

О Т З Ы В

официального оппонента д.ф.-м.н. Баскина И. И. о диссертационной работе Ромеро Рейес Илякай Владиславовны «Оценка аффинности комплексов белок-лиганд с применением нейронных сетей», представленной на соискание ученой степени кандидата физико-математических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ.

Актуальность темы. Представленная работа находится на пересечении таких научных направлений как химическая информатика, математическое моделирование, биофизика и фармакология. В ней предлагается подход для решения одной из ключевых проблем, встающих при разработке новых лекарственных препаратов – прогнозирования сродства органического лиганда к биологической мишени. Ее решение дает возможность оптимизировать структуру молекул разрабатываемых органических соединений под определенную мишень, связанную с лечением определенных заболеваний.

Предлагаемый подход основан на использовании аппарата искусственных нейронных сетей, сама структура которых допускает высокую степень параллелизма вычислений. Применение для этого специализированных плат-нейроускорителей для эмуляции нейронных сетей крайне невыгодно по экономическим соображениям ввиду высокой стоимости и поэтому малой распространенности последних. Кроме того, само построение моделей с помощью методов машинного обучения, а также оценка их прогнозирующей способности, требует весьма интенсивных вычислений, которые могут быть эффективно распараллелены. В связи с этим очень перспективным является использование очень распространенных, производительных и относительно дешевых графических процессоров с технологией CUDA. Поэтому задача разработки на основе этой технологии высокопараллельных эмуляторов нейронных сетей и основанных на них программных средств также весьма актуальна.

Научная новизна исследования. Прежде всего, несомненной научной новизной обладает разработанный в диссертационной работе подход к прогнозированию сродства органических лигандов к белкам, основанный на сочетании методов химической информатики, предполагающих построение количественных моделей «структура-свойство» с помощью методов машинного обучения с использованием базы данных по измеренным в эксперименте

значениям сродства набора органических лигандов к белку, с методами молекулярного моделирования (с использованием основанных на эмпирических силовых полях методов молекулярной механики и динамики) комплексов лигандов с сайтом связывания биологической макромолекулы. В диссертации найдено удачное сочетание (1) набора физико-химических дескрипторов (т.е. признаков), описывающих структуры химических соединений, (2) набора физических характеристик взаимодействия лиганд-белок, оцениваемых с помощью методов ММ/PBSA и ММ/GBSA по результатам моделирования с помощью молекулярной динамики, (3) искусственных нейронных сетей в качестве метода машинного обучения для построения регрессионных моделей «структура-свойство», и (4) методов снижения размерности пространства молекулярных дескрипторов, что привело к построению моделей с высокой прогнозирующей способностью.

Хотя общая идея сочетать линейные и нелинейные методы снижения размерности пространства признаков с регрессионными и классификационными методами машинного обучения не нова, и к ней, в сущности, сводится находящаяся в центре внимания в последние годы задача «обучения признакам» (feature learning), однако в области химической информатики таких работ очень мало. До сих пор при обработке химических баз данных нелинейные методы снижения размерности применялись почти исключительно с целью визуализации химического пространства, и лишь в последние два года некоторые из них начали использоваться при построении классификационных моделей. Именно поэтому проведенное в диссертационной работе систематическое исследование возможности сочетания при построении моделей «структура-свойство» искусственных нейронных сетей с разнообразными методами снижения размерности пространства дескрипторов обладает высокой степенью научной новизны и представляет несомненный интерес для химической информатики.

В рассматриваемой работе впервые метод вложения для новых точек был применен для количественной оценки сродства лиганда к белку. Ранее при обработке химической информации для осуществления вложения для новых точек применялся лишь глобально-нелинейный подход, основанный на дополнительной нейронной сети. Используемый в диссертационной работе локально-линейный подход является в вычислительном плане значительно более эффективным.

Обоснованность и достоверность научных положений и выводов.

Основные научные положения и выводы, сделанные в диссертационной работе, представляются вполне обоснованными. Высокая прогнозирующая способность построенных моделей подтверждена как с помощью процедуры перекрестного контроля, так и с помощью прогноза на внешней независимой выборке. Особенно следует отметить то, что результаты прогноза по построенным моделям прошли экспериментальную проверку, подтвердившую их правильность. Очень высокое ускорение, достигнутое благодаря использованию графических процессоров, подтверждает правильность выбранного способа распараллеливания вычислений.

Практическая значимость полученных результатов. Разработанный в диссертационной работе новый метод прогнозирования сродства лигандов к биологическим рецепторам и реализующее их программное обеспечение, а также построенные с их помощью модели по связыванию с двумя типами ядерных рецепторов стероидных гормонов, могут быть использованы в исследованиях, направленных на создание новых лекарственных препаратов.

Оценка языка и стиля изложения работы. Диссертационная работа читается легко, поскольку написана доступным языком и отличается ясностью формулировок. Текст достаточно иллюстрирован схемами, таблицами и рисунками.

Недостатки и замечания. Представленная работа не лишена недостатков, которые, однако, не являются принципиальными.

1. Процесс формирования хэш-кодов для химических структур на стр. 57-58 диссертации описан не полностью корректно.

2. Непонятно, почему в работе не проводилось сравнение с моделями, получаемыми с использованием исходного набора дескрипторов без применения методов снижения размерности.

3. В диссертационной работе не описаны детали реализации процедуры с исключением по одному LOO в случае нейронных сетей, для которых декларирован трехвыборочный подход к обучению.

4. В ряде случаев для сравнения качества моделей используется коэффициент детерминации R^2 , вычисляемый на обучающей выборке, что в данном случае никак не характеризует прогнозирующую способность моделей ввиду склонности нейронных сетей к переобучению.

5. В работе не объяснено, какие преимущества могут дать нелинейные методы снижения размерности по сравнению с линейными, если, как следует из представленных таблиц, они приводят только к ухудшению моделей.

6. Непонятно, почему при варьировании числа нейронов в скрытом слое не рассматривались сети с числом скрытых нейронов меньшим, чем число входных нейронов минус три. Таким путем можно пропустить оптимальные модели, основанные на нейронных сетях с минимальным размером скрытого слоя.

Выводы. Сделанные замечания не могут умалить достоинства предложенной работы. В целом, диссертация представляет законченное исследование в рамках выдвинутых целей и имеет как научное, так и практическое значение. Автореферат и публикации полностью отражают содержание диссертационной работы. По актуальности, объёму и научному уровню она полностью соответствует требованиям Положения о присуждении ученых степеней, предъявляемым к кандидатским диссертациям, а ее автор – Ромеро Рейес И. В. заслуживает присуждения искомой степени кандидата физико-математических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ.

Официальный оппонент,
доктор физико-математических наук,
ведущий научный сотрудник
физического факультета Московского
государственного университета
им. М.В.Ломоносова

И. И. Баскин

12.05.2014

Декан физического факультета Московского
государственного университета
им. М.В.Ломоносова,
д.ф.-м.н., профессор



Н. Н. Сысоев