О Т З Ы В

официального оппонента к.ф.-м.н. Д. А. Филимонова о диссертационной работе Евгения Игоревича Прохорова «Адаптивная двухфазная схема решения задачи «структура – свойство»», представленной на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 - теоретические основы информатики.

Актуальность темы.

Биологическая активность занимает особое место среди различных свойств химических соединений, поскольку создание новых высокоэффективных и более безопасных лекарственных средств является одним из приоритетных направлений науки и технологии. Однако, число различных структур химических соединений, зарегистрированных крупнейшим поставщиком химической информации CAS, в настоящее время свыше 60 млн, и изучаемых молекулярных мишеней лекарств более 2 тысяч, а с учетом альтернативного сплайсинга, посттрансляционных модификаций, белок-белковых взаимодействий и наличия различных сайтов связывания «лиганд-белок» число потенциальных фармакологических мишеней превосходит 2 млн. Поэтому экспериментальное тестирование десятков миллионов органических соединений на тысячи видов биологической активности невозможно и необходимо использование компьютерных методов для поиска и оптимизации новых фармакологически активных веществ. Среди компьютерных методов поиска и конструирования новых лекарств методы машинного обучения образуют одно из актуальных направлений. Несмотря на его более чем полувековую историю, оно продолжает активно развиваться, а сложность возникающих практических проблем обусловливает актуальность создания новых эффективных подходов.

Проблема применения методов машинного обучения для поиска зависимостей между структурой молекул химических соединений и их свойствами включает методы описания структур молекул и алгоритмы поиска зависимостей на основе анализа обучающих выборок ограниченного объема. Пространство структур молекул химических соединений дискретно и в нем не применимы понятия гладкости и даже непрерывности зависимостей, что создает особые проблемы интерполяции и экстраполяции свойств от одних соединений к другим. Именно это создает особые сложности данной предметной области и требует разработки адекватных проблеме подходов, особенно для поиска потенциально полезных соединений в информационных массивах из сотен тысяч, миллионов и даже миллиардов структур молекул химических соединений. Поэтому разработка метода прогнозирования свойств химических соединений с использованием

ограничений допустимости и адаптивного описания структур химических соединений в задаче поиска количественных корреляций «структура – свойство» весьма актуальна.

Степень обоснованности научных положений выводов и рекомендаций сформулированных в диссертации.

Диссертационная работа содержит ряд новых результатов, научная достоверность и оригинальность которых не вызывают сомнения. В работе автор последовательно и взвешенно использует различные методы теоретической информатики, теории машинного обучения и теории сложности.

Достоверность и новизна исследований и полученных результатов.

В работе получены следующие основные результаты.

Введено и формализовано понятие эффективных ограничений допустимости для математического моделирования «структура – свойство».

Предложены нетрадиционные подходы к оптимизации дескрипторного описания молекулярных графов для целей прогнозирования свойств неизученных химических соединений.

Разработана и обоснована новая методика построения и использования ограничений допустимости для моделей «структура – свойство», позволяющая получать более достоверный прогноз свойств молекулярных графов, а также ускорить процесс просмотра больших баз соединений.

Показаны возможности практического использования полученных результатов для прогнозирования свойств химических соединений.

Достоверность и обоснованность результатов, полученных автором, обеспечивается математическими формулировками и доказательствами, а также тестами и экспериментальными исследованиями. Теорема и утверждения диссертации доказаны. Найдена явная оценка показателя качества прогнозирования для моделей «структура – свойство», использующих предложенные автором ограничения допустимости. Приводятся оценки вычислительной сложности разработанных алгоритмов.

Значимость для науки и практики полученных автором результатов.

Виртуальный скрининг является одним из основных практических приложений моделей «структура — свойство», суть которого состоит в том, чтобы отыскать в большом информационном массиве структур химических соединений те, которые потенциально обладают целевой активностью или необходимым физико-химическим свойством.

На практике в выборке, содержащей сотни тысяч и миллионы структур химических соединений, находятся, для большинства биологических активностей, несколько десятков структур соединений, обладающих, в некоторой степени, целевой активностью. Задача скрининга заключается в том, чтобы эти соединения в базе идентифицировать.

Модели «структура — свойство» (QSAR-модели) позволяют вычислять прогноз активности соединений и отбирать структуры потенциально активных соединений. Успех такого отбора автор диссертации определяет через количество соединений, которые отобраны как активные и действительно проявили данную активность в результате экспериментальной проверки. Однако, что чем точнее прогноз, тем меньше будет предложено соединений, представляющих интерес для экспериментов.

При проведении массового виртуального скрининга ограничения допустимости играют роль фильтра, который позволяет сократить число ложноположительных прогнозов, что повышает качество отбора потенциальных соединений (для экспериментов будет предложено меньше неактивных структур, что сэкономит затраты на проведение экспериментальной проверки, и доля верных прогнозов, среди предложенных соединений возрастает). В тоже время, вместе с ложноположительными прогнозами фильтром могут быть отброшены также и реально активные соединения, ценность обнаружения которых высока. Поэтому использованию любых ограничений и фильтров сопряжено с риском потерять нужные соединения при поиске.

Понятие эффективности ограничений допустимости, формализованное автором, в целом позволяет говорить о том, что эффективные ограничения допустимости с большой вероятностью отсеивают ложные прогнозы (как ложноположительные, так и ложноотрицательные), и с малой вероятностью отказываются от прогнозов верных. Также эффективность предложенных ограничений допустимости в смысле повышения показателя качества моделей не только теоретически доказана, но и продемонстрирована на небольших выборках реальных соединений.

Однако для практических применений данного подхода имеет значение также и то, насколько эффективно построенные модели будут обнаруживать активные соединения в выборках большого и сверхбольшого объема, а не только в обучающих и контрольных выборках, где количество активных и неактивных соединений примерно одинаково.

Конкретные рекомендации по использованию результатов и выводов диссертации.

Построенные автором модели и полученные теоретические результаты могут быть использованы в практических задачах прогнозирования биологической активности,

которые возникают при конструировании лекарственных средств и во многих других не менее актуальных исследованиях.

Оценка содержания диссертации и ее завершенности.

Работа носит завершенный характер. Подход автора получил свое выражение в формальных понятиях и утверждениях. Получены точные оценки показателей качества моделей в предложенной автором терминологии.

Диссертационная работа построена по традиционной схеме и состоит из введения, трех глав, заключения и списка литературы.

Во введении и первой главе автор дает достаточно полное изложение представлений о QSAR-моделировании и связанных процессах, освещает методы виртуального скрининга, дает обзор подхода к решению задачи на базе фрагментных дескрипторов особых точек. Автор также отмечает особенности задачи и недостатки рассматриваемого подхода и трудности, связанные с его применением, формализует ряд понятий, на базе которых формулируется теоретическая часть полученных результатов.

Вторая глава содержит изложение теоретических результатов, в том числе методов построения моделей «структура – свойство» и ограничений допустимости для них. Глава содержит математические доказательства ряда оценок на качество моделирования, а также оценки вычислительной сложности рассматриваемых алгоритмов.

В третьей главе автор приводит описание исследований, проведенных совместно со специалистами-химиками и отражающих практическую часть результатов работы.

Диссертация читается легко, написана доступным языком и отличается ясностью формулировок. Текст хорошо иллюстрирован схемами, таблицами и рисунками.

Недостатки и замечания.

По тексту диссертации имеются следующие замечания.

1. Более полно говорить о практической значимости работы позволили бы результаты тестирования построенных моделей на больших базах соединений, где число активных соединений чрезвычайно мало по сравнению с общим числом соединений.

2. Показатель качества моделей «структура – свойство», рассмотренный автором, безусловно, заслуживает внимания, однако, интерес представляют также и другие характеристики моделей, в том числе измеренные в массовых тестах.

3. Изложение материала диссертации имеет ряд редакционных недостатков. Так выбранная терминологии не всегда удачна, например, понятие «уровней» встречается как в описании процесса построения дескрипторов молекулярных графов, так и в изложении

двухфазной схемы для различения классифицирующих моделей, осуществляющих прогнозирования и моделей, реализующих ограничения допустимости. Есть и другие мелкие недочеты, которые в той или иной степени могут затруднить восприятие полученных результатов.

Выводы.

Сделанные замечания не снижают общего высокого уровня работы. Диссертация, в целом, представляет законченное научное исследование в соответствии с поставленными автором целями и задачами и имеет как научное, так и практическое значение для теоретических основ информатики и ее приложений. Результаты диссертации являются новыми, работа снабжена математическими доказательствами, результаты получены автором самостоятельно. Автореферат и публикации полностью отражают содержание диссертационной работы. По актуальности, объему и научному уровню выполненных исследований она полностью удовлетворяет всем требованиям Минобрнауки РФ, предъявляемым к диссертациям на соискание ученой степени кандидата наук, а её автор Е. И. Прохоров заслуживает присвоения искомой степени кандидата физико-математических наук по специальности 05.13.17 - теоретические основы информатики.

Официальный оппонент,
кандидат физико-математических наук, ведущий научный сотрудник
Федерального государственного бюджетного учреждения
«Научно исследовательский институт биомедицинской химии имени В. Н. Ореховича»
Российской академии медицинских наук
119121, Москва, Погодинская ул., 10/8
Тел. 2-499-2553029; E-mail: dmitry.filimonov@ibmc.msk.ru

Д. А. Филимонов

Подпись ведущего научного сотрудника ФБГУ «ИБМХ» РАМН,
кандидата физико-математических наук Д. А. Филимонова заверяю.
Ученый секретарь,
к.х.н.

Е. А. Карпова