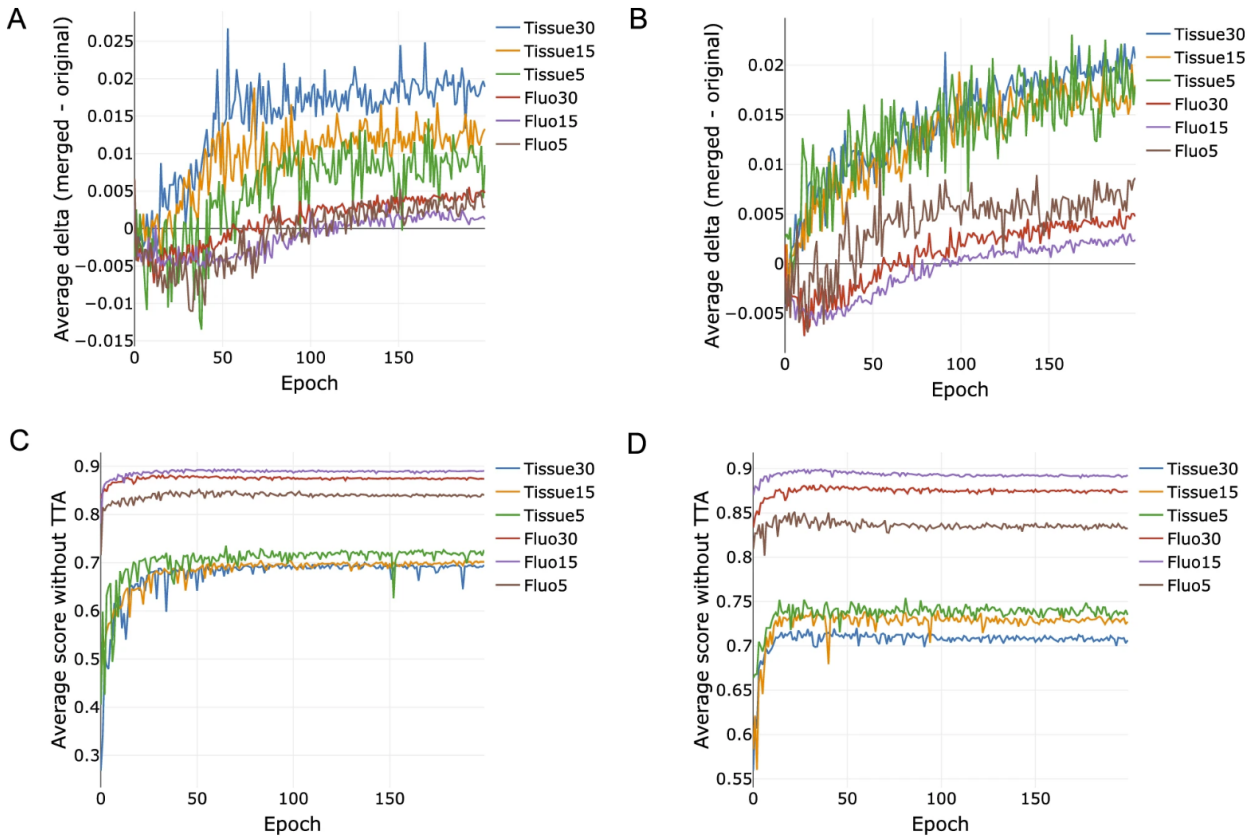


Рис. 14: Средний индекс Жаккара на тестовых данных и влияние аугментации во время тестирования для U-Net ($delta$). A. Средняя $delta$, аугментации во время обучения не использовались. B. Средняя $delta$, аугментации во время обучения использовались. C. Средний индекс Жаккара, аугментации во время обучения не использовались. D. Средний индекс Жаккара, аугментации во время обучения не использовались. Источник рисунка [20].

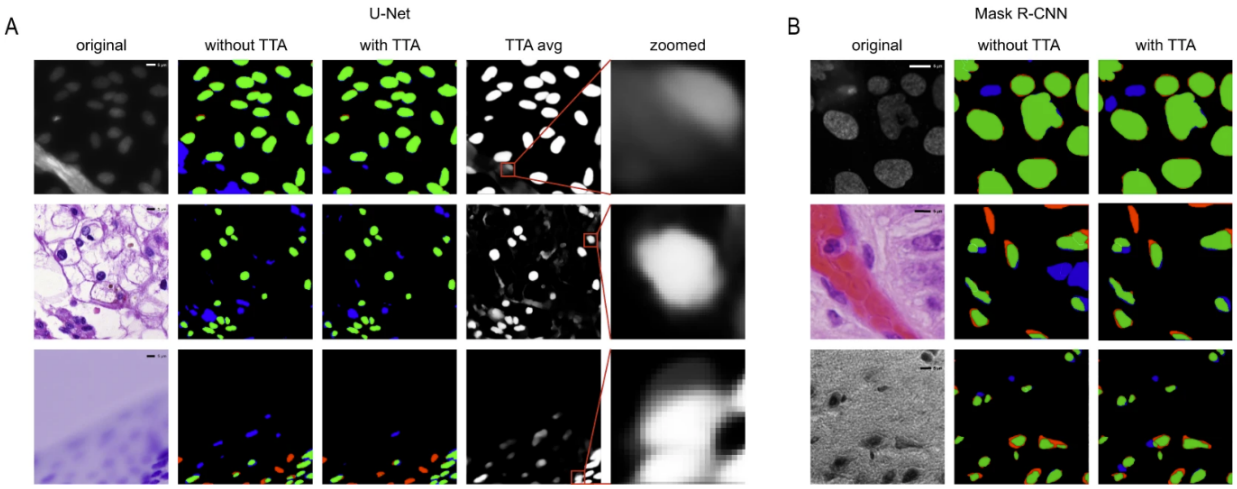


Рис. 15: Примеры сегментирования. A. Сегментирования U-Net. Первый столбец - оригинальное изображение, второй столбец - предсказания без аугментации во время тестирования в сравнении с аннотированной маской, третий столбец - предсказания с аугментацией во время тестирования в сравнении с аннотированной маской. Красным цветов показаны ложно отрицательные сегментации, зеленым показаны истинно положительные сегментации и синим ложно положительные сегментации. Четвертый столбец - предсказания с аугментацией во время тестирования до фильтрации по пороговому значению, пятый столбец - увеличенные фрагменты из предыдущего столбца. Строки - примеры изображений. B. Сегментирования Mask R-CNN. Столбцы аналогичны первым трем столбцам из A, строки - примеры изображений. Источник рисунка [20].

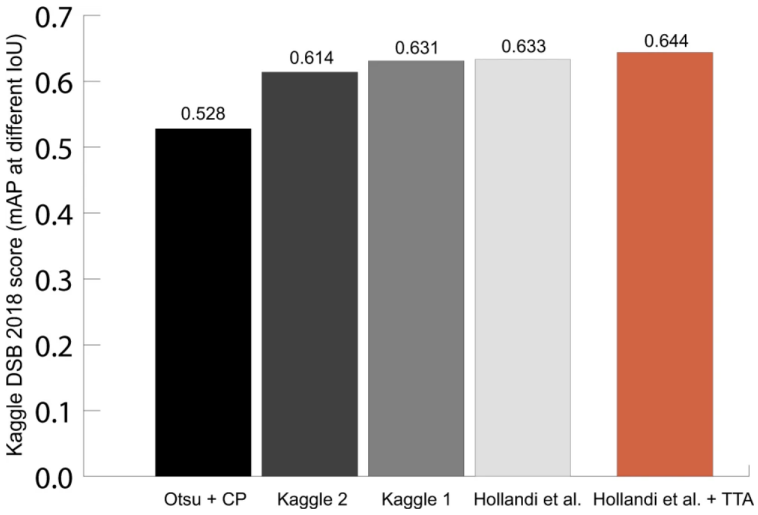


Рис. 16: Результаты на тестовой выборке соревнования DSB 2018 стадии 2 аугментации во время тестирования, комбинированного с методом [37]. Источник рисунка [20].

3.4 Обучение представлениям для профилирования пертурбаций на основе изображений[1]

В этом разделе кратко пересказывается [82].

Фенотипический подход к поиску лекарств основывается на наблюдениях за действием лекарств на объектах, в данном конкретном случае мы рассматриваем отдельные клетки. Эта проблема требует не только значительных лабораторных усилий, но и вычислительных подходов для обработки полученных данных. Одной из первых попыток измерить эффект лечения с помощью характеристик, извлеченных из данных флуоресцентной микроскопии, была [9]. Позже был выпущен CellProfiler [19], стандартный подход для извлечения представлений одиночных клеток. Он генерирует человекочитаемые признаки и их полезность была доказана в различных последующих задачах [16].

Теперь возникает вопрос, что если мы сможем извлечь еще более биологически релевантные представления клеток из изображений с помощью глубинного обучения? Вдохновленные обучением представлениям и популярными архитектурами глубокого обучения для классификации изображений, исследователи начали искать методологию, которая позволила бы им извлекать такие биологически релевантные представления.

Одна из первых попыток использования переносного обучения (использование предварительно обученных сетей классификации изображений с набором данных ImageNet [15]) для морфологического профилирования была предпринята в [51] на полных изображениях, то есть размер полного изображения был изменен до размера входа сети и запущен в режиме предсказания.

Обучение моделей непосредственно на изображениях отдельных клеток было изучено в пробных экспериментах [50]. Они были основаны на слабо-контроллируемом обучении (weakly-supervised learning, далее WSL), которое не требует ручного аннотирования данных для обучения представлениям признаков. Вместо этого используются известные пертурбации клетки (например, лекарства) в качестве косвенных признаков для интересующих фенотипов. Эти разметка являются слабой, поскольку нет уверенности в том, что все пертурбации имеют фенотип, достаточно сильно отличающийся от клеток, на которых не было воздействия (отрицательных контролей), или результирующие фенотипы не похожи для разных методов лечения.

Здесь проводится систематическая оценка трех больших публичных наборов данных Cell Painting. Эти наборы данных содержат тысячи лекарств, сотни планшетов и миллионы отдельных клеток. Тестируемые представления извлекаются предварительно обученными моделями и моделями, обученными с WSL, и сравниваются с классическими признаками. Для проведения экспериментов по обучению и извлечению представлений (наборов признаков) был разработан общедоступный инструмент *DeepProfiler*.

В настоящее время найдены лучшие практики, позволяющие улучшить результаты в по-

[1]Статья размещена в качестве препринта и будет подана в журнал

следующей биологической задаче методами глубокого обучения. Для интерпретации полученных результатов с помощью обученных моделей и рассуждений о задачах используется фреймворк причинно-следственного моделирования [83] [84].

3.4.1 Наборы данных Cell Painting

В данном исследовании использовалось пять наборов данных:

- BBBC037 (также известный как TA-ORF) набор данных [85], опубликованный в 2017 году для тестирования морфологического профилирования с использованием сверхэкспрессии в клетках человека в качестве общего подхода для аннотирования функции генов и аллелей.
- BBBC022 набор данных [47], опубликованный в 2013 году, скрининг 1600 биологически активных соединений.
- Набор данных BBBC036 (также известный как CDRP-Bioactives) [86], опубликован в 2017 году, скрининг примерно 2000 биологически активных соединений.
- BBBC043 dataset (также известный как LUAD) [87], тестирование вариантов аденокарциномы легких (всего 375 вариантов).
- Набор данных LINCS [88], скрининг 1300 соединений.

Изображения во всех вышеперечисленных наборах данных были получены при 20-кратном увеличении и пятиканальном режиме (все снимки сделаны с помощью протокола Cell Painting). Первые три набора данных в списке выше используются в качестве бенчмарков, последние два используются только для построения комбинированного набора данных Cell Painting (обсуждается в раздел 3.4.5).

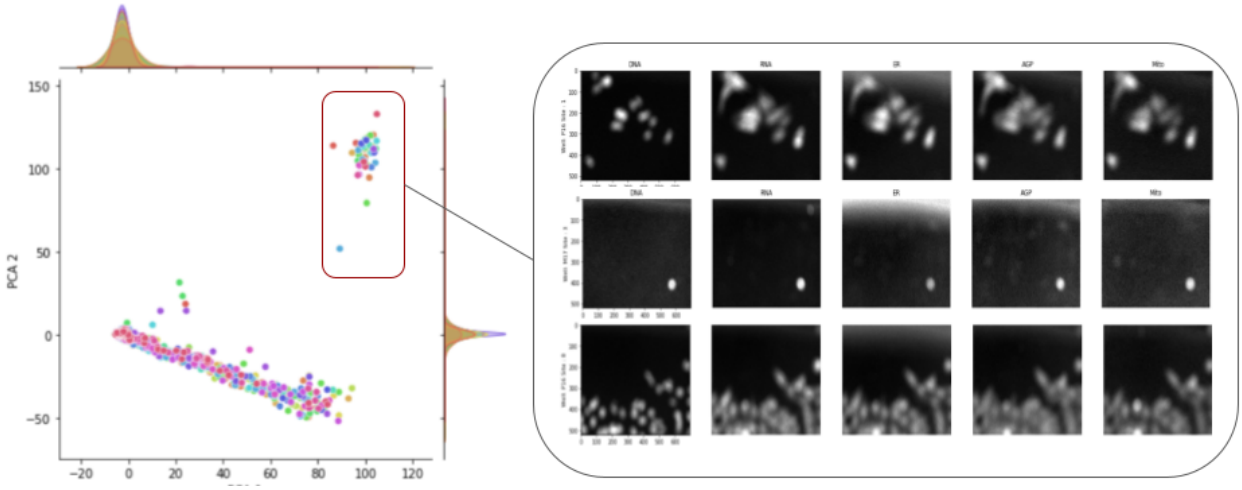


Рис. 17: Пример контроля качества для набора данных BBBC022. Слева: График PCA для первых двух принципиальных компонент. Каждая точка - лунка, цвета обозначают планшеты. Наблюдается кластер с выбросами. Справа: примеры изображений из лунок-выбросов. Видно, что эти изображения не в фокусе.

Для вышеуказанных наборов данных был проведен контроль качества, чтобы удалить очень шумные или расфокусированные изображения, так как они не могут содержать достоверную фенотипическую информацию и в конечном итоге могут исказить агрегированные представления. Для этого представления были извлечены с помощью DeepProfiler (модель EfficientNet-B0, предварительно обученная на наборе данных ImageNet, (см. 3.4.3), затем эти признаки были объединены по изображениям, по лункам (как описано в [48]) и было применено преобразование сферическое преобразование (см. 3.4.4.2). Затем к этим агрегированным профилям был применен анализ принципиальных компонент (PCA). Выбросы, наблюдаемые на графиках PCA, проверялись вручную на наличие технических проблем (Рисунок 17).

Наборы данных имеют метки классов: пертурбации генов в BBBC037 и соединения с концентрацией для BBBC022 и BBBC036, далее общее название пертурбации. В случае наборов данных BBBC036 и BBBC022, соединения, которые присутствовали более одного раза, были отфильтрованы, оставив только записи с максимальными концентрациями. Для последующего анализа важны аннотации суперклассов: сигнальные пути генов (набор данных BBBC037) или механизмы действия (для наборов данных BBBC036 и BBBC022). Аннотации суперклассов были взяты из [85] и затем доработаны. Координаты клеток были получены с помощью CellProfiler.

3.4.2 DeepProfiler

Было разработано ПО под названием DeepProfiler, которое помогает обучать модели с WSL и извлекать представления одиночных клеток из высокопроизводительных экспериментов по визуализации. DeepProfiler представляет стандартизированный рабочий процесс для использования сверточных нейронных сетей для извлечения представлений одиночных клеток из объемных коллекций изображений.

С помощью DeepProfiler можно либо обучить модель и затем выполнить извлечение признаков, либо использовать предварительно обученную модель для извлечения признаков. Входными данными DeepProfiler являются изображения, соответствующие метаданные и конфигурация эксперимента. DeepProfiler извлекает изображения одиночных клеток заданного размера (вырезанные изображения отдельных клеток, DeepProfiler не осуществляет сегментирование, но может вырезать объекты, если предоставлена маска сегментирования) из полноразмерных изображений, которые являются входными данными для сетей глубокого обучения. Рабочий процесс показан на рисунке 18. Извлеченные признаки могут быть использованы в последующем анализе, который обычно уникален для набора данных и зависит от биологических вопросов. Помимо обучения и извлечения признаков, DeepProfiler имеет дополнительные функции для сжатия изображений и извлечения вырезанных и отдельных клеток из полноразмерных изображений в отдельные наборы изображений.

Фреймворк реализован в Tensorflow [79] (для версий 1 и 2). Исходный код, документация и обсуждения доступны на странице GitHub (`https://github.com/cytomining/DeepProfiler/`).

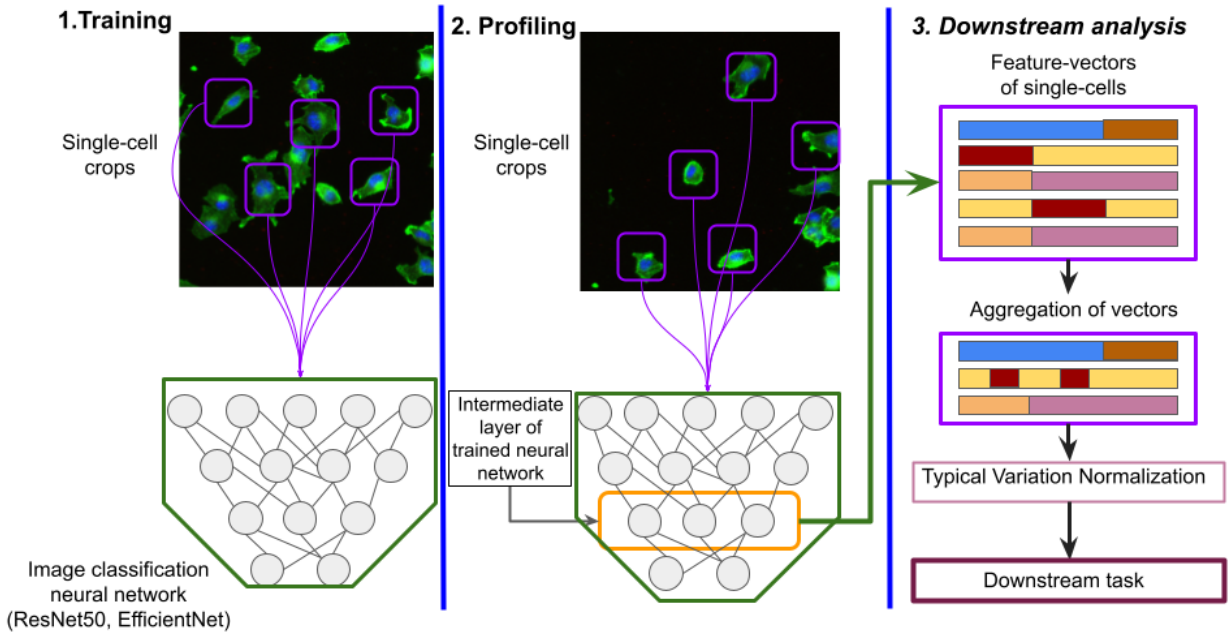


Рис. 18: Типичное использование DeepProfiler. 1. Обучение сети классификации изображений 2. Обученная модель используется для извлечения представлений 3. Представления используются для последующих задач по анализу данных. Шаги 1 и 2 на изображении выполняются с помощью DeepProfiler, шаг 3 не входит в DeepProfiler и является предпочтением пользователя. Использованы изображения микроскопии из набора данных BBBC021 [72].

3.4.3 Экспериментальная установка

3.4.3.1 EfficientNet

Для экспериментов по глубокому обучению использовалась архитектура EfficientNet [89], в частности, ее базовая версия EfficientNet-B0. Выбор был обусловлен ее вычислительной эффективностью и продемонстрированной точностью на наборе данных ImageNet [15], превосходящей ResNet50 [41]. EfficientNet использовалась в нескольких предыдущих публикациях, связанных с обработкой изображений клеток, для извлечения признаков и профилирования [87], а также для обучения модели на объединенном наборе данных изображений клеток [90]. Некоторые решения задачи Recursion Pharmaceuticals по классификации изображений клеток `https://www.kaggle.com/competitions/recursion-cellular-image-classification/` были основаны на различных модификациях EfficientNet.

3.4.3.2 Эксперименты с предварительно обученными моделями

В этом подходе, предварительно обученные на наборе данных ImageNet [15]. Поскольку предварительно обученные сети требуют 3-канального входа, каждый из каналов реплицируется три раза и передается в модель отдельно. В качестве входных данных использовались вырезанные изображения отдельных клеток размером 128×128. Предварительная обработка для используемой модели также потребовала изменения размера до 224×224 и min-max

нормализация для получения входных данных в диапазоне $[-1, 1]$. Признаки были извлечены из слоя $block6a\_activation$. Для каждого канала длина вектора признаков составляет 672 признака, соответственно, полный вектор признаков для клетки составляет 3360 признаков.

3.4.3.3 Эксперименты с WSL

Обучение моделей и последующее извлечение признаков проводилось с помощью DeepProfiler. В качестве входных данных использовались предварительно вырезанные изображения одиночных клеток, сохраненные в виде последовательности из пяти каналов и преобразованные в процессе обучения, так что на вход модели поступает $128 \times 128 \times 5$. Во время обучения использовались дополнения:

- Случайное обрезание изображения с вероятностью 50%, размер каждой стороны обрезанного изображения составляет не менее 80% от исходного размера, после этого оно масштабируется до размера исходного изображения.
- Случайный переворот по горизонтали и затем случайный поворот (на 90 градусов).
- Изменение цвета: яркость (до 10% отклонения от оригинала), затем контраст (до 20% отклонения от оригинала). Каждый канал обрабатывается отдельно на обоих этапах.

Поскольку количество отдельных клеток варьируется между классами, автобалансировка производится в каждой эпохе обучения. Для всех наборов данных были выбраны следующие параметры: функция потерь categorical cross-entropy, размер пакета 32, постоянная скорость обучения 0.005 с оптимизатором SGD, включенные дополнения, отсутствие сглаживания меток и 30 эпох. Модели инициализируются предварительно обученными весами на данных ImageNet.

Использовались две схемы разбиения данных на тренировочные и валидационные:

- Leave-plates-out - одиночные клетки из одного подмножества планшетов используются для обучения, а из другого - для валидации.
- Leave-cells-out - отдельные клетки из каждого планшета и каждой лунки используются как в тренировке, так и в валидации, примерно 60% клеток из каждой лунки используются в тренировке, 40% в валидации.

Используя обученные модели, признаки были извлечены из слоя $block6a\_activation$ (размер вектора признаков - 672).

3.4.3.4 Вычислительная эффективность

Вычислительная эффективность оценивалась с точки зрения времени вычислений и дискового пространства, необходимого в сравнении с классическим подходом. Предлагаемый подход быстрее, чем классический, так как использует распараллеливание на GPU. Для всех экспериментов по глубокому обучению использовалась NVIDIA V100. Время обучения в среднем по

наборам данных занимает 3.3 часа, профилирование занимает примерно 0.58 часа с предварительно обученной моделью и 0.22 часа на планшет с обученной моделью. Предварительно обученная модель требует больше времени, так как для одного изображения требуется пять циклов предсказания. Сравнение доступно на рисунке 19. Цена вычислений в денежном эквиваленте здесь не сравнивается, однако обычно облачные вычисления на GPU обходятся дороже, чем на CPU.

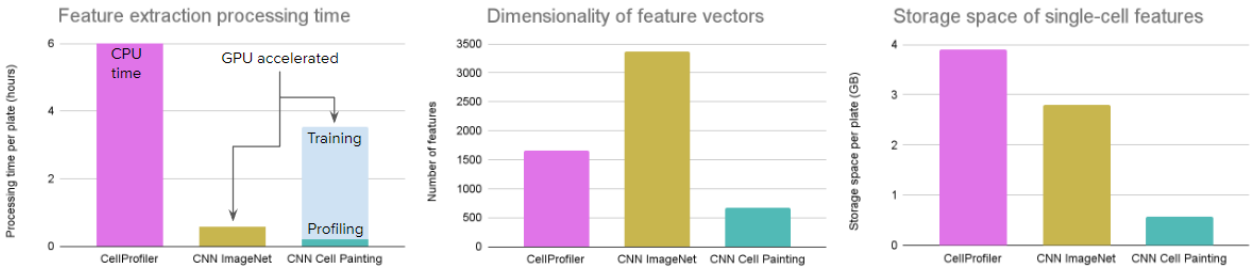


Рис. 19: Вычислительная затраты на стратегии профилирования. Источник рисунка [82].

3.4.4 Рабочий процесс и оценка профилирования

3.4.4.1 Агрегация признаков и поиск сходства

Агрегирование признаков – это процесс получения профилей уровня пертурбаций из профилей одиночных клеток [48]. Существуют промежуточные уровни, такие как уровень изображения и уровень лунки. Векторы признаков отдельных клеток агрегируются с помощью медианы на уровень изображения, а затем профили уровня изображения агрегируются с помощью среднего на уровень лунок. В данной работе шаги агрегирования признаков одинаковы, независимо от источника признаков - CellProfiler или модели глубокого обучения.

Для оценки сходства между процедурами могут использоваться различные метрики [48], здесь используется косинусное сходство (также используется в других работах [91]).

3.4.4.2 Корректировка с использованием сферического преобразования

Одной из попыток сокращения нежелательных технических вариаций является нормализация типичных вариаций (TVN), предложенная в [92], также использованная в [93]. Она рассчитывает оси вариаций с помощью анализа принципиальных компонент на профилях отрицательных контрольных лунок. Полученные оси нормализуются, в результате чего оси с большой вариацией сокращаются, а оси с малой вариацией увеличиваются. Затем преобразование применяется ко всем профилям уровня лунок.

Здесь преобразование ZCA-сферизации используется аналогично TVN. В качестве входных данных используется матрица профилей отрицательных контролей $X^{n \times d}$, где n - количество контрольных лунок, а d - размерность вектора признаков. Ковариационная матрица для X - $\Sigma = \frac{X^T X}{n}$, ее собственное разложение - $Q = U \Delta U^T$, где Δ - собственные значения. Для получения окончательной ZCA-трансформации [94] (сферическое преобразование), является следующее $U(\Delta + \lambda)^{-1} U^T$, где λ - параметр регуляризации.

3.4.4.3 Оценка и метрики

Как описано в [16], биологической задачей является проверка того, принадлежат ли наиболее похожие пертурбации примененные к клеткам к одному и тому же генетическому сигнальному пути или механизму действия (MoA), в соответствии с использованной метрикой сходства. Для оценки использовалось несколько метрик, все они кратко описаны ниже. В дальнейшем тексте *пертурбации-запросы* называются пертурбации, которые имеют по крайней мере две пертурбации в одном и том же механизме действия или пути и используются в качестве запросов в задаче ранжирования.

Первая метрика, которая была использована, это коэффициент обогащения (folds of enrichement). Отношение шансов рассчитывается аналогично [16], с той лишь разницей, что здесь это делается только для 1% порога и для каждого запроса отдельно. Затем вычисляется среднее полученных значений для всех запросов, что и является результирующей метрикой.

В качестве другой метрики была использована интерполированная кривая Precision-Recall (точность-полнота) и среднее значение средней точности для задачи ранжирования. Эта метрика рассчитывается следующим образом: каждая пертурбация является запросом, и проверяются пертурбации наиболее похожие на запрос. $Precision@K$ в этой задаче ранжирования - это отношение пертурбаций, которые принадлежат к тому же MoA/пути, что и запрос, к K наиболее похожим пертурбациям. Та же интуиция применима для $Precision@Recall$: для одной пертурбации-запроса перебираются все пертурбации-ответы, ранжированные по метрика похожести, пока не достигнем полноты равной 1 (найдем все положительные совпадения). Поскольку каждый MoA/путь имеет различное количество связанных с ним истинных пертурбаций, $Precision@Recall$ интерполируется для покрытия максимального количества точек полноты, интерполированная точность определяется как $p_{interpolated}(r) = max_{r' \geq r} p(r')$ [95]. Средняя точность (average precision, площадь под интерполированной кривой Precision-Recall) - это среднее значение $p_{interpolated}$ во всех точках полноты. Здесь mAP - простое среднее значение средней точности для всех запросов.

$Hits\ in\ the\ top\ 1\%$ метрика моделирует задачу поиска 'попадания' в наиболее перспективных пертурбциях-кандидатах. Метрика применима на нескольких уровнях профилирования:

- Уровень пертурбации: Измеряет количество пертурбаций-запросов, которые имеют пертурбации-ответ с тем же MoA/путем среди 1% наиболее похожих.

- Уровень лунки: Профиль лунки используется в качестве запроса (могут быть использованы все пертурбации). Подсчитывается количество пертурбаций, в которых лунки-запросы и лунки-ответы одной пертурбации входят в 1% наиболее похожих.

- Уровень изображения: Профиль изображения используется в качестве запроса (могут быть использованы все пертурбации). Подсчитывается пертурбаций, у которых изображения-запрос и изображения-ответ одной и той же пертурбации входят в топ-1% наиболее похожих изображений. Изображения из той же лунки, что и изображение-запрос, исключаются из возможных ответов.

3.4.5 Выбор сильных пертурбаций и комбинированный набор данных Cell Painting

Для расширения потенциального пространства признаков с учетом биологических и технических вариаций были собраны пертурбации, приводящие к сильным фенотипам, из пяти наборов данных Cell Painting. Сильная пертурбатция – пертурбация приводящяя к фенотипу, который отличается от фенотипа необработанных клеток. Для оценки силы фенотипа используется пространство признаков CellProfiler (со сферической коррекцией и параметром регуляризации $1e-2$) и измеряется Евклидово расстояние между профилями лунок пертурбаций и профилями лунок отрицательных контролей (алгоритм 1).

Algorithm 1 Выбор сильных пертурбаций

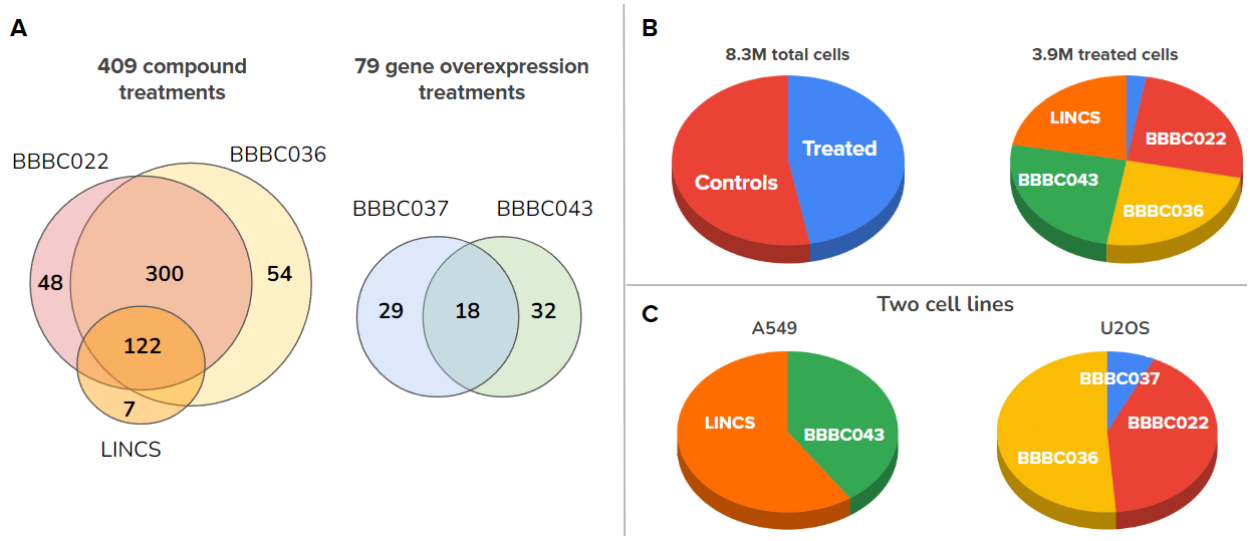
for each p in $Plates$ **do**
    Расчет медианного профиля отрицательных контролей в планшете - MCP_p
    Вычислить Евклидово расстояние между профилями для лунок пертурбаций и MCP_p, получем расстояние EDT_p
    Вычислить Евклидово расстояние между профилями для лунок отрицательных контролей и MCP_p, получем расстояния ECT_p
    Вычислить μ и σ от ECT_p
    Использовать μ и σ для Z-score EDT_p.
end for
for each t в $Treatments$ **do**
    $Z(t) \leftarrow \sum_p^{Plates} EDT_p(t)$, где Z хранит конечные расстояния для каждого лечения
end for

Выбор сильных пертурбаций для объединенного набора данных Cell Painting включал следующие шаги:

- Выбрать 500 самых сильных соединений в соответствии с алгоритмом 1 из BBBC022.
- Пересечь их с BBBC036, включить пересечение в комбинированный набор данных Cell Painting.
- Дополнительно выбрать 50 из BBBC022 и 62 из BBBC036 самых сильных соединений и добавить их в набор данных.
- Выбрать 7 случайных соединений из LINCS, из 20 лучших (по числу связанных обработок) MoA, и добавить их в набор данных.
- Выбрать 28 пересекающихся генов дикого типа между наборами данных BBBC043 и BBBC037 и добавьте их в набор данных.
- Дополнительно выбрать 29 самых сильных пертурабций из BBBC037 и 32 из BBBC043 и добавить их в набор данных.

- Отфильтровать классы с менее чем 100 клетками.
- Добавить контрольные клетки один класс для наборов данных скрининга соединений (BBBC022, BBBC036, LINCS) и другой для наборов данных сверхэкспрессии генов (BBBC037, BBBC043). Контрольные клетки из BBBC036 и LINCS отобраны частично.

Полученный набор данных содержит 8.3 миллиона клеток из 232 планшетов, 488 пертурбаций и 2 типов отрицательных контролей. Более подробная информация о наборе данных представлена на рисунке 20.



.

Рис. 20: Описание комбинированного набора данных Cell Painting. A. Источники пертурбаций в объединенном наборе данных. B. Распределение обработанных и контрольных клеток и источники клеток с пертурбациями. C. Источники клеток внутри каждой клеточной линии. Источник рисунка [82].

3.4.6 Причинно-следственные связи в скрининговых экспериментах

Меняя экспрессию генов клеток или применя на них лекарства, биологи пытаются воздействовать на их состояние и наблюдают за ответной реакцией. Причинный граф для такого эксперимента включает четыре переменные: пертурбации T, изображения O, фенотипы Y и пакетные эффекты (batch-effects) C. В терминах моделирования каузальности это вмешательства, наблюдения, результаты и конфаундеры соответственно. T и O - наблюдаемые переменные, а Y и C - скрытые переменные. Целью является извлечение Y, многомерного представления о пертурбации, которое может быть использовано в дальнейших задачах. Чтобы быть полезным в задаче последующего анализа, Y должен отражать биологически значимое представление, однако в действительности технические вариации, пакетные эффекты C влияют на все остальные элементы этой каузальной модели. C влияет на изображения в результате технических вариаций в процессе получения изображений, на пертурбации - в результате дизайна компоновки планшетов (шаблон расположения пертурбаций в планшетах в скрининговом эксперименте) и на фенотипы - в результате условий окружающей среды. Эти связи показаны на графике (Рисунок 21).

Предполагается, что пертурбация является основной причиной изменения фенотипа клетки. Для извлечения представления об изменениях фенотипа используется WSL с претекстовой задачей классификации лечения. Представления, извлеченные из промежуточных слоев CNN, кодируют все визуальные вариации, в данном случае как пакетные эффекты, так и фенотипы. WSL вместе с пакетной коррекцией (batch correction) поможет отделить фенотипическую вариацию от технической.

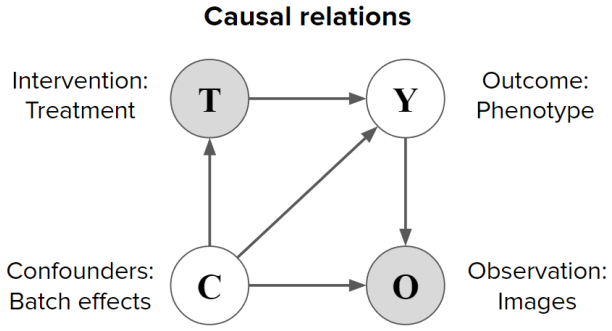


Рис. 21: Причинно-следственная модель для скринингового эксперимента. T обозначает пертурбации, O - изображения (наблюдения), Y - фенотипы (результаты) и C - пакетные эффекты (конфаундеры). Источник рисунка [82].

3.4.7 Результаты и наблюдения

В подразделе обсуждаются результаты, полученные с помощью WSL на объединенном наборе данных Cell Painting *CNN Cell Painting* и моделях, обученных наборах данных-бенчмарках. Предварительно обученная модель на наборе данных ImageNet (также называемом *CNN ImageNet*) и классические признаки, извлеченные с помощью CellProfiler, служат в качестве базы для сравнения.

3.4.7.1 Представления из WSL моделей подчеркивают биологически-релевантные признаки

CNN Cell Painting модель показывает лучшие количественные результаты в биологической задаче (Рисунок 22, голубые точки), чем базовые подходы. Этого следовало ожидать, так как созданные вручную признаки могут упустить некоторую информацию, а модель ImageNet обучена на совершенно других изображениях и не оптимизирована для изображений клеток. Модели, обученные только на соответствующих наборах данных-бенчмарках, не показали последовательного улучшения результатов по сравнению с базовыми подходами (Рисунок 22, зеленые точки).

Для качественной оценки использовалась UMAP-проекция [96] пространства признаков, полученного с помощью *CNN Cell Painting* (рис. 23). В наборе данных BBBC037 пертурбации сгруппированы вместе в соответствии с аннотациями их сигнальных путей, что воспроизводит наблюдения из [85]. В проекциях BBBC022 и BBBC036 многие пертурбации (соединения) также сгруппированы вместе в соответствии с их механизмом действия.

CNN ImageNet демонстрирует схожие или более низкие показатели по сравнению с признаками CellProfiler (Рисунок 22, желтые и розовые точки).

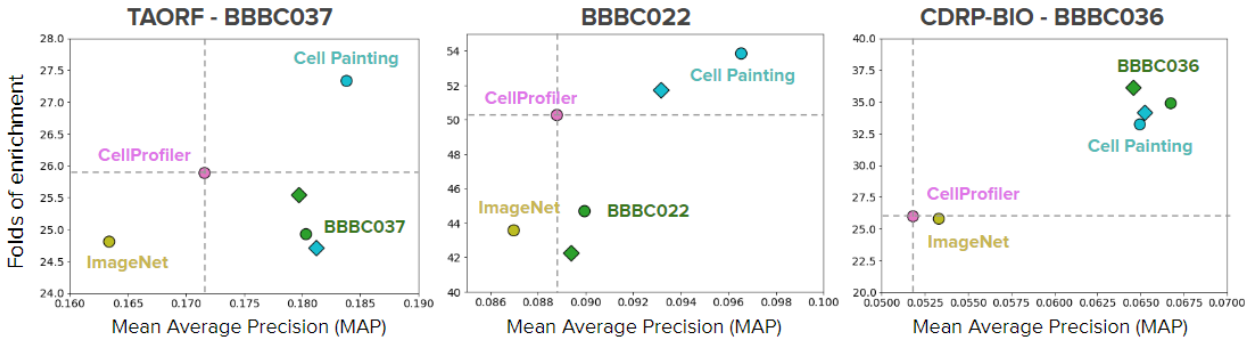


Рис. 22: Количественная оценка представлений для трех бенчмарков в двух метриках: mAP (ось X) и полнота (ось Y). На графике базовыми являются обученные модели CellProfiler (розовый) и CNN ImageNet (желтый): CNN Cell Painting модель (голубой), обученная на соответствующем эталонном наборе данных (зеленый). Схема обучения-валидации 'leave-cells-out' показана кружками, а 'leave-plates-out' ромбами. Источник рисунка [82].

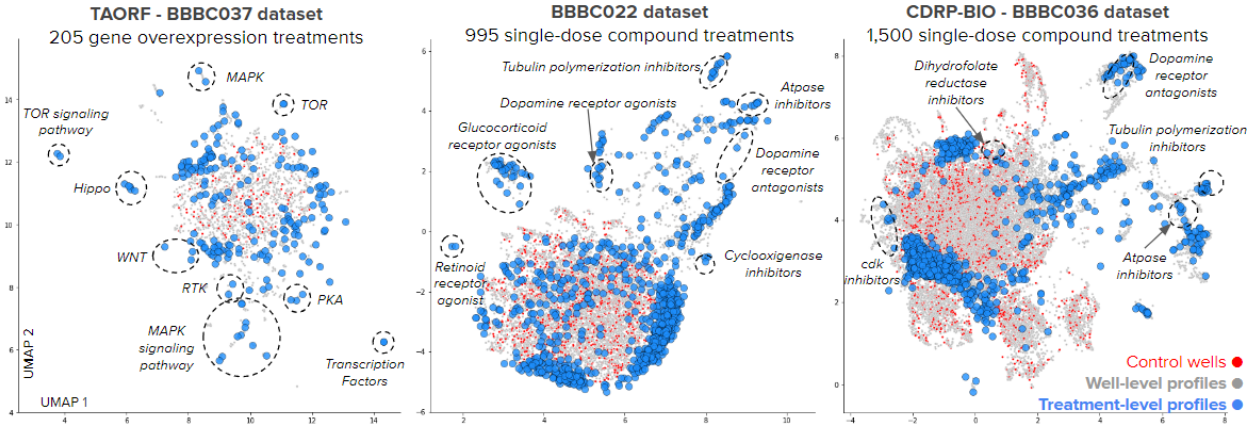


Рис. 23: Графики UMAP признаков уровня лунок, извлеченных с помощью Cell Painting CNN для трех бенчмарков. Серые точки: профили лунок пертурбаций, красные точки: профили уровня лунок отрицательного контроля, синие точки: профили уровня пертурбаций. Пунктирными эллипсами выделены кластеры профилей на уровне пертурбаций с одинаковой биологической аннотацией. Источник рисунка [82].

3.4.7.2 Слабо-контроллируемое обучение распознает как фенотипы, так и пакетные эффекты.

Различные схемы валидации 'leave-plates-out' и 'leave-cells-out' (см. раздел Экспериментальная установка 3.4.3) помогают понять содержание признаков, изученных на основе изображений Cell Painting. В схеме валидации 'leave-cells-out' модель имеет доступ к полному распределению биологической вариации (лечение T) и технической вариации (пакетные эффекты C), в то время как в схеме 'leave-plates-out' модель по-прежнему имеет доступ к полному распределению биологической вариации, но только к части технической вариации.

Значительная разница в точности наблюдалась в задаче классификации пертурбаций для двух разбиений даннных. При разбиении ’leave-cells-out’ обученная CNN может точно классифицировать одиночные клетки как из тренировочного, так и из валидационного набора, в то время как при разбиении ’leave-plates-out’ обученная модель практически не может классифицировать отдельные клетки из валидационного набора (Рисунок 24). Тем не менее, две модели, обученные по разным схемам валидации, демонстрируют схожий результат в последующей биологической задаче (Рисунок 22). Это наблюдение позволяет сделать вывод, что модели пытаются использовать любую информацию, которая может объяснить связь между изображениями и пертурбациями, включая пакетные эффекты. Точность на валидационных данных в ’leave-cells-out’ слишком оптимистична (пакетные эффекты в значительной степени используются для построения связи между наблюдением и вмешательством), напротив, в модели ’leave-plates-out’ слишком пессимистична, поскольку в этом случае модель не знает о технической вариации в планшетах из валиднационной части данных.

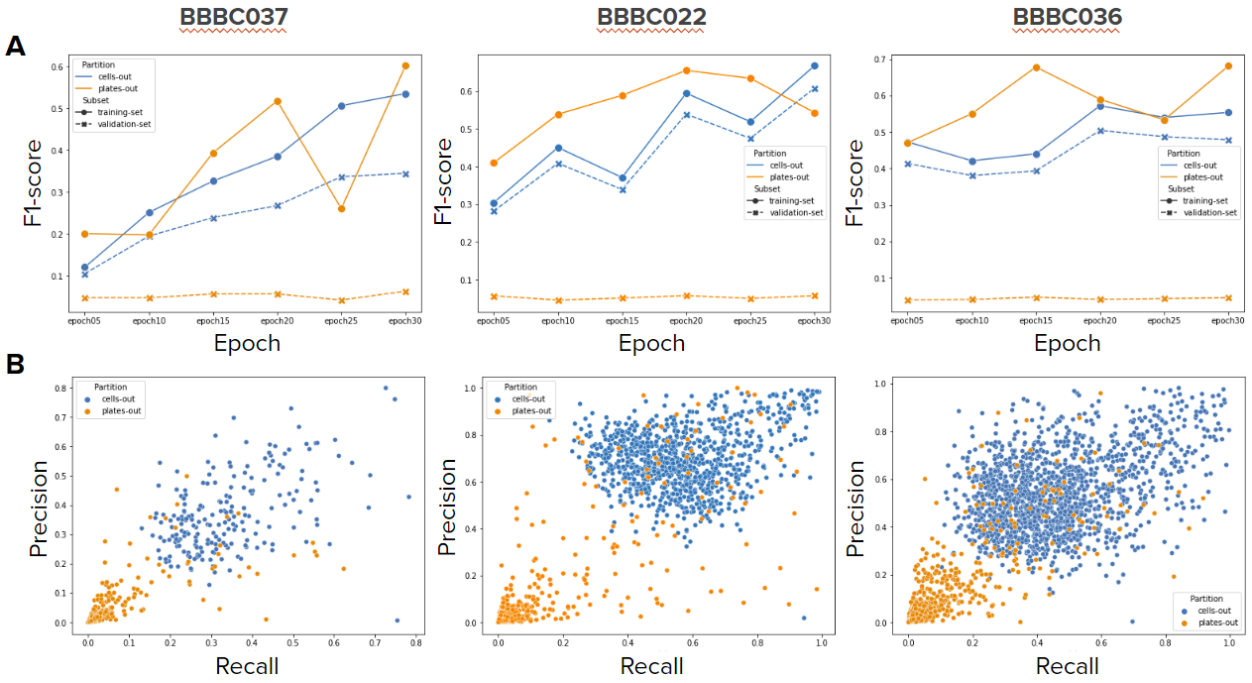


Рис. 24: Классификации в претекстовой задаче (классификация пертурбаций) в бенчмарках при разбиении данных ’leave-plates-out’ (оранжевый) и ’leave-cells-out’ (синий). A. F1-score для обучающего набора (сплошная линия) и валидационного набора (пунктирная линия) для каждой пятой эпохи. B. Полнота (ось X) и точность (ось Y) для последней контрольной точки. Каждая точка - это класс (лечение, включая отрицательный контроль). Источник рисунка [82].

3.4.7.3 Обучение с использованием сильных фенотипов улучшает результаты в биологической задаче

Поскольку в предыдущем разделе было замечено, что контроль распределения вмешивающихся факторов C не изменяет результаты в биологической задаче, пришло время исследо-

вать, что произойдет, если фенотипическое распределение Y будет ограничено. Интуитивно понятно, что WSL минимизирует ошибку в претекстовой задаче, используя конфаундные факторы для правильной классификации пертурбаций со слабым фенотипическим сигналом. Такие пертурбации могут иметь более сильный технический сигнал, а не биологически значимый фенотипический сигнал.

Сильные пертурбации были отобраны путем измерения Евклидова расстояния между профилями отрицательного контроля и пертурбаций, полученными с помощью CellProfiler (см. раздел 3.4.5). Это приближенное значение среднего эффекта лечения (average treatment effect, ATE), причинного параметра для результатов пертурбации. Поскольку мы не можем наблюдать состояние до и после применения пертурбации в одной и той же лунке, это можно рассматривать только как приближенное значение ATE. Для оценки ATE были выбраны характеристики CellProfiler, поскольку они не требуют обучения и поэтому могут служить в качестве независимого приора.

WSL только на сильных пертурбациях в было оценено только в схеме обучения-валидации 'leave-plates-out'. Результаты демонстрируют незначительное улучшение результатов по сравнению с обучением на полных наборах данных (Рисунок 25, синие точки).

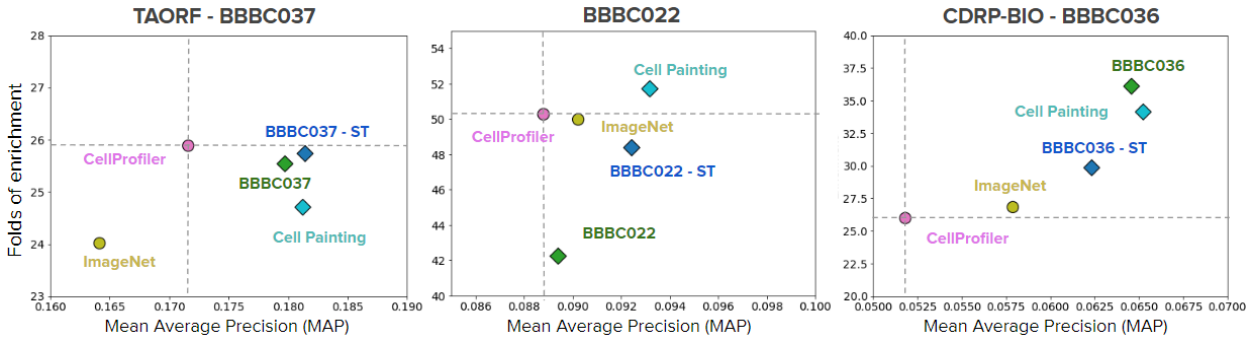


Рис. 25: Количественный результат работы представлений для бенчмарков в двух метриках: mAP (ось X) и степень обогащения (ось Y). На графике CellProfiler (розовый) и CNN ImageNet (желтый): обученная CNN Cell Painting модель (голубой), обученные модели на соответствующем наборе данных (зеленый), обученная на сильных пертурбациях из соответствующего набора данных (синий). Во всех обучающих экспериментах использовалась схема обучения-валидации 'leave-plates-out'. Источник рисунка [82].

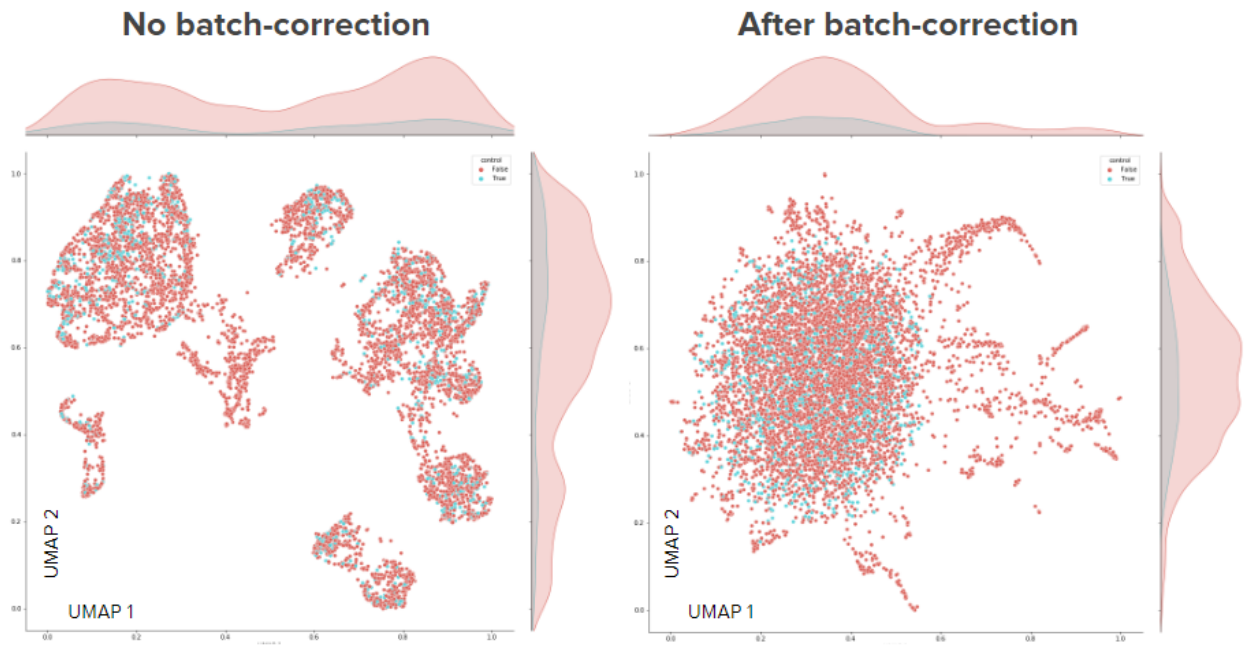


Рис. 26: Качественное влияние пакетной коррекции на графиках UMAP. Левый график показывает UMAP-представление набора данных BBBC022 без пакетной коррекции, а правый - после пакетной коррекции. Точки представляют собой вложения профили уровня лунок (голубой - отрицательный контроль, красный - пертурбация). Графики плотности представлены в верхней и правой частях графиков. Признаки извлечены с помощью модели *Cell Painting CNN*.

3.5 Прогнозирование активности соединений на основе фенотипических профилей и химических структур[2]

В этом разделе кратко пересказывается [97].

Поиск лекарств - дорогой и очень медленный процесс, поскольку существует слишком много теоретически возможных химических соединений, которые невозможно проверить в реальном физическом эксперименте. Даже если фармацевтические компании могут позволить себе тестировать миллионы таких соединений в своих экспериментах, это охватывает лишь небольшую их часть. Кроме того, тестирование этих соединений является дорогостоящим (поскольку они содержат дорогостоящие биологические материалы: первичные клетки, антитела и т.д.) Для выявления соединений-кандидатов используются системы фенотипического анализа. Наконец, этот процесс занимает много времени и требует времени экспертов для проведения анализов.

Для снижения стоимости скрининговых исследований при поиске лекарств, существует возможность использования вычислительных методов, например, современное глубинное обучение может позволить точно предсказать активность соединений в анализах. Предыдущие работы пытались использовать методы машинного обучения только с морфологическими

[2]Статья размещена в качестве препринта и подана в журнал

данными [98] [99].

В данном проекте целью является оценка предсказательной способности представлений химических структур, морфологических профилей клеток и профилей экспрессии генов для вычислительного прогнозирования результатов анализов в промышленных масштабах. Гипотеза заключается в том, что прогностические возможности этих источников данных дополняют друг друга, и эти источники данных могут быть использованы вместе для дальнейшего повышения успешности процесса скрининга лекарств. Кроме того, тестируются базовые методы объединения данных, хотя это не является основной задачей проекта, и этот вопрос может быть исследован в дальнейшем.

3.5.1 Материалы и методы

Набор данных состоит из четырех частей: матрица взаимодействий между анализами и соединениями, морфологические профили, профили экспрессии генов и представления химических структур. Вся информация была собрана из анализов, полученных в ходе экспериментов по поиску лекарств, проводимых в Broad Institute [86].

Матрица взаимодействия анализов и соединений является основной частью данных. Строки - это соединения (представленные в виде SMILES-строк), а столбцы - пробы. Ячейки заполнены 1 (есть взаимодействие, 'попадание'), 0 (отсутствие взаимодействия, 'непопадание') и могут быть пустыми (это соединение не тестировалось с данным пробой). 'Попадания' и 'непопадания' вместе взятые также называются показаниями. Только часть соединений была протестирована с каждой конкретной пробой, что означает, что матрица разреженная. Первоначально матрица содержала 496 анализов, но мы отфильтровали их с помощью следующей процедуры:

- Применили все фильтры Pan-assay Interference (PAINS) [100], реализованные в RDKit, которые удалили 786 соединений, в результате чего осталось 16 210 соединений.
- Удалены все анализы без взаимодействий, их количество уменьшилось с 496 до 437.
- Рассчитали индекс Жаккара для взаимодействий между анализами, для выявления анализов несущих избыточную информацию. Матрица с индексами Жаккара (437×437) была преобразована в двоичную по пороговому значению 0.7, а затем была применена иерархическая кластеризация с использованием косинусного расстояния, которая была использована для фильтрации.
- Заключительное удаление часто взаимодействующих соединений, определяемых как соединения, которые взаимодействуют с 10% анализов (30 или более) и, наконец, удаление анализов не содержащих ни одного взаимодействия. Окончательный набор данных состоит из 16 170 соединений и 270 анализов.

Большинство анализов в окончательном наборе данных – клеточные, другие представленные типы анализов: биохимические, бактериальные и дрожжевые, а также есть слабо пред-

ставленные категории анализов, такие как грибковые, гомогенные, вирусы и черви (Рисунок 27).

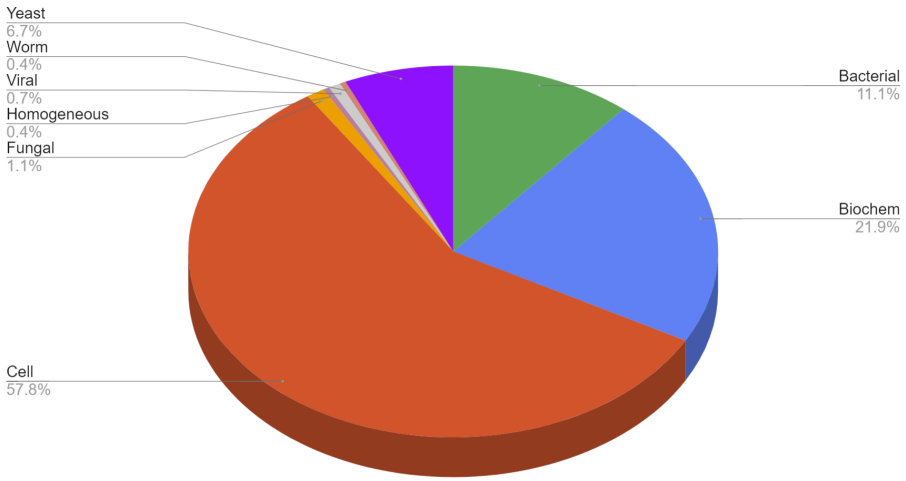


Рис. 27: Распределение типов анализов в окончательном наборе данных. Источник рисунка [97].

Эксперименты с методом Cell Painting [10] [47] [101] [102] проводились для получения пятиканальных изображений высокого разрешения. Эти изображения обрабатывались программой CellProfiler для сегментирования и получения ~ 1700 морфологических признаков на уровне отдельных клеток. Затем они были агрегированы до уровня лунок, как в [48]. К профилям на уровне лунок применялось сферическое преобразование (см. также 3.4.4.2) для коррекции эффектов партии. Для расчета сферического преобразования использовали ячейки с DMSO из всех планшетов. Затем профили были агрегированы до уровня пертурбации (обозначены как MO, за исключением Таблицы 3.5.2 и Таблицы 2). Эксперименты также проводились с профилями без применения сферического преобразования, хотя дополнительный прирост в метриках в этом случае, вероятнее всего, из-за пакетных эффектов (Рисунок 28).

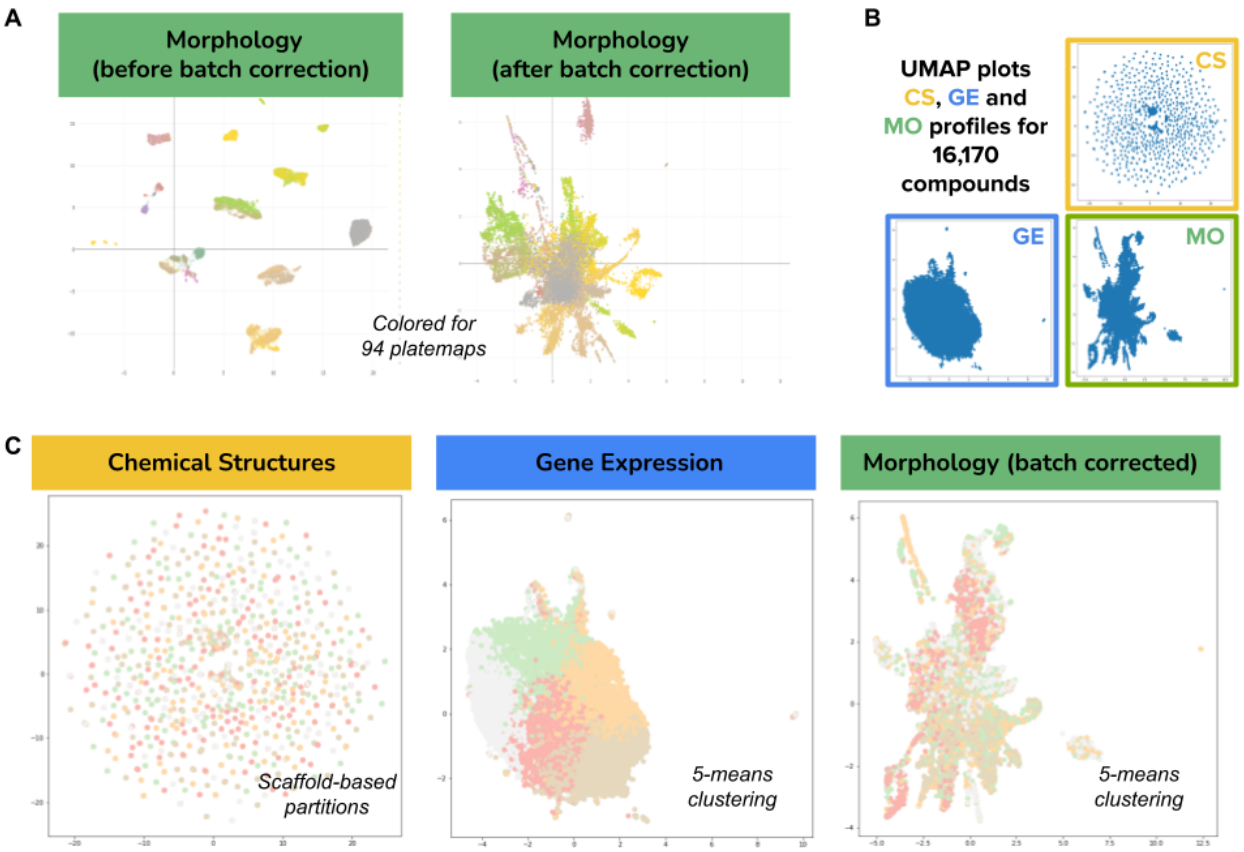


Рис. 28: Представления на уровне соединений в трех различных типах данных. Визуализация построена с помощью UMAP. A. Пространство признаков морфологии изначально было сгруппировано по технической вариации (группам планшетов), которая была скорректирована с помощью сферического преобразования. Цветовая палитра для 94 групп планшетов является непрерывной и может иметь схожие тона для групп планшетов. B. Представления в трех различных типах данных. C. Те же представления, что и в B, раскрашенные по кластерам, полученными в ходе экспериментов с кросс-валидацией (см. раздел 'Эксперименты и результаты'). Источник рисунка [97].

3.5.2 Эксперименты и результаты

Эксперименты проводились используя нескольких подходов для разбиения данных на тренировочные и тестовые. Все эти подходы имеют одну и ту же идею: мы хотим предсказать взаимодействие анализа с соединением для соединений, которые отличаются по сравнению с тренировочными данными. С практической точки зрения, поиск схожих химических структур для соединения с известной активностью имеет мало смысла. Наиболее близким к такому реальному сценарию является разбиение на основе скаффолдов (для кросс-валидации из пяти разбиений), полученное с помощью кластеризации Бемиса-Мурко [58] [103].

В дополнение к разбиениям на основе скаффолдов, были сделаны разбиения на основе морфологических признаков и признаков экспрессии генов. Для разбиения данных по экспрессии генов были кластеризованы признаки экспрессии генов, а для разбиения по морфологии были кластеризованы признаки морфологии, прошедшие коррекцию по партиям

(для кросс-валидации из пяти разбиений) с использованием одноразмерной кластеризации K-средних (реализация [104]), см. кластеризацию на рисунке 28.

В качестве основной метрики мы используем медианную AUROC с порогом 0.9; этот же порог использовался в более ранних работах по предсказанию активности соединений в пробах [60] [105] [106].

При обучении моделей, использовалась логистическая регрессия как функция потерь для каждого анализа, а общая потеря - это сумма потерь для всех анализов. Пакет содержит информацию о 50 соединениях. Если нет данных о взаимодействии между анализом и соединением, оно игнорируется при обновлении градиента. В каждом обучении оптимизация гиперпараметров проводилась перед обучением (см. 3.5.1).

Результаты показывают, что с помощью данных о морфологии можно точно предсказать наибольшее количество анализов с медианой $AUROC > 0.9$ по результатам кросс-валидации (28 для морфологии, 19 для экспрессии генов и 16 для химических структур), см. рисунок 30. Хотя при более низких порогах AUROC (0.7) химические структуры сравнялись с морфологией (также см. рисунок 32). Интересно, что все три типа данных не имеют общих хорошо предсказанных анализов (Рисунок 30), а каждая пара типов данных имеет несколько общих хорошо предсказанных анализов, что означает, что различные источники данных содержат значительно взаимодополняющую информацию.

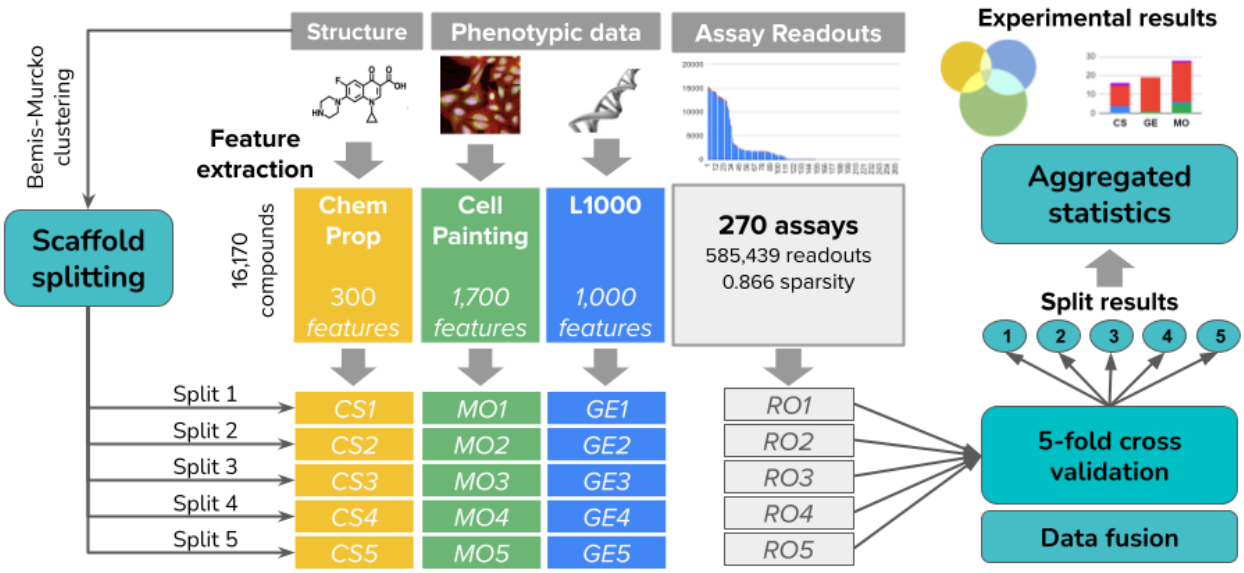


Рис. 29: Иллюстрация экспериментальной методики. Источник рисунка [97].

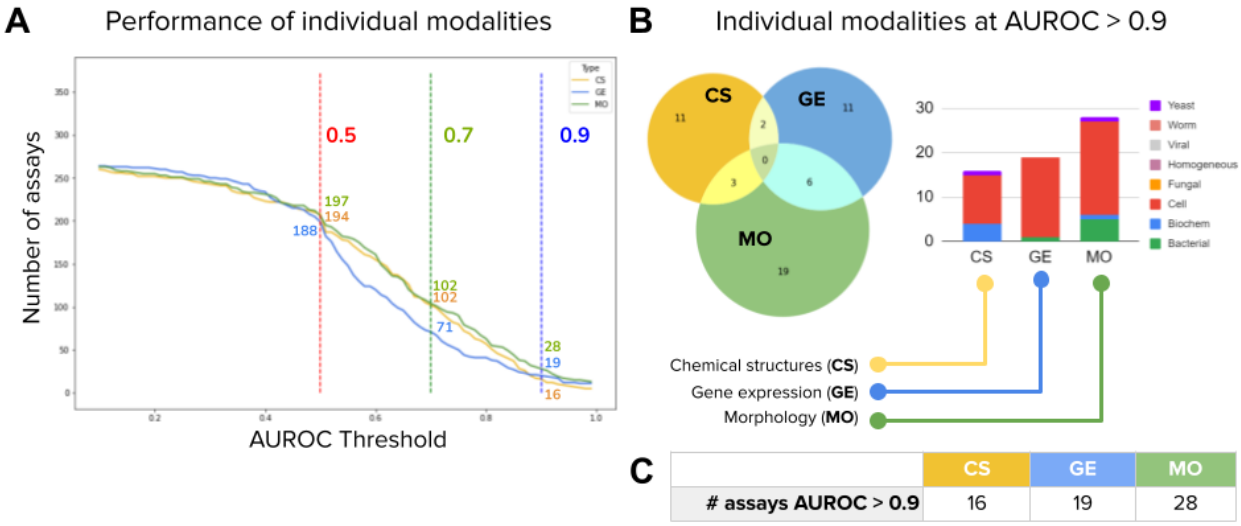


Рис. 30: A. Эффективность отдельных типов данных измеряется как количество анализов (вертикальная ось), предсказанных с AUROC выше определенного порога (горизонтальная ось). B. Диаграммы Венна показывают количество точно предсказанных анализов (медиана $AUROC > 0.9$ по результатам кросс-валидации), которые являются общими или уникальными для каждого типа данных. Гистограмма показывает распределение типов анализов, точно предсказанных отдельными типами данных. C. Количество точно предсказанных (медиана $AUROC > 0,9$ по результатам кросс-валидации) анализов по каждой отдельного типа данных. Источник рисунка [97].

Можно использовать несколько типов данных для предсказания взаимодействия анализа и соединения. Для их объединения в один предиктор, мы использовали два подхода: а) ***Раннее слияние*** - векторы признаков объединяются в один вектор и используются в качестве входа для нейронной сети. б) ***Позднее слияние*** - для каждого типа данных обучается отдельная модель, затем результаты предсказаний объединяются, используя максимальную вероятность среди предсказаний для каждой пары 'соединение-анализ'.

Согласно экспериментам (Таблица 2), раннее слияние данных не дало никакой дополнительной эффективности, более того, оно ухудшило показатели. Результаты для отдельных типов данных показали, что они не имеют много общих хорошо предсказанных анализов (Рисунок 30), и когда мы объединяем векторы характеристик, мы вносим шум в анализы, которые могут быть хорошо предсказаны для одного типа данных, но не могут быть предсказаны для другого. Позднее слияние работает лучше на практике, но улучшение незначительное (31 хорошо предсказанных анализов при комбинации CS+MO против 28 при использовании только MO). Протестированные нами подходы к слиянию довольно просты, и необходимо провести дополнительные исследования для поиска более эффективных методов слияния. В качестве дополнительной метрики мы измерили ретроспективную оценку, которая представляет собой симуляцию наилучшего возможного слияния данных. В этом анализе мы знаем прогнозы заранее. Использование слияния вместе с индивидуальными типами данных может дать 7-17% прироста предсказанных анализов с высокой точностью (Рисунок 31).

	Scaffold-based splits — Real world setting					
Avg. assays tested: 233.2	MO	MO-BC	GE	GE-S	CS-GC	CS-MF
Mean AUPRC	**0.261**	0.252	0.234	0.231	0.232	0.223
Mean AUROC	**0.657**	0.637	0.592	0.587	0.630	0.610
$AUC > 0.5$	**160.0**	151.4	139.2	138.8	150.2	146.8
$AUC > 0.7$	**91.2**	83.2	57.2	59.4	88.4	81.6
$AUC > 0.9$	27.0	**28.0**	21.8	18.4	21.6	21.0
	Gene expression splits (simulation)					
Avg. assays tested: 232.0	MO	MO-BC	GE	GE-S	CS-GC	CS-MF
Mean AUPRC	**0.263**	0.248	0.222	0.201	0.246	0.244
Mean AUROC	**0.664**	0.642	0.577	0.561	0.647	0.658
$AUC > 0.5$	155.6	150.2	127.6	127.2	153.2	**157.4**
$AUC > 0.7$	94.4	86.2	45.4	46.6	94.2	**99**
$AUC > 0.9$	**27.4**	23.6	14.2	12.6	22.6	22.4
	Morphology(bc)-based splits (simulation)					
Avg. assays tested: 179.8	MO	MO-BC	GE	GE-S	CS-GC	CS-MF
Mean AUPRC	0.224	0.207	0.199	0.198	0.225	**0.245**
Mean AUROC	0.634	0.600	0.562	0.564	0.631	**0.652**
$AUC > 0.5$	142	128.6	125.4	126.2	140.8	**143.6**
$AUC > 0.7$	**72.8**	63.0	49.2	49.2	81.0	82.6
$AUC > 0.9$	**21.6**	17.0	14.4	13.6	19.4	22.6
	Random splits (simulation)					
Avg. assays tested: 232.4	MO	MO-BC	GE	GE-S	CS-GC	CS-MF
Mean AUPRC	**0.259**	0.247	0.234	0.228	0.244	0.242
Mean AUROC	**0.670**	0.643	0.601	0.595	0.659	0.651
$AUC > 0.5$	**163.6**	154.2	145.6	144.0	157.6	157.8
$AUC > 0.7$	**97.2**	88.4	61.8	66.0	94.8	94.0
$AUC > 0.9$	**26.2**	22.0	20.4	17.4	25.8	23.4

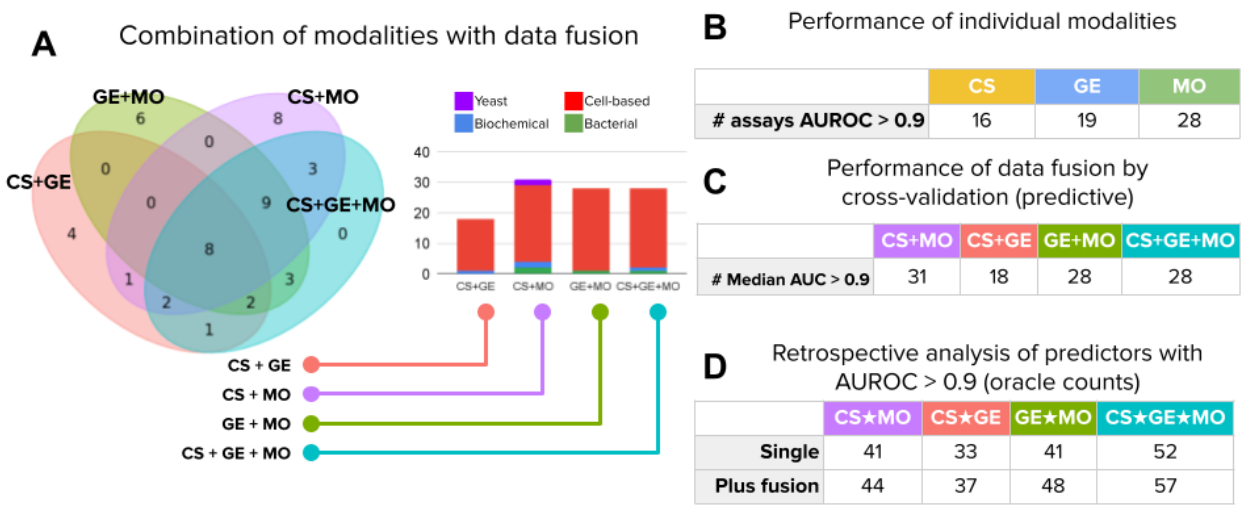
Таблица 1: Результаты экспериментов с кросс-валидацией (5 разбиений). В таблицах представлены средние результаты экспериментов для кросс-валидации в соответствии с различными подходами к разбиению данных. Метрики: средняя площадь под кривой 'точность-полнота' (Mean AUPRC) для 5 разбиений, средняя площадь под ROC-кривой (Mean AUROC) для 5 разбиений, среднее количество предсказанных анализов, пороговое значение площади под ROC-кривой ($AUC > 0.5$, $AUC > 0.7$, $AUC > 0.9$) для 5 разбиений. Источники используемых данных: MO: морфологические признаки без пакетной коррекции. MO-BC: морфологические признаки с коррекцией. GE: особенности экспрессии генов. CS-GC: графовые сверточные характеристики (GC). CS-MF: отпечатки Моргана. Среднее количество анализов в тестовом наборе отличается в разных типах данных, так как невозможно оценить анализ без 'попаданий' в тестовом наборе (которые отличаются, так как мы использовали разные подходы к разбиению на тренировочный и тестовые). Источник таблицы [97].

	Baseline: independent modalities (scaffold-based partitions)					
	MO		GE		CS	
	Mean	Std	Mean	Std	Mean	Std
Mean AUPRC	0.252	0.021	0.234	0.038	0.232	0.036
Mean AUROC	0.637	0.021	0.592	0.034	0.630	0.018
$AUC > 0.5$	151.4	13.502	139.2	13.773	150.2	13.255
$AUC > 0.7$	83.2	11.100	57.2	16.316	88.4	6.066
$AUC > 0.9$	28.0	4.848	21.8	8.198	21.6	6.229

	Early fusion — concatenation (scaffold-based partitions)							
	GE-MO		MO-CS		GE-CS		GE-MO-CS	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Mean AUPRC	0.214	0.045	0.251	0.021	0.219	0.028	0.221	0.021
Mean AUROC	0.586	0.038	0.632	0.031	0.577	0.061	0.582	0.038
$AUC > 0.5$	138.8	18.377	151.8	19.905	138.6	26.773	137.2	22.928
$AUC > 0.7$	59.2	12.215	87.8	15.531	63.4	21.663	59.8	14.516
$AUC > 0.9$	16.0	4.743	23.6	4.159	17.0	2.292	20.4	4.278

	Late fusion — max pooling (scaffold-based partitions)							
	GE-MO		MO-CS		GE-CS		GE-MO-CS	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Mean AUPRC	0.261	0.026	0.267	0.034	0.251	0.039	0.265	0.032
Mean AUROC	0.652	0.028	0.661	0.027	0.645	0.026	0.665	0.031
$AUC > 0.5$	157.4	11.845	157.8	13.773	155.6	16.637	159.0	15.017
$AUC > 0.7$	86.0	9.670	98.8	7.430	87.0	9.566	96.4	10.877
$AUC > 0.9$	29.4	6.618	29.4	5.128	23.8	8.843	28.0	5.148

Таблица 2: Оценка отдельных и комбинированных типов данных для моделей, обученных с помощью разбиения на фрагменты. Показатели: средняя площадь под кривой 'точность-полнота' (Mean AUPRC) для 5 разбиений, средняя площадь под ROC-кривой (Mean AUROC) для 5 разбиений, среднее количество предсказанных анализов, пороговое значение площади под ROC-кривой ($AUC > 0.5$, $AUC > 0.7$, $AUC > 0.9$) для 5 разбиений. Также приведены стандартные отклонения. Источник таблицы [97].

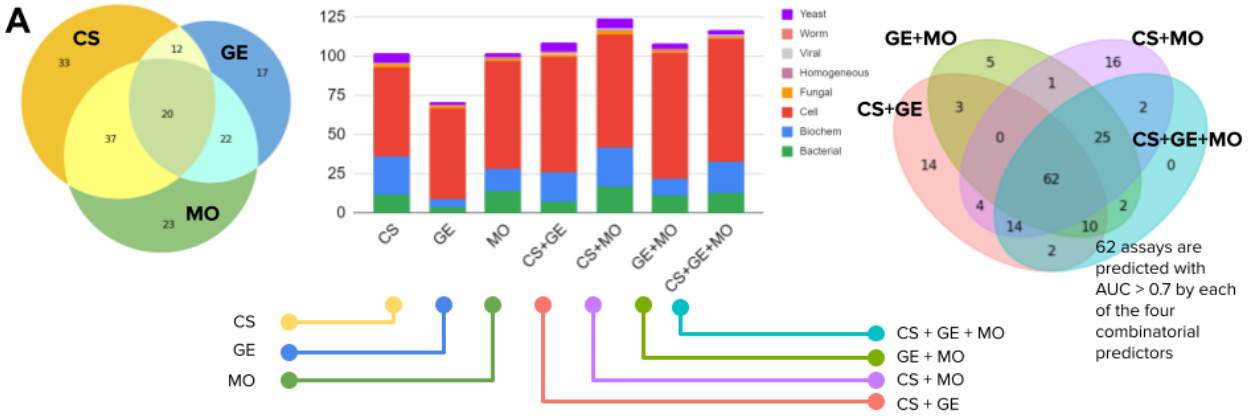


	CS	GE	MO
# assays AUROC > 0.9	16	19	28

	CS+MO	CS+GE	GE+MO	CS+GE+MO
# Median AUC > 0.9	31	18	28	28

	CS★MO	CS★GE	GE★MO	CS★GE★MO
Single	41	33	41	52
Plus fusion	44	37	48	57

Рис. 31: Количество точно предсказанных анализов (медиана AUROC по разделениям выше 0.9). A. Слева - диаграмма Венна точно предсказанных анализов с использованием позднего слияния, справа - гистограммы, показывающие распределение точно предсказанных типов анализов с поздним слиянием. B. Количество точно предсказанных анализов по отдельным типам данных. C. Количество точно предсказанных анализов для комбинированных методов с использованием позднего слияния. Подсчеты приведены для медианного и средней площади под ROC-кривой по разбиениям. D. Количество точно предсказанных анализов для ретроспективного анализа. 'Single' - простое объединение точно предсказанных анализов по отдельным типам данных. 'Plus fusion' - объединение точно предсказанных анализов с отдельными типами данных плюс предиктор позднего слияния. Источник рисунка [97].



B

	CS	GE	MO	CS+GE	CS+MO	GE+MO	CS+GE+MO
Mean AUC	0.630	0.592	0.637	0.645	0.661	0.652	0.665
Mean AUC > 0.7	98	63	100	97	115	106	111
Median AUC > 0.7	102	71	102	109	124	108	117
Retrospective ★	-	-	-	151	157	139	174

Рис. 32: Количество предсказанных анализов с умеренной точностью (медианная площадь под ROC-кривой по расщеплениям выше 0.7). А. Слева диаграмма Венна предсказанных анализов с отдельными типами данных, в центре гистограмма предсказанных типов анализов по отдельным типам данных и в сочетании (позднее слияние), справа диаграмма Венна предсказанных анализов с объединенными типами данных (позднее слияние). В. Таблица эффективности отдельными типами данных и комбинированных (позднего слияния). Показатели: средняя площадь под ROC-кривой (Mean AUC) для 5 разбиений, среднее количество предсказанных анализов, пороговое значение площади под ROC-кривой ($AUC > 0.7$) для 5 разбиений. Источник рисунка [97].

	CS	GE	MO	CS+GE	CS+MO	GE+MO	CS+GE+MO	Evaluated assays
Cell-based	7.05%	11.54%	13.46%	10.90%	16.03%	17.31%	16.67%	156
Biochemical	6.78%	0.00%	1.69%	1.69%	3.39%	0.00%	1.69%	59
Bacterial	0.00%	3.33%	16.67%	0.00%	6.67%	3.33%	3.33%	30
Yeast	5.56%	0.00%	5.56%	0.00%	11.11%	0.00%	0.00%	18
Fungal	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	3
Viral	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2
Worm	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1
Homogeneous	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1

Таблица 3: Количественно предсказанных анализов по типам с площадью под ROC-кривой (AUROC) более 0.9 (медианная по разбиениям). Источник таблицы [97].

	CS	GE	MO	CS+GE	CS+MO	GE+MO	CS+GE+MO	Evaluated assays
Cell-based	36.54%	37.18%	44.23%	47.44%	46.15%	51.28%	50.00%	156
Biochemical	40.68%	8.47%	23.73%	32.20%	42.37%	18.64%	33.90%	59
Bacterial	40.00%	13.33%	46.67%	23.33%	56.67%	36.67%	43.33%	30
Yeast	33.33%	11.11%	11.11%	33.33%	33.33%	16.67%	16.67%	18
Fungal	66.67%	33.33%	33.33%	33.33%	66.67%	33.33%	33.33%	3
Viral	50.00%	0.00%	0.00%	50.00%	50.00%	0.00%	50.00%	2
Worm	0.00%	100.00%	100.00%	100.00%	0.00%	100.00%	0.00%	1
Homogeneous	0.00%	0.00%	100.00%	0.00%	100.00%	100.00%	100.00%	1

Таблица 4: Количественно предсказанных анализов по типам с площадью под ROC-кривой (AUROC) более 0.7 (медианная по разбиениям). Источник таблицы [97].

4 Выводы

Прогресс в области вычислительной эффективности и высокопроизводительных методов биологии обоюден: рост вычислительных мощностей проложил путь для развития высокопроизводительных методов в биологии. Это, в свою очередь, задействует вычислительные мощности, создавая новые массивы данных, которые нуждаются в анализе. В рамках этих процессов возникли новые научные направления и вычислительные методы. Обработка изображений ждала своей очереди, совершенствуя свои лабораторные и вычислительные методы в течении некоторого времени, хотя первые попытки анализа изображений были успешными и привели к возникновению нового направления [9] [19].

В середине 2010-х годов, когда начался переход от классического анализа изображений к анализу изображений на основе глубокого обучения и GPU-вычисления стали доступными, анализ биологических изображений пережил стремительный взлет. К этому времени протоколы для микроскопии устоялись, и появились новые специфические протоколы [10] и методы (например, микроскопия сверхвысокого разрешения)[107]. Методы классификации, обнаружения и сегментирования изображений были быстро приняты сообществом вычислительных биологов для решения конкретных задач [7][108].

За последние несколько лет научная область сегментирования клеток (ядер) стала более продвинутой, помимо новых методов (включая попытки построить общий метод сегментирования клеток) и дополнительных методов постобработки, также были опубликованы новые крупномасштабные наборы данных и инструменты аннотации [7]. В настоящее время разрабатываются новые методы, обычно специфичные для конкретной категории данных, но сообщество стремится к общим моделям сегментирования и сегментированию в 3D [7].

В рамках перерождения фенотипического подхода к поиску лекарств [46], одной из областей анализа биологических изображений, представляющих особый интерес и широко применяющих методы глубокого обучения [48] [51], является профилирование на основе изображений [10]. Ожидается, что в ближайшем будущем оно будет развиваться как с биологической, так и с вычислительной стороны [52]. С вычислительной стороны все внимание приковано к методам глубокого обучения без учителя. Гипотеза в том, что они будут в большей степени способны распознавать биологически значимые признаки отдельных клеток.

Данная диссертация посвящена использованию методов глубокого обучения для сегментирования и фенотипического профилирования одиночных клеток. Со стороны сегментирования, в диссертации представлен обзор состояния направления сегментирования клеточных ядер, создан инструмент аннотирования для создания наборов данных сегментирования клеток (ядер) и оценка одного из методов постобработки для сегментирования ядер. С точки зрения фенотипирования, в диссертации представлены слабо-контролируемое обучение для масштабного профилирования на основе изображений и оценка предсказательной силы различных типов клеточных данных в задаче поиска лекарств.

1. В обзорной статье мы представили описание методов глубокого обучения для 2D и 3D данных, описание наборов данных и инструментов аннотации. Было высказано

несколько важных замечаний относительно текущего состояния области сегментирования клеточных ядер с расчетом на то, что сообщество примет их во внимание. Был разработан помощник для принятия решений по выбору метода сегментирования.

2. Плагин AnnotatorJ разработан для популярной программы обработки изображений ImageJ/Fiji, который использует предварительно обученные модели на основе U-Net для облегчения аннотирования изображений клеточных ядер. Эксперименты с экспертами-аннотаторами показали, что AnnotatorJ сокращает время необходимое для аннотирования и повышает точность созданных аннотаций.

3. Подход аугментация во время тестирования был экспериментально оценен для двух популярных методов глубокого обучения: U-Net и Mask R-CNN. Согласно полученным результатам, в среднем можно получить дополнительную точность сегментирования с помощью аугментации во время тестирования, хотя в отдельных случаях это не гарантировано. В случаях с недостаточно оптимизированными моделями качество сегментирования незначительно ухудшается. Просмотр изображений также показал, что в основном изменяются маски на границах объектов, хотя в редких случаях, особенно в случае Mask R-CNN, так как он основан на сегментировании объектов в целом (улучшая сегментирование путем удалением ложных или добавления истинных объектов). Рекомендацией будет использование этого подхода для анализа неопределенных областей при сегментировании. Вычислительные затраты на предсказания увеличиваются, но это является проблемой только в очень больших масштабах или если предсказания выполняются на CPU.

4. Сверточные нейросети, обученные с использованием подхода слабо-контролируемого обучения, были сравнены с классическими подходом и предварительно обученной моделью на ImageNet в задаче профилирования на трех бенчмарках. Основной вывод заключается в том, что, максимизируя техническую и фенотипическую вариаиции, модель при слабо-контролируемом обучении более способна впоследствии генерировать биологически значимые представления. Пакетная коррекция оказалась одним из решающих элементов для этого. В ходе данного проекта был создан комбинированный набор данных Cell Painting и разработан программный инструмент DeepProfiler для профилирования изображений моделями глубокого обучения. В результате экспериментов на этих данных была обучена модель для извлечения признаков из данных Cell Painting.

5. Мы оценили предсказательную способность различных типов данных: морфологии, профилей экспрессии генов и химических структур для предсказания взаимодействия проб с химическими соединениями. Результаты показывают, что эти три типа данных по отдельности могут предсказать 6-10% анализов с высокой точностью. Согласно нашим экспериментам, эти методы оказались взаимодополняющими и в сочетании могут обеспечить до 21% анализов, которые могут быть предсказаны с высокой точностью

или до 64%, если приемлема более низкая точность.

Список иллюстраций

1 Примеры сегментирования слева направо: исходное изображение, семантическое сегментирование, сегментирование отдельных объектов. Источник изображения и аннотации сегментирования: набор данных Data Science Bowl 2018 [11]. 5

2 Микроскопический снимок и маска сегментирования, полученная с помощью пороговой обработки Otsu [31] из пакета Scikit-image [32]. Источник изображения: набор данных Data Science Bowl 2018 [11]. 9

3 Стандартная архитектура U-Net. 10

4 Стандартная архитектура Mask R-CNN. 11

5 Пример изображения, полученного методом Cell Painting, органоиды (подписи сверху) и красители (подписи внизу). Изображение из набора данных BBBC022 [47]. 12

6 A. SMILES представление Ибупрофена и его сгенерированное графическое представление. B. Скаффолд Бемиса-Мурко молекулы Ибупрофена. графическое представление и скаффолд были сгенерированы программным пакетом RDKit (`https://www.rdkit.org/`). 13

7 Интерфейс ассистента для выбора методов сегментирования. Слева расположено дерево типов микроскопии. В правом верхнем углу расположены элементы управления фильтрацией для выбора 2D/3D-методов и конкретных методов для решения сложностей сегментирования. В правом нижнем углу находится список методов сегментирования. 15

8 Первый шаг аннотирования с помощью помощника для построения контуров: инициализация контура путем рисования линии на объекте. Цифрами и зелеными рамками показаны шаги, которые необходимо выполнить в интерфейсе. Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11]. 17

9 Инициализированный контур с помощью предварительно обученной модели сегментирования глубокого обучения (справа). Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11]. 17

10 Ручное уточнение контура объекта. Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11]. 18

11 После уточнения границ, объект добавляется нажатием клавиши ‘Q‘. Источник изображения микроскопии: набор данных Data Science Bowl 2018 [11]. . . 18

12 Принцип работы аугментации во время тестирования. Входные данные: предсказание на нескольких аугментированных экземплярах одних и тех же тестовых изображений с обученными моделями. Для объединения предсказаний использовалось голосование на уровне пикселей (U-Net), или комбинация совмещения объектов и метода голосования (Mask R-CNN). Источник рисунка [20]. 20
13 Результаты аугментации во время тестирования для Mask R-CNN (delta) для разбиений данных на тренировочные и тестовые. Каждая точка - изображение. Столбцы - эпоха оубчения. Штриховая линия - среднее, обычная линия - медиана. A. Fluorescent_5. B. Разбиение Fluorescent_15 (кросс-валидация 1) C. Fluorescent_30. D. Tissue_5. E. Tissue_15 (кросс-валидация 1) F. Разбиение Tissue_30. Источник рисунка [20]. 23
14 Средний индекс Жаккара на тестовых данных и влияние аугментации во время тестирования для U-Net ($delta$). A. Средняя $delta$, аугментации во время обучения не использовались. B. Средняя $delta$, аугментации во время обучения использовались. C. Средний индекс Жаккара, аугментации во время обучения не использовались. D. Средний индекс Жаккара, аугментации во время обучения не использовались. Источник рисунка [20]. 24
15 Примеры сегментирования. A. Сегментирования U-Net. Первый столбец - оригинальное изображение, второй столбец - предсказания без аугментации во время тестирования в сравнении с аннотированной маской, третий столбец - предсказания с аугментацией во время тестирования в сравнении с аннотированной маской. Красным цветов показаны ложно отрицательные сегментации, зеленым показаны истинно положительные сегментации и синим ложно положительные сегментации. Четвертый столбец - предсказания с аугментацией во время тестирования до фильтрации по пороговому значению, пятый столбец - увеличенные фрагменты из предыдущего столбца. Строки - примеры изображений. B. Сегментирования Mask R-CNN. Столбцы аналогичны первым трем столбцам из A, строки - примеры изображений. Источник рисунка [20]. 25
16 Результаты на тестовой выборке соревнования DSB 2018 стадии 2 аугментации во время тестирования, комбинированного с методом [37]. Источник рисунка [20]. 25
17 Пример контроля качества для набора данных BBBC022. Слева: График PCA для первых двух принципиальных компонент. Каждая точка - лунка, цвета обозначают планшеты. Наблюдается кластер с выбросами. Справа: примеры изображений из лунок-выбросов. Видно, что эти изображения не в фокусе. . . 27

18 Типичное использование DeepProfiler. 1. Обучение сети классификации изображений 2. Обученная модель используется для извлечения представлений 3. Представления используются для последующих задач по анализу данных. Шаги 1 и 2 на изображении выполняются с помощью DeepProfiler, шаг 3 не входит в DeepProfiler и является предпочтением пользователя. Использованы изображения микроскопии из набора данных BBBC021 [72]. 29
19 Вычислительная затраты на стратегии профилирования. Источник рисунка [82]. 31
20 Описание комбинированного набора данных Cell Painting. A. Источники пертурбаций в объединенном наборе данных. B. Распределение обработанных и контрольных клеток и источники клеток с пертурбациями. C. Источники клеток внутри каждой клеточной линии. Источник рисунка [82]. 34
21 Причинно-следственная модель для скринингового эксперимента. T обозначает пертурбации, O - изображения (наблюдения), Y - фенотипы (результаты) и C - пакетные эффекты (конфаундеры). Источник рисунка [82]. 35
22 Количественная оценка представлений для трех бенчмарков в двух метриках: mAP (ось X) и полнота (ось Y). На графике базовыми являются обученные модели CellProfiler (розовый) и CNN ImageNet (желтый): CNN Cell Painting модель (голубой), обученная на соответствующем эталонном наборе данных (зеленый). Схема обучения-валидации 'leave-cells-out' показана кружками, а 'leave-plates-out' ромбами. Источник рисунка [82]. 36
23 Графики UMAP признаков уровня лунок, извлеченных с помощью Cell Painting CNN для трех бенчмарков. Серые точки: профили лунок пертурбаций, красные точки: профили уровня лунок отрицательного контроля, синие точки: профили уровня пертурбаций. Пунктирными эллипсами выделены кластеры профилей на уровне пертурбаций с одинаковой биологической аннотацией. Источник рисунка [82]. 36
24 Классификации в претекстовой задаче (классификация пертурбаций) в бенчмарках при разбиении данных 'leave-plates-out' (оранжевый) и 'leave-cells-out' (синий). A. F1-score для обучающего набора (сплошная линия) и валидационного набора (пунктирная линия) для каждой пятой эпохи. B. Полнота (ось X) и точность (ось Y) для последней контрольной точки. Каждая точка - это класс (лечение, включая отрицательный контроль). Источник рисунка [82]. . . 37
25 Количественный результат работы представлений для бенчмарков в двух метриках: mAP (ось X) и степень обогащения (ось Y). На графике CellProfiler (розовый) и CNN ImageNet (желтый): обученная CNN Cell Painting модель (голубой), обученные модели на соответствующем наборе данных (зеленый), обученная на сильных пертурбациях из соответствующего набора данных (синий). Во всех обучающих экспериментах использовалась схема обучения-валидации 'leave-plates-out'. Источник рисунка [82]. 38

26 Качественное влияние пакетной коррекции на графиках UMAP. Левый график показывает UMAP-представление набора данных BBBC022 без пакетной коррекции, а правый - после пакетной коррекции. Точки представляют собой вложения профили уровня лунок (голубой - отрицательный контроль, красный - пертурбация). Графики плотности представлены в верхней и правой частях графиков. Признаки извлечены с помощью модели *Cell Painting CNN*. 39
27 Распределение типов анализов в окончательном наборе данных. Источник рисунка [97]. 41
28 Представления на уровне соединений в трех различных типах данных. Визуализация построена с помощью UMAP. A. Пространство признаков морфологии изначально было сгруппировано по технической вариации (группам планшетов), которая была скорректирована с помощью сферического преобразования. Цветовая палитра для 94 групп планшетов является непрерывной и может иметь схожие тона для групп планшетов. B. Представления в трех различных типах данных. C. Те же представления, что и в B, раскрашенные по кластерам, полученными в ходе экспериментов с кросс-валидацией (см. раздел 'Эксперименты и результаты'). Источник рисунка [97]. 42
29 Иллюстрация экспериментальной методики. Источник рисунка [97]. 43
30 A. Эффективность отдельных типов данных измеряется как количество анализов (вертикальная ось), предсказанных с AUROC выше определенного порога (горизонтальная ось). B. Диаграммы Венна показывают количество точно предсказанных анализов (медиана $AUROC > 0.9$ по результатам кросс-валидации), которые являются общими или уникальными для каждого типа данных. Гистограмма показывает распределение типов анализов, точно предсказанных отдельными типами данных. C. Количество точно предсказанных (медиана $AUROC > 0,9$ по результатам кросс-валидации) анализов по каждой отдельного типа данных. Источник рисунка [97]. 44
31 Количество точно предсказанных анализов (медиана AUROC по разделениям выше 0.9). A. Слева - диаграмма Венна точно предсказанных анализов с использованием позднего слияния, справа - гистограммы, показывающие распределение точно предсказанных типов анализов с поздним слиянием. B. Количество точно предсказанных анализов по отдельным типам данных. C. Количество точно предсказанных анализов для комбинированных методов с использованием позднего слияния. Подсчеты приведены для медианного и средней площади под ROC-кривой по разбиениям. D. Количество точно предсказанных анализов для ретроспективного анализа. 'Single' - простое объединение точно предсказанных анализов по отдельным типам данных. 'Plus fusion' - объединение точно предсказанных анализов с отдельными типами данных плюс предиктор позднего слияния. Источник рисунка [97]. 47

32 Количество предсказанных анализов с умеренной точностью (медианная площадь под ROC-кривой по расщеплениям выше 0.7). А. Слева диаграмма Венна предсказанных анализов с отдельными типами данных, в центре гистограмма предсказанных типов анализов по отдельным типам данных и в сочетании (позднее слияние), справа диаграмма Венна предсказанных анализов с объединенными типами данных (позднее слияние). В. Таблица эффективности отдельными типами данных и комбинированных (позднего слияния). Показатели: средняя площадь под ROC-кривой (Mean AUC) для 5 разбиений, среднее количество предсказанных анализов, пороговое значение площади под ROC-кривой ($AUC > 0.7$) для 5 разбиений. Источник рисунка [97]. 48

Список таблиц

1 Результаты экспериментов с кросс-валидацией (5 разбиений). В таблицах представлены средние результаты экспериментов для кросс-валидации в соответствии с различными подходами к разбиению данных. Метрики: средняя площадь под кривой 'точность-полнота' (Mean AUPRC) для 5 разбиений, средняя площадь под ROC-кривой (Mean AUROC) для 5 разбиений, среднее количество предсказанных анализов, пороговое значение площади под ROC-кривой ($AUC > 0.5$, $AUC > 0.7$, $AUC > 0.9$) для 5 разбиений. Источники используемых данных: MO: морфологические признаки без пакетной коррекции. MO-BC: морфологические признаки с коррекцией. GE: особенности экспрессии генов. CS-GC: графовые сверточные характеристики (GC). CS-MF: отпечатки Моргана. Среднее количество анализов в тестовом наборе отличается в разных типах данных, так как невозможно оценить анализ без 'попаданий' в тестовом наборе (которые отличаются, так как мы использовали разные подходы к разбиению на тренировочный и тестовые). Источник таблицы [97]. 45

2 Оценка отдельных и комбинированных типов данных для моделей, обученных с помощью разбиения на фрагменты. Показатели: средняя площадь под кривой 'точность-полнота' (Mean AUPRC) для 5 разбиений, средняя площадь под ROC-кривой (Mean AUROC) для 5 разбиений, среднее количество предсказанных анализов, пороговое значение площади под ROC-кривой ($AUC > 0.5$, $AUC > 0.7$, $AUC > 0.9$) для 5 разбиений. Также приведены стандартные отклонения. Источник таблицы [97]. 46

3 Количественно предсказанных анализов по типам с площадью под ROC-кривой (AUROC) более 0.9 (медианная по разбиениям). Источник таблицы [97]. 48

4 Количественно предсказанных анализов по типам с площадью под ROC-кривой (AUROC) более 0.7 (медианная по разбиениям). Источник таблицы [97]. 49

Список литературы

[1] Peter Horvath, Nathalie Aulner, Marc Bickle, Anthony M Davies, Elaine Del Nery, Daniel Ebner, Maria C Montoya, Päivi Östling, Vilja Pietiäinen, Leo S Price, Spencer L Shorte, Gerardo Turcatti, Carina von Schantz, and Neil O Carragher. Screening out irrelevant cell-based models of disease. *Nat. Rev. Drug Discov.*, 15(11):751–769, November 2016.

[2] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, 14(9):618–630, September 2013.

[3] Lukas Badertscher, Thomas Wild, Christian Montellese, Leila T Alexander, Lukas Bammert, Marie Sarazova, Michael Stebler, Gabor Csucs, Thomas U Mayer, Nicola Zamboni, Ivo Zemp, Peter Horvath, and Ulrike Kutay. Genome-wide RNAi screening identifies protein modules required for 40S subunit synthesis in human cells. *Cell Rep.*, 13(12):2879–2891, December 2015.

[4] Olaf Wolkenhauer, Peter Wellstead, Kwang-Hyun Cho, Dhanya Mullassery, Caroline A Horton, Christopher D Wood, and Michael R H White. Single live-cell imaging for systems biology 9. *Essays Biochem.*, 45:121–134, September 2008.

[5] Jeffrey M Levsky, Shailesh M Shenoy, Rossanna C Pezo, and Robert H Singer. Single-cell gene expression profiling. *Science*, 297(5582):836–840, August 2002.

[6] Nikolai Slavov. Single-cell protein analysis by mass spectrometry. *Curr. Opin. Chem. Biol.*, 60:1–9, February 2021.

[7] Reka Hollandi, Nikita Moshkov, Lassi Paavolainen, Ervin Tasnadi, Filippo Piccinini, and Peter Horvath. Nucleus segmentation: towards automated solutions. *Trends Cell Biol.*, January 2022.

[8] Ben T Grys, Dara S Lo, Nil Sahin, Oren Z Kraus, Quaid Morris, Charles Boone, and Brenda J Andrews. Machine learning and computer vision approaches for phenotypic profiling. *J. Cell Biol.*, 216(1):65–71, January 2017.

[9] Zachary E. Perlman, Michael D. Slack, Yan Feng, Timothy J. Mitchison, Lani F. Wu, and Steven J. Altschuler. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198, 2004.

[10] Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. January 2016.

[11] Juan C Caicedo, Allen Goodman, Kyle W Karhohs, Beth A Cimini, Jeanelle Ackerman, Marzieh Haghighi, Cherkeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E Carpenter. Nucleus segmentation across imaging experiments: the 2018 data science bowl. *Nat. Methods*, 16(12):1247–1253, December 2019.

[12] Réka Hollandi, Ákos Diósdi, Gábor Hollandi, Nikita Moshkov, and Péter Horváth. AnnotatorJ: an ImageJ plugin to ease hand annotation of cellular compartments. *Mol. Biol. Cell*, 31(20):2179–2186, September 2020.

[13] Johannes Schindelin, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, Curtis Rueden, Stephan Saalfeld, Benjamin Schmid, Jean-Yves Tinevez, Daniel James White, Volker Hartenstein, Kevin Eliceiri, Pavel Tomancak, and Albert Cardona. Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, 9(7):676–682, June 2012.

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for Large-Scale image recognition. September 2014.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[16] Mohammad H Rohban, Hamdah S Abbasi, Shantanu Singh, and Anne E Carpenter. Capturing single-cell heterogeneity via data fusion improves image-based profiling. *Nat. Commun.*, 10(1):2082, May 2019.

[17] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.*, 59(8):3370–3388, August 2019.

[18] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root,

Bang Wong, and Todd R Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017.

[19] Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100, October 2006.

[20] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Sci. Rep.*, 10(1):5068, March 2020.

[21] Aditya Pratapa, Michael Doron, and Juan C Caicedo. Image-based cell phenotyping with deep learning. *Curr. Opin. Chem. Biol.*, 65:9–17, December 2021.

[22] Erik Meijering. Cell segmentation: 50 years down the road [life sciences]. *IEEE Signal Process. Mag.*, 29(5):140–145, September 2012.

[23] Christoph Sommer, Christoph Straehle, Ullrich Köthe, and Fred A Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 230–233, March 2011.

[24] Pascal Bamford and Brian Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing*, 71(2):203–213, December 1998.

[25] Jozsef Moinar, Adam Istvan Szucs, Csaba Molnar, and Peter Horvath. Active contours for selective object segmentation. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, March 2016.

[26] Csaba Molnar, Ian H Jermyn, Zoltan Kato, Vesa Rahkama, Päivi Östling, Piia Mikkonen, Vilja Pietiäinen, and Peter Horvath. Accurate morphology preserving segmentation of overlapping cells based on active contours. *Sci. Rep.*, 6:32412, August 2016.

[27] Adrien Hallou, Hannah G Yevick, Bianca Dumitrascu, and Virginie Uhlmann. Deep learning for bioimage analysis in developmental biology. *Development*, 148(18), September 2021.

[28] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.*, 20(2):145–159, February 2021.

[29] Riddhiman Dhar, Alsu M Missarova, Ben Lehner, and Lucas B Carey. Single cell functional genomics reveals the importance of mitochondria in cell-to-cell phenotypic variation. *Elife*, 8, January 2019.

[30] Tomohiro Hayakawa, V B Surya Prasath, Hiroharu Kawanaka, Bruce J Aronow, and Shinji Tsuruoka. Computational nuclei segmentation methods in digital pathology: A survey. *Arch. Comput. Methods Eng.*, 28(1):1–13, January 2021.

[31] Nobuyuki Otsu. A threshold selection method from Gray-Level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66, January 1979.

[32] Stéfan van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and scikit-image contributors. scikit-image: image processing in python. *PeerJ*, 2:e453, June 2014.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 234–241. Springer International Publishing, Cham, 2015.

[34] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods*, 18(1):100–106, January 2021.

[35] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, Lecture notes in computer science, pages 265–273. Springer International Publishing, Cham, 2018.

[36] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017.

[37] Reka Hollandi, Abel Szkalisity, Timea Toth, Ervin Tasnadi, Csaba Molnar, Botond Mathe, Istvan Grexa, Jozsef Molnar, Arpad Balind, Mate Gorbe, Maria Kovacs, Ede Migh, Allen Goodman, Tamas Balassa, Krisztian Koos, Wenyu Wang, Juan Carlos Caicedo, Norbert Bara, Ferenc Kovacs, Lassi Paavolainen, Tivadar Danka, Andras Kriston, Anne Elizabeth Carpenter, Kevin Smith, and Peter Horvath. NucleAIzer: A parameter-free deep learning framework for nucleus segmentation using image style transfer. *Cell Syst.*, 10(5):453–458.e6, May 2020.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[40] Andrew Y Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Twenty-first international conference on Machine learning - ICML '04*, New York, New York, USA, 2004. ACM Press.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, June 2017.

[43] Ellen L Berg. The future of phenotypic drug discovery. *Cell Chem Biol*, 28(3):424–430, March 2021.

[44] David C Swinney and Jonathan A Lee. Recent advances in phenotypic drug discovery. *F1000Res.*, 9, August 2020.

[45] Jörg Eder, Richard Sedrani, and Christian Wiesmann. The discovery of first-in-class drugs: origins and evolution. *Nat. Rev. Drug Discov.*, 13(8):577–587, August 2014.

[46] Fabien Vincent, Arsenio Nueda, Jonathan Lee, Monica Schenone, Marco Prunotto, and Mark Mercola. Phenotypic drug discovery: recent successes, lessons learned and new directions. *Nature Reviews Drug Discovery*, may 2022.

[47] Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman A Carrel, Todd R Golub, Stuart L Schreiber, Paul A Clemons, Anne E Carpenter, and Alykhan F Shamji. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One*, 8(12):e80999, December 2013.

[48] Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G Linington, and Anne E Carpenter. Data-analysis strategies for image-based cell profiling. *Nat. Methods*, 14(9):849–863, August 2017.

[49] Joseph Boyd. *Deep learning for computational phenotyping in cell-based assays*. PhD thesis, Université Paris sciences et lettres, June 2020.

[50] Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of Single-Cell feature embeddings. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018:9309–9318, June 2018.

[51] Nick Pawlowski, Juan C Caicedo, Shantanu Singh, Anne E Carpenter, and Amos Storkey. Automating morphological profiling with generic deep convolutional networks. November 2016.

[52] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D. Boyd, and Anne E. Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, dec 2020.

[53] Mingyue Zheng, Jihui Zhao, Chen Cui, Zunyun Fu, Xutong Li, Xiaohong Liu, Xiaoyu Ding, Xiaoqin Tan, Fei Li, Xiaomin Luo, Kaixian Chen, and Hualiang Jiang. Computational chemical biology and drug design: Facilitating protein structure, function, and modulation studies. *Med. Res. Rev.*, 38(3):914–950, May 2018.

[54] Douglas B Kell, Soumitra Samanta, and Neil Swainston. Deep learning and generative methods in cheminformatics and chemical biology: navigating small molecule space intelligently. *Biochem. J*, 477(23):4559–4580, December 2020.

[55] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, February 1988.

[56] B Zagidullin, Z Wang, Y Guan, E Pitkänen, and J Tang. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Brief. Bioinform.*, 22(6), November 2021.

[57] H L Morgan. The generation of a unique machine description for chemical Structures-A technique developed at chemical abstracts service. *J. Chem. Doc.*, 5(2):107–113, May 1965.

[58] G W Bemis and M A Murcko. The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.*, 39(15):2887–2893, July 1996.

[59] Karren Yang, Samuel Goldman, Wengong Jin, Alex Lu, Regina Barzilay, Tommi Jaakkola, and Caroline Uhler. Improved conditional flow models for molecule to image synthesis. June 2020.

[60] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, Victoria M Tran, Anush Chiappino-Pepe, Ahmed H Badran, Ian W Andrews, Emma J Chory, George M Church, Eric D Brown, Tommi S Jaakkola, Regina Barzilay, and James J Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, February 2020.

[61] Kai Yao, Kaizhu Huang, Jie Sun, Linzhi Jing, Dejian Huang, and Curran Jude. Scaffold-A549: A benchmark 3D fluorescence image dataset for unsupervised nuclei segmentation. *Cognit. Comput.*, November 2021.

[62] Christoffer Edlund, Timothy R Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. LIVECell-A large-scale dataset for label-free live cell segmentation. *Nat. Methods*, 18(9):1038–1045, September 2021.

[63] Florin C Walter, Sebastian Damrich, and Fred A Hamprecht. Multistar: Instance segmentation of overlapping objects with star-convex polygons. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2021.

[64] Soham Mandal and Virginie Uhlmann. Splinedist: Automated cell segmentation with spline curves. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1082–1086, April 2021.

[65] Linfeng Yang, Rajarshi P Ghosh, J Matthew Franklin, Simon Chen, Chenyu You, Raja R Narayan, Marc L Melcher, and Jan T Liphardt. NuSeT: A deep learning tool for reliably separating and analyzing crowded cells. *PLoS Comput. Biol.*, 16(9):e1008193, September 2020.

[66] Ulysse Rubens, Romain Mormont, Lassi Paavolainen, Volker Bäcker, Benjamin Pavie, Leandro A Scholz, Gino Michiels, Martin Maška, Devrim Ünay, Graeme Ball, Renaud Hoyoux, Rémy Vandaele, Ofra Golani, Stefan G Stanciu, Natasa Sladoje, Perrine Paul-Gilloteaux, Raphaël Marée, and Sébastien Tosi. BIAFLOWS: A collaborative framework to reproducibly deploy and benchmark bioimage analysis workflows. *Patterns (N Y)*, 1(3):100040, June 2020.

[67] Ruchika Verma, Neeraj Kumar, Abhijeet Patil, Nikhil Cherian Kurian, Swapnil Rane, Simon Graham, Quoc Dang Vu, Mieke Zwager, Shan E Ahmed Raza, Nasir Rajpoot, Xiyi Wu, Huai Chen, Yijie Huang, Lisheng Wang, Hyun Jung, G Thomas Brown, Yanling Liu, Shuolin Liu, Seyed Alireza Fatemi Jahromi, Ali Asghar Khani, Ehsan Montahaei, Mahdieh Soleymani Baghshah, Hamid Behroozi, Pavel Semkin, Alexandr Rassadin, Prasad Dutande, Romil Lodaya, Ujjwal Baid, Bhakti Baheti, Sanjay Talbar, Amirreza Mahbod, Rupert Ecker, Isabella Ellinger, Zhipeng Luo, Bin Dong, Zhengyu Xu, Yuehan Yao, Shuai Lv, Ming Feng, Kele Xu, Hasib Zunair, Abdessamad Ben Hamza, Steven Smiley, Tang-Kai Yin, Qi-Rui Fang, Shikhar Srivastava, Dwarikanath Mahapatra, Lubomira Trnavska, Hanyun Zhang, Priya Lakshmi Narayanan, Justin Law, Yinyin Yuan, Abhiroop Tejomay, Aditya Mitkari, Dinesh Koka, Vikas Ramachandra, Lata Kini, and Amit Sethi. MoNuSAC2020: A multi-organ nuclei segmentation and classification challenge. *IEEE Trans. Med. Imaging*, pages 1–1, 2021.

[68] Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. March 2017.

[69] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. April 2018.

[70] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, 2019.

[71] Csilla Brasko, Kevin Smith, Csaba Molnar, Nora Farago, Lili Hegedus, Arpad Balind, Tamas Balassa, Abel Szkalisity, Farkas Sukosd, Katalin Kocsis, Balazs Balint, Lassi Paavolainen, Marton Z Enyedi, Istvan Nagy, Laszlo G Puskas, Lajos Haracska, Gabor Tamas, and Peter Horvath. Intelligent image-based in situ single-cell isolation, 2018.

[72] Peter D Caie, Rebecca E Walls, Alexandra Ingleston-Orme, Sandeep Daya, Tom Houslay, Rob Eagle, Mark E Roberts, and Neil O Carragher. High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Mol. Cancer Ther.*, 9(6):1913–1926, June 2010.

[73] Luis Pedro Coelho, Aabid Shariff, and Robert F Murphy. Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms, 2009.

[74] Kevin Smith, Yunpeng Li, Filippo Piccinini, Gabor Csucs, Csaba Balazs, Alessandro Bevilacqua, and Peter Horvath. CIDRE: an illumination-correction method for optical microscopy, 2015.

[75] Peter Naylor, Marick Lae, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map, 2019.

[76] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imaging*, 36(7):1550–1560, July 2017.

[77] Juan C. Caicedo, Jonathan Roth, Allen Goodman, Tim Becker, Kyle W. Karhohs, Matthieu Broisin, Csaba Molnar, Claire McQuin, Shantanu Singh, Fabian J. Theis, and Anne E. Carpenter. Evaluation of deep learning strategies for nucleus segmentation in fluorescence images. *Cytometry Part A*, 95(9):952–965, 2019.

[78] zhixuhao. zhixuhao/unet. `https://github.com/zhixuhao/unet`. Accessed: 2019-10-7.

[79] Tensorflow Developers. TensorFlow, 2021.

[80] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[81] matterport. matterport/Mask_RCNN. `https://github.com/matterport/Mask_RCNN`. Accessed: 2019-10-7.

[82] Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Claire McQuin, Matthew Smith, Allen Goodman, Rebecca Senft, Yu Han, Mehrtash Babadi, Peter Horvath, Beth A. Cimini, Anne E. Carpenter, Shantanu Singh, and Juan C Caicedo. Learning representations for image-based profiling of perturbations. *bioRxiv*, 2022.

[83] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688–701, October 1974.

[84] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In Maria Florina Balcan and Kilian Q Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA, 2016. PMLR.

[85] Mohammad Hossein Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S Boehm, and Anne E Carpenter. Systematic morphological profiling of human gene and allele function via cell painting. *Elife*, 6, March 2017.

[86] Mark-Anthony Bray, Sigrun M Gustafsdottir, Mohammad H Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L Sokolnicki, Joshua A Bittker, Nicole E Bodycombe, Vlado Dancík, Thomas P Hasaka, Cindy S Hon, Melissa M Kemp, Kejie Li, Deepika Walpita, Mathias J Wawer, Todd R Golub, Stuart L Schreiber, Paul A Clemons, Alykhan F Shamji, and Anne E Carpenter. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *Gigascience*, 6(12):1–5, December 2017.

[87] J C Caicedo, J Arevalo, F Piccioni, and others. Cell painting predicts impact of lung cancer variants. *Mol. Biol. Cell*, 2022.

[88] Gregory P Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C Caicedo, Beth A Cimini, Kyle Karhohs, David J Logan, Mohammad Rohban, Maria Kost-Alimova, Kate Hartland, Michael Bornholdt, Niranj Chandrasekaran, Marzieh Haghighi, Shantanu Singh, Aravind Subramanian, and Anne E Carpenter. Morphology and gene expression profiling provide complementary information for mapping cell state. October 2021.

[89] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking model scaling for convolutional neural networks. May 2019.

[90] Stanley Bryan Z Hua, Alex X Lu, and Alan M Moses. CytoImageNet: A large-scale pretraining dataset for bioimage transfer learning. November 2021.

[91] Vebjorn Ljosa, Peter D Caie, Rob Ter Horst, Katherine L Sokolnicki, Emma L Jenkins, Sandeep Daya, Mark E Roberts, Thouis R Jones, Shantanu Singh, Auguste Genovesio, Paul A Clemons, Neil O Carragher, and Anne E Carpenter. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.*, 18(10):1321–1329, December 2013.

[92] D Michael Ando, Cory Y McLean, and Marc Berndl. Improving phenotypic measurements in High-Content imaging screens. July 2017.

[93] Alexis Perakis, Ali Gorji, Samriddhi Jain, Krishna Chaitanya, Simone Rizza, and Ender Konukoglu. Contrastive learning of Single-Cell phenotypic representations for treatment classification. In *Machine Learning in Medical Imaging*, pages 565–575. Springer International Publishing, 2021.

[94] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *Am. Stat.*, 72(4):309–314, October 2018.

[95] Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing,, 2008.

[96] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, December 2018.

[97] Nikita Moshkov, Tim Becker, Kevin Yang, Peter Horvath, Vlado C Dancik, Bridget K Wagner, Paul C Clemons, Shantanu Singh, Anne E Carpenter, and Juan C Caicedo. Predicting compound activity from phenotypic profiles and chemical structures. December 2020.

[98] M Hofmarcher, E Rumetshofer, D A Clevert, and others. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of chemical*, 2019.

[99] Gregory P Way, Maria Kost-Alimova, Tsukasa Shibue, William F Harrington, Stanley Gill, Federica Piccioni, Tim Becker, William C Hahn, Anne E Carpenter, Francisca Vazquez, and Shantanu Singh. Predicting cell health phenotypes using image-based morphology profiling. July 2020.

[100] Jonathan B Baell and Georgina A Holloway. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, 53(7):2719–2740, April 2010.

[101] Mathias J Wawer, David E Jaramillo, Vlado Dančík, Daniel M Fass, Stephen J Haggarty, Alykhan F Shamji, Bridget K Wagner, Stuart L Schreiber, and Paul A Clemons. Automated Structure-Activity relationship mining: Connecting chemical structure to biological profiles. *J. Biomol. Screen.*, 19(5):738–748, June 2014.

[102] Mathias J Wawer, Kejie Li, Sigrun M Gustafsdottir, Vebjorn Ljosa, Nicole E Bodycombe, Melissa A Marton, Katherine L Sokolnicki, Mark-Anthony Bray, Melissa M Kemp, Ellen Winchester, Bradley Taylor, George B Grant, C Suk-Yee Hon, Jeremy R Duvall, J Anthony Wilson, Joshua A Bittker, Vlado Dančík, Rajiv Narayan, Aravind Subramanian, Wendy Winckler, Todd R Golub, Anne E Carpenter, Alykhan F Shamji, Stuart L Schreiber, and Paul A Clemons. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proceedings of the National Academy of Sciences*, 111(30):10911–10916, July 2014.

[103] Sebastian G Rohrer and Knut Baumann. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.*, 49(2):169–184, February 2009.

[104] jingw. GitHub - jingw2/size_constrained_clustering: Implementation of size constrained clustering algorithm. `https://github.com/jingw2/size_constrained_clustering`. Accessed: 2022-4-3.

[105] Jaak Simm, Günter Klambauer, Adam Arany, Marvin Steijaert, Jörg Kurt Wegner, Emmanuel Gustin, Vladimir Chupakhin, Yolanda T Chong, Jorge Vialard, Peter Buijnsters, Ingrid Velter, Alexander Vapirev, Shantanu Singh, Anne E Carpenter, Roel Wuyts, Sepp Hochreiter, Yves Moreau, and Hugo Ceulemans. Repurposing High-Throughput image assays enables biological activity prediction for drug discovery. *Cell Chem Biol*, 25(5):611–618.e3, May 2018.

[106] Maris Lapins and Ola Spjuth. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. July 2019.

[107] Malte Renz. Fluorescence microscopy—a historical and technical perspective. *Cytometry Part A*, 83(9):767–779, 2013.

[108] Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature Methods*, 16(12):1233–1246, may 2019.