ЖУРНАЛ ФИЗИЧЕСКОЙ ХИМИИ, 2023, том 97, № 2, с. 262–269

ХЕМОИНФОРМАТИКА И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 544-971.62,54-32

ВЛИЯНИЕ РАЗЛИЧНЫХ ФАКТОРОВ НА ПРЕДСКАЗАНИЕ КОНСТАНТ КИСЛОТНОСТИ НИЗКОМОЛЕКУЛЯРНЫХ ОРГАНИЧЕСКИХ СОЕДИНЕНИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

© 2023 г. Д. Д. Матюшин[a], А. Ю. Шолохова[a,*], А. К. Буряк[a]

[a] *Институт физической химии и электрохимии им. А.Н. Фрумкина РАН, 119071 Москва, Россия*

**e-mail: shonastya@yandex.ru*

Поступила в редакцию 11.08.2022 г.
После доработки 11.08.2022 г.
Принята к публикации 25.08.2022 г.

Изучено влияние способа стандартизации структуры молекулы и параметров расчета молекулярных отпечатков пальцев на точность предсказания константы кислотности. Показано, что стандартизация, т.е. выбор таутомерной формы и способа записи структуры молекулы, с помощью OpenEye QuacPac дает наилучшие результаты, однако библиотека RDKit позволяет достигнуть сравнимой точности. Установлено, что способ выбора зарядового состояния оказывает большое влияние на точность предсказания. Исследована точность предсказания в зависимости от радиуса (размера подструктур) круговых молекулярных отпечатков пальцев, лучшие результаты достигаются при использовании радиуса $r = 2$. Использован случайный лес – один из алгоритмов машинного обучения. Кроме того, показано, что метод опорных векторов также дает достаточно высокую точность при оптимизации гиперпараметров.

Ключевые слова: константа кислотности, машинное обучение, таутомеры

DOI: 10.31857/S0044453723020152, **EDN:** ECXZUI

Константа кислотности (pK_a) – важная физико-химическая характеристика молекулы. Измерение константы кислотности – трудоемкая задача, и общедоступные базы данных содержат экспериментальные данные о pK_a для нескольких тысяч молекул [1–3]. Собственные внутренние базы данных, используемые фармацевтическими компаниями, и коммерческие базы данных с ограниченным доступом содержат в несколько раз больше данных [4, 5]. Значение pK_a для подавляющего большинства органических молекул неизвестно. Эта величина – важное свойство молекулы при использовании в медицине и при изучении биофармацевтических свойств. Быстрое предсказание значений pK_a *in silico* необходимо для разработки лекарств и виртуального скрининга. Виртуальный скрининг, т.е. предсказание физико-химических и биофармацевтических свойств для большого числа молекул с целью отбора перспективных кандидатов при разработке лекарств, приобрел большое значение в последние годы [6, 7].

Предсказание констант кислотности с помощью машинного обучения, может быть применено в жидкостной хроматографии. За последние годы разработаны точные способы предсказания времен удерживания в жидкостной хроматографии по структуре соединения с помощью глубокого обучения [8, 9]. Кислотно-основная диссоциация оказывает сильнейшее влияние на хроматографическое удерживание [10, 11], и предсказание констант кислотности может быть использовано и при предсказании времен удерживания.

Машинное обучение произвело революцию во многих областях науки и техники [12]. В частности, в химии машинное обучение применяется для предсказания свойств химических веществ по их структуре [1, 2, 13], в химическом анализе [8, 9], а также в других областях. Один из важных методов машинного обучения – случайный лес, который представляет собой ансамбль деревьев решений, каждое из которых делает не слишком точное предсказание искомой величины, однако усреднение большого количества предсказаний позволяет уменьшить погрешность и спрогнозировать величину достаточно точно [14].

Один из основных способов представить молекулу в виде вектора для последующего применения метода машинного обучения – так называемые молекулярные отпечатки пальцев (molecular fingerprints, MF). Обычно MF представляют собой вектор, состоящий из единиц и нулей или из целых чисел, длиной от нескольких десятков до нескольких тысяч значений [4, 15, 16]. Каждое

число обычно характеризует наличие того или иного фрагмента (субструктуры) в молекуле или (в случае целочисленных, называемых также аддитивными, MF) количество вхождений этого фрагмента. Существуют разные алгоритмы генерации MF. Один из наиболее важных и распространенных алгоритмов – так называемые круговые молекулярные отпечатки пальцев [16]. Существует два варианта этого алгоритма, обозначаемые ECFP и FCFP [16]. Различие между этими двумя алгоритмами объяснено в работе [16].

Предсказание pK_a с помощью машинного обучения активно изучается последние несколько лет [1–5, 17, 18]. Многие коммерческие пакеты программного обеспечения включают в себя возможность предсказания pK_a: BioByte, ACD/Labs, Simulations Plus, ChemAxon Marvin, Epik. Есть и программное обеспечение с открытым исходным кодом для предсказания pK_a [1–3]. Многие модели pK_a ограничены отдельными классами соединений, например, только первичными аминами [4]. Помимо машинного обучения для предсказания pK_a используются и квантово-химические методы [19–21], однако, требуя значительно больше вычислительных ресурсов, эти методы не сильно превосходят по точности методы, основанные на машинном обучении.

В последние годы появился ряд работ, использующих графовые нейронные сети (graph neural networks) для предсказания pK_a [3, 17, 18]. Графовая нейронная сеть способна предсказывать как свойства молекулы в целом, так и свойства отдельных атомов в молекуле, и, как следствие, способна предсказывать и микроскопические, и макроскопические константы кислотности. Такие модели используются в программном обеспечении MolGpka [18], pkasolver [3], Graph-pKa [17]. Следует отметить, что в ряде случаев, при применении трансферного обучения, такие работы используют для обучения не только публично доступные базы данных pK_a, но и коммерческое программное обеспечение, и, как следствие, неявно используют базы данных, использованные при разработке соответствующих моделей. Так, например, в работе [3] для предварительного обучения графовой нейронной сети использовалось программное обеспечение Epik (предсказанные с помощью него значения pK_a), при этом, возможно, наборы данных, использованные для тестирования, пересекались с наборами данных, использованными авторами Epik. Сравнение таких моделей с моделями, обученными по публично доступному набору данных и не использующими трансферное обучение, с использованием одних и тех же тестовых наборов не вполне корректно.

Несмотря на определенный прогресс, достигнутый при применении графовых нейронных сетей, изучение методов предсказания pK_a с помощью традиционных методов машинного обучения по-прежнему актуально. В последние годы появилось несколько работ [1, 2, 4], посвященных сравнению методов машинного обучения для предсказания значений pK_a и изучению различных факторов, влияющих на точность, таких как тип MF. Однако некоторые факторы, сильно влияющие на точность предсказания, не рассматриваются в указанных работах, и их учет может повлиять на сделанные выводы. Один из них – предварительная стандартизация структур, выбор формы молекулы, которая будет использована при расчете MF. Другой важный фактор – выбор алгоритма для расчета круговых молекулярных отпечатков пальцев – ECFP или FCFP, а также радиуса (размера) рассматриваемых субструктур. Кроме того, сравнение методов машинного обучения зачастую проводится без описания процедуры подбора значений гиперпараметров.

Цель данной работы – изучить влияние способа стандартизации структур (выбора таутомерной формы и зарядового состояния) и параметров расчета MF на точность предсказания pK_a с помощью случайного леса, а также выяснить, возможно ли с помощью метода опорных векторов достигнуть такой же точности предсказания, как и с помощью случайного леса.

МЕТОДЫ ИССЛЕДОВАНИЯ

Использованное программное обеспечение

Влияние различных факторов на точность предсказания pK_a с помощью случайного леса исследовали с помощью языка программирования Python (версия 3.10.0), пакета scikit-learn [22] (версия 1.1.1), с использованием библиотеки RDKit [23] (версия 2021.9.4). Для расчета MF использовали функции AllChem.GetMorganFingerprintAsBitVect и AllChem.GetHashedMorganFingerprint из библиотеки RDKit (бинарные и аддитивные круговые MF соответственно). В качестве метрик точности использовали среднюю абсолютную ошибку и среднеквадратичную ошибку (в единицах pK_a).

Для изучения возможности применения метода опорных векторов использовали библиотеку LIBSVM [24] (версия 3.25 для языка Java). Для генерации молекулярных дескрипторов наряду с RDKit также использовали библиотеку CDK [25] (версия 2.7.1). Подбор гиперпараметров выполняли с помощью собственного программного обеспечения на языке Java.

Наборы данных и стандартизация структур

В целях сравнения использовали в точности такие же наборы данных, как и в работе [1]. Наборы данных загружали из сети Интернет [26] в

формате SDF. Рассматривали три набора данных, обозначенных как Main, Novartis и Literature. Эти наборы данных содержали информацию о константах кислотности (pK_a) для 5994, 280 и 123 молекул соответственно. Наборы данных содержали только молекулы с одной функциональной группой, подверженной кислотно-основной диссоциации, с pK_a в диапазоне 2–12. Многоосновные кислоты и основания исключали из рассмотрения. Все структуры приведены авторами работы [1] к зарядовому состоянию, доминирующему при pH 7.4, с помощью коммерческого программного обеспечения OpenEye QuacPac, и для всех молекул выбрана оптимальная таутомерная форма с помощью этого же программного обеспечения. И для кислых, и для основных соединений рассматривалась константа кислотности, соответствующая кислотно-основному переходу при pH в диапазоне 2–12. Для основных соединений, таким образом, рассматривалась величина pK_a протонированной формы. Кислые и основные соединения рассматривались вместе.

Набор данных Main использовался для обучения, также для него определялась точность предсказания с помощью кросс-валидации (так же, как в работе [1]). Использовалась пятикратная (5-fold) кросс-валидация (перекрестная проверка) со случайным разбитием набора данных. Наборы данных Novartis и Literature использовались в качестве тестовых, при этом использовалась модель, обученная с помощью всего набора Main. Для оценки воспроизводимости каждый раз перед обучением или кросс-валидации из набора Main случайным образом удалялась информация об 1% молекул.

В данной работе мы исследовали влияние зарядового состояния и таутомерной формы на точность предсказания. Наборы данных использовались или в неизменном виде, стандартизированные с помощью OpenEye QuacPac, или стандартизация средствами RDKit. Рассматривались два варианта стандартизации таутомеров: с помощью конвертации в строку InChI [27] и последующей обратной конвертации, а также с помощью функции MolStandardize.canonicalize_tautomer_smiles из библиотеки RDKit. Эти варианты стандартизации обозначены ниже как InChI и RDKit соответственно. Рассматривались также комбинации указанных методов. Каждый из рассмотренных методов приводит к одному и тому же результату, вне зависимости от того, какая таутомерная форма исходно была использована. В то же время результаты работы методов стандартизации различаются между собой. Для исследования влияния зарядового состояния структур также использовалась нейтрализация всех заряженных групп в молекуле с помощью класса rdMolStandardize.Uncharger из библиотеки RDKit перед дальнейшим применением методов стандартизации.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Влияние различных методов стандартизации молекул на точность предсказания

Структура молекулы может быть представлена в виде различных таутомерных форм, например, виниловый спирт при нормальных условиях в водном растворе находится в равновесии с ацетальдегидом (равновесие смещено в сторону ацетальдегида). Для других кетонов, например, для ацетилацетона, в равновесном растворе енольная форма присутствует в значительных концентрациях. С точки зрения предсказания макроскопической pK_a, эти формулы эквивалентны, и результат предсказания для кетонной и для енольной форм должен быть одинаковым, так как в водном растворе обе формы находятся в равновесии и представляют одно и то же химическое вещество. Однако MF для кетонной и для енольной форм будут различаться и, с точки зрения машинного обучения, это различные соединения. Большие молекулы могут иметь множество таутомерных форм, соотношение между равновесными концентрациями которых неизвестно.

Модель машинного обучения представляет собой эмпирическое выражение, связывающее входное представление молекулы (например, MF) с искомой величиной (например, pK_a). Простого и эффективного способа учесть все таутомерные формы во входном представлении молекулы не существует, поэтому в наборах данных используется одна таутомерная форма для каждой молекулы. Более того, значения MF могут отличаться и для разных форм записи структурной формулы, представляющих собой одну и ту же молекулу. Поэтому требуется так называемая стандартизация молекулы: выбор таутомерной формы и формы записи структуры с помощью определенного алгоритма. Такой алгоритм, получая на вход любую из таутомерных форм, приводит молекулу к определенному виду. При этом могут использоваться эвристики или эмпирические формулы для выбора наиболее правдоподобной формы.

В работе [1] стандартизация выполнялась при помощи коммерческого программного обеспечения OpenEye QuacPac, использующего достаточно сложный алгоритм. В других случаях, например, в работе [4] способ стандартизации не указан. Стандартизация структур представляется одним из ключевых шагов [1], влияние которого не было ранее изучено. В данной работе использовались пять вариантов стандартизации, описанных выше. В том случае, когда никакой дополнительной стандартизации не проводилось,

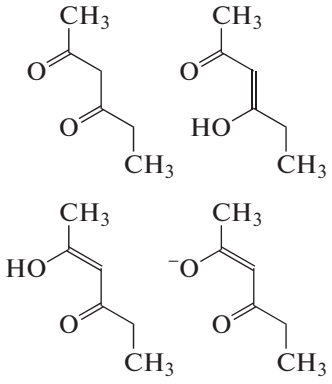

Рис. 1. Различные структурные формулы, описывающие одно и то же (с точки зрения предсказания макроскопической pK_a) химическое вещество, находящееся в водном растворе.

фактически речь идет о структурах, стандартизированных с помощью OpenEye QuacPac, взятых из работы [1].

Помимо выбора таутомерной формы отдельную проблему представляет выбор зарядового состояния. В работе [1] выбор осуществлялся с помощью программного обеспечения OpenEye QuacPac при pH 7.4. Рассматриваемые наборы данных содержат как pK_a кислот, так и pK_a протонированных форм органических оснований, и эти величины предсказываются одной и той же моделью, хотя фактически речь идет о различных задачах. С точки зрения кислотно-основного равновесия, для молекул с одной группой, подверженной диссоциации, протонированная и депротонированная формы представляют собой одну и ту же молекулу, в том смысле, что этой паре соответствует одна константа кислотности pK_a. Тем не менее, как и в случае с таутомерными формами, иону и нейтральной молекуле соответствуют различные MF. Таким образом, приведение всех структур к зарядовой форме при фиксированном pH имеет большое значение. Для изучения важности этого эффекта осуществлялось приведение всех молекул к нейтральной форме перед стандартизацией и вычислением MF.

Во всех случаях (нейтрализация, стандартизация) одинаковые преобразования проводились со всеми молекулами, во всех наборах данных. На рис. 1 показан пример различных форм структуры молекулы, имеющих различные MF, однако на самом деле представляющих одну и ту же молекулу, которой должно соответствовать одно значение pK_a.

В табл. 1 показаны метрики точности для различных способов преобразования структур. Отсутствие преобразования означает структуры, взятые из работы [1], стандартизированные с помощью программного обеспечения OpenEye QuacPac. Использовались бинарные молекулярные отпечатки пальцев FCFP, длиной 4096 бит, радиус (размер субструктур) был равен 3. В качестве метода машинного обучения использован случайный лес, имплементация из пакета scikit-learn, использовано 1000 деревьев, остальные значения гиперпараметров взяты по умолчанию. Доверительный интервал рассчитан для трех величин по результатам трех расчетов. Точность предсказания с помощью машинного обучения представляет собой случайную величину вследствие стохастической природы использованного алгоритма, случайного разбития набора данных на пять подмножеств для кросс-валидации и случайного исключения 1% молекул.

Из табл. 1 видно, что влияние метода стандартизации значимо (при тестировании с использованием наборов данных Novartis и Literature), и стандартизация с помощью OpenEye QuacPac дает наилучшие результаты. Также видно, что результаты стандартизации с помощью комбинаций методов InChI-RDKit и RDKit-InChI отличаются от полученных с помощью RDKit и InChI, примененных по отдельности. Это связано с тем, что разные методы в некоторых случаях по-разному рассматривают структуры как таутомерные формы одной молекулы или как разные молекулы. Следует отметить, что различие при использовании разных методов стандартизации не слишком велико, и программное обеспечение с открытым исходным кодом может быть использовано для этой цели.

Кроме того, был рассмотрен следующий подход. С помощью библиотеки RDKit для каждой молекулы были сгенерированы все возможные таутомеры. При обучении каждый таутомер (соответствующие ему MF) рассматривался как отдельное соединение. Одно и то же значение pK_a рассматривалось для каждого из таутомеров. При тестировании с использованием внешних тестовых наборов Novartis и Literature результат предсказания для всех таутомеров усреднялся. Для наборов Novartis и Literature значение среднеквадратичной ошибки составило 1.66 и 0.92 единиц pK_a соответственно. Средняя абсолютная ошибка составила 1.27 и 0.60 соответственно.

Данные табл. 1 показывают, что использование зарядового состояния, доминирующего при pH 7.4, сгенерированного с помощью OpenEye QuacPac, приводит к существенно лучшим результатам по сравнению с нейтральной формой. Однако генерация такой формы включает в себя неявным образом оценку кислотности функциональной группы. Таким образом сравнение алгоритма, обученного с помощью публично доступного набора данных и только открытого программного обеспечения, с другим алгоритмом, включающим в себя приведение структуры к за-

Таблица 1. Точность предсказания pK_a с помощью случайного леса (1000 деревьев) и бинарных молекулярных отпечатков пальцев FCFP с радиусом 3 и длиной 4096 бит, при использовании различных методов стандартизации структуры

Стандартизация	Среднеквадратичная ошибка			Средняя абсолютная ошибка		
	Main, кросс-валидация	Novartis	Literature	Main, кросс-валидация	Novartis	Literature
Нет (OpenEye QuacPac)	1.10 (0.02)	1.53 (0.02)	0.78 (0.01)	0.72 (0.01)	1.15 (0.01)	0.52 (0.01)
InChI	1.13 (0.01)	1.59 (0.01)	0.86 (0.02)	0.74 (0.01)	1.20 (0.01)	0.59 (0.01)
RDKit	1.13 (0.01)	1.58 (0.01)	0.95 (0.03)	0.73 (0.00)	1.17 (0.01)	0.62 (0.01)
InChI, затем RDKit	1.15 (0.01)	1.66 (0.01)	0.91 (0.01)	0.74 (0.00)	1.24 (0.01)	0.62 (0.01)
RDKit, затем InChI	1.14 (0.03)	1.59 (0.01)	0.90 (0.01)	0.74 (0.02)	1.19 (0.00)	0.62 (0.01)
Нейтрализация	1.20 (0.02)	1.72 (0.03)	1.06 (0.03)	0.78 (0.01)	1.30 (0.03)	0.67 (0.01)
Нейтрализация, затем InChI	1.22 (0.01)	1.78 (0.01)	1.04 (0.01)	0.80 (0.01)	1.37 (0.01)	0.68 (0.01)
Нейтрализация, затем RDKit	1.22 (0.00)	1.79 (0.02)	1.14 (0.02)	0.80 (0.01)	1.34 (0.01)	0.76 (0.01)

Примечание. Значения в скобках представляют собой половину длины двустороннего доверительного интервала, рассчитанной с помощью критерия Стьюдента при $p = 0.95$.

рядовой форме, доминирующей при pH 7.4, с помощью коммерческого программного обеспечения, не вполне корректно. Возможным способом достижения лучшей точности без использования OpenEye QuacPac или другого программного обеспечения для выбора зарядового состояния служит рассмотрение кислых и основных соединений по отдельности, с разбитием набора данных на наборы кислых и основных молекул [2]. Однако при этом уменьшается общий размер набора данных, что негативно влияет на точность предсказания.

Влияние параметров расчета молекулярных отпечатков пальцев на точность предсказания

Существуют два варианта алгоритма расчета круговых молекулярных отпечатков пальцев: ECFP и FCFP, также оба алгоритма могут работать в варианте аддитивных MF (вектор целых чисел) и бинарных MF (вектор нулей и единиц) и имеют два параметра: длину вектора l (количество битов или целых чисел) и так называемый радиус r, характеризующий размер рассматриваемых фрагментов. Работа [4] посвящена в том числе сравнению различных видов MF для предсказания pK_a. При этом используется $r = 7$ и алгоритм ECFP, и делается вывод, что круговые MF обладают меньшей эффективностью по сравнению с другими видами MF. Однако такое сравнение может быть не вполне корректным, если радиус r и вариант алгоритма выбраны неоптимально. В работе [1] используется алгоритм FCFP, $r = 3$.

На рис. 2 показана зависимость средней абсолютной ошибки D предсказания pK_a (набор Main, кросс-валидация) от r для четырех вариантов алгоритма и различных значений l. Видно, что точности, достигнутые с помощью бинарных и аддитивных MF, близки. Вероятно, это связано с тем, что рассматриваются только молекулы с одной функциональной группой, подверженной диссоциации, и на значение pK_a влияет преимущественно ее непосредственное окружение. Однако аддитивные MF несут в себе больше информации и позволяют достигнуть несколько лучшую точность по сравнению с бинарными MF. Также видно, что алгоритм FCFP позволяет достичь существенно лучших результатов по сравнению с ECFP.

Другое неожиданное наблюдение – наилучшие результаты достигаются при $r = 2$. При больших значениях r увеличиваются количество и размер рассматриваемых фрагментов, но вырастает вероятность коллизий – ситуаций, когда один и тот же бит соответствует совершенно разным фрагментам. Наблюдаются существенное падение точности при росте r для маленьких значений l и сравнительно небольшое падение точности при росте r для больших значений l. Так как при больших значениях l вероятность коллизий падает, наблюдаемое поведение означает, что наибольшее значение для предсказания pK_a имеют небольшие фрагменты молекулы, а коллизии существенно влияют на точность. При фиксированном r точность растет с ростом l. В целом сходное поведение наблюдается и для точности для наборов данных Novartis и Literature. Соответствующие данные приведены на рис. 3 и в табл. 2.

Применение метода опорных векторов для предсказания констант кислотности

В работе [1] показано, что при использовании гиперпараметров по умолчанию и круговых

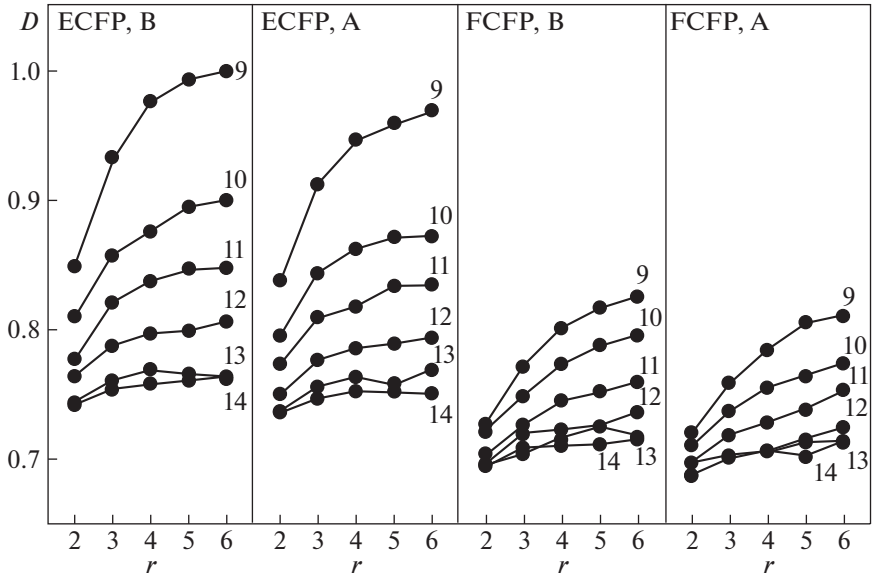

Рис. 2. Зависимости средней абсолютной ошибки D предсказания pK_a от параметра r (радиус или размер субструктур) алгоритма вычисления круговых молекулярных отпечатков пальцев для различных вариантов алгоритма (ECFP, FCFP) и различных значений длины молекулярного отпечатка пальцев l: 512, 1024, 2048, 4096, 8192, 16384 ($\log_2 l = 9$, 10, 11, 12, 13, 14). Буквы B, A после названия алгоритма обозначают бинарные и аддитивные молекулярные отпечатки пальцев соответственно, цифрами рядом с кривыми обозначены значения $\log_2 l$.

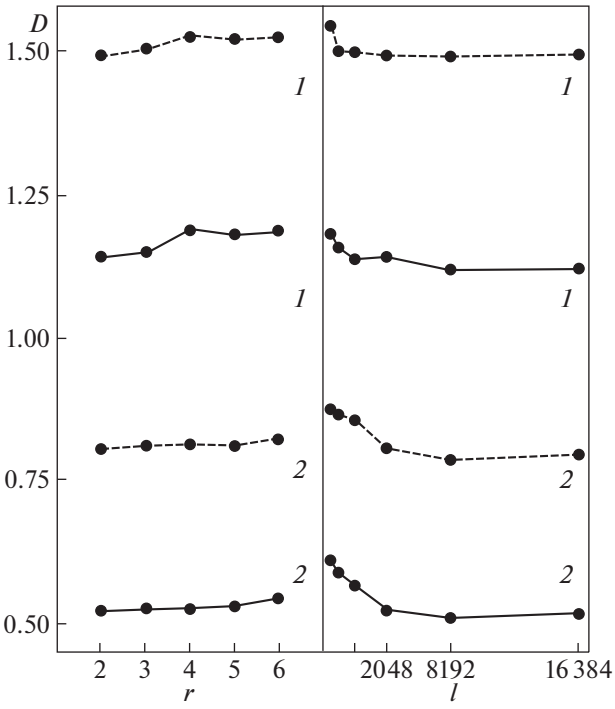

Рис. 3. Зависимости ошибки D предсказания pK_a от параметра r (при $l = 4096$) и параметра l (при $r = 2$). Для предсказания использован алгоритм вычисления круговых молекулярных отпечатков пальцев FCFP, аддитивные молекулярные отпечатки пальцев. Сплошной линией обозначена средняя абсолютная ошибка, пунктирной – среднеквадратичная ошибка. Цифрами показаны тестовые наборы данных: *1* – Novartis, *2* – Literature.

молекулярных отпечатков пальцев метод опорных векторов дает существенно худшие результаты по сравнению со случайным лесом. Однако после оптимизации гиперпараметров и выбора более удачного входного представления молекул удается добиться сравнимой точности. В качестве входного представления молекулы были использованы все доступные 2D молекулярные дескрипторы, рассчитываемые с помощью библиотеки RDKit (класс MoleculeDescriptors.MolecularDescriptorCalculator, функция rdMolDescriptors.MQNs_), а также набор молекулярных дескрипторов, рассчитанный с помощью библиотеки CDK и описанный в работе [28]. MF не использовались. Использовались следующие значения гиперпараметров: $\gamma = 0.153$, $c = 15.9$, $\varepsilon = 0.0042$. Достигнутая точность приведена в табл. 3, сопоставление предсказанных и референсных значений показано на рис. 4.

ЗАКЛЮЧЕНИЕ

Множество параметров влияет на точность предсказания pK_a с помощью машинного обучения с использованием одних и тех же наборов данных. Эти параметры могут быть оптимизированы для достижения наилучшей точности предсказания. Кроме того, при публикации результатов таких исследований в случае, когда исходные коды не публикуются, важно указывать, какие именно параметры были использованы. Изучено влияние различных факторов на предсказание

Таблица 2. Точность предсказания pK_a с помощью случайного леса (1000 деревьев) и различных вариантов круговых молекулярных отпечатков пальцев (l = 4096, r = 2)

Набор данных	Бинарные FP		Аддитивные FP	
	ECFP	FCFP	ECFP	FCFP
Novartis (среднеквадратичная ошибка)	1.68	1.48	1.66	1.49
Literature (среднеквадратичная ошибка)	1.00	0.79	0.99	0.81
Novartis (средняя абсолютная ошибка)	1.30	1.12	1.30	1.14
Literature (средняя абсолютная ошибка)	0.65	0.52	0.66	0.52

Таблица 3. Точность предсказания pK_a с помощью метода опорных векторов

Набор данных	Среднеквадратичная ошибка	Средняя абсолютная ошибка
Main (кросс-валидация)	1.04	0.66
Novartis	1.66	1.30
Literature	0.93	0.64

pK_a для органических молекул, содержащих одну подверженную диссоциации функциональную группу. Показано, что важными являются способ выбора зарядовой и таутомерной формы молекулы, параметры круговых молекулярных отпечатков пальцев. Установлено, что использование программного обеспечения OpenEye QuacPac позволяет достичь лучших результатов по сравнению со средствами библиотеки RDKit. Наилучшие результаты (при использовании случайного леса) достигаются при использовании радиуса (размера подструктур) r = 2. Метод опорных векторов может давать весьма точные результаты при оптимизации гиперпараметров. Полученные результаты могут быть использованы при дальнейшей разработке наиболее точных методов предсказания pK_a низкомолекулярных соединений.

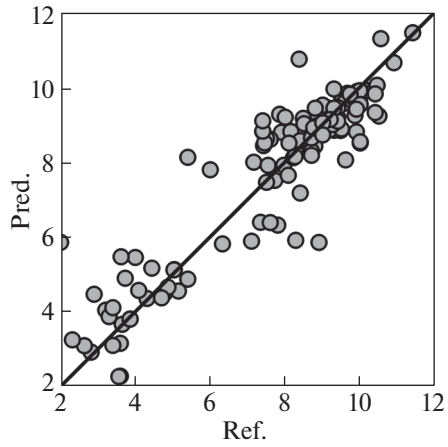

Рис. 4. Сопоставление предсказанных (Pred.) с помощью метода опорных векторов и референсных (Ref.) значений pK_a для тестового набора данных Literature.

Исследование выполнено за счет гранта Российского научного фонда (проект № 22-13-00266), предоставленного Институту физической химии и электрохимии имени А.Н. Фрумкина Российской академии наук.

СПИСОК ЛИТЕРАТУРЫ

1. *Baltruschat M., Czodrowski P.* // F1000Res. 2020. V. 9. P. 113. https://doi.org/10.12688/f1000research.22090.2
2. *Mansouri K., Cariello N.F., Korotcov A. et al.* // J. Cheminform. 2019. V. 11. № 1. P. 60. https://doi.org/10.1186/s13321-019-0384-1
3. *Mayr F., Wieder M., Wieder O. et al.* // Front. Chem. 2022. V. 10. P. 866585. https://doi.org/10.3389/fchem.2022.866585
4. *Lu Y., Anand S., Shirley W. et al.* // J. Chem. Inf. Model. 2019. V. 59. № 11. P. 4706. https://doi.org/10.1021/acs.jcim.9b00498
5. *Rupp M., Korner R., Tetko I.* // CCHTS. 2011. V. 14. № 5. P. 307. https://doi.org/10.2174/138620711795508403
6. *Lionta E., Spyrou G., Vassilatis D. et al.* // CTMC. 2014. V. 14. № 16. P. 1923. https://doi.org/10.2174/1568026614666140929124445
7. *Bahi M., Batouche M.* // 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS). Tebessa: IEEE, 2018. P. 1–5. https://doi.org/10.1109/PAIS.2018.8598488
8. *Yang Q., Ji H., Fan X. et al.* // J. Chromatogr. A. 2021. V. 1656. P. 462536. https://doi.org/10.1016/j.chroma.2021.462536
9. *Fedorova E.S., Matyushin D.D., Plyushchenko I.V. et al.* // J. Chromatogr. A. 2022. V. 1664. P. 462792. https://doi.org/10.1016/j.chroma.2021.462792
10. *Milyushkin A.L., Matyushin D.D., Buryak A.K.* // J. Chromatogr. A. 2020. V. 1613. P. 460724. https://doi.org/10.1016/j.chroma.2019.460724
11. *Zenkevich I.G., Nikitina D.A.* // Russ. J. Phys. Chem. A. 2021. V. 95. № 2. P. 395. https://doi.org/ *Зенкевич И.Г., Никитина Д.А.* // Журн. физ. химии. 2021. Т. 95. № 2. С. 285. https://doi.org/10.1134/S003602442102028X
12. *Angra S., Ahuja S.* // 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC). Chirala, Andhra Pradesh, India: IEEE, 2017. P. 57. https://doi.org/10.1109/ICBDACI.2017.8070809
13. *Mansouri K., Grulke C.M., Judson R.S. et al.* // J. Cheminform. 2018. V. 10. № 1. P. 10. https://doi.org/10.1186/s13321-018-0263-1

14. *Parmar A., Katariya R., Patel V.* // International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018 / Ed. Hemanth J., Fernando X., Lafata P. et al. Cham: Springer International Publishing, 2019. V. 26. P. 758. https://doi.org/10.1007/978-3-030-03146-6_86
15. *Cereto-Massagué A., Ojeda M.J., Valls C. et al.* // Methods. 2015. V. 71. P. 58. https://doi.org/10.1016/j.ymeth.2014.08.005
16. *Rogers D., Hahn M.* // J. Chem. Inf. Model. 2010. V. 50. № 5. P. 742. https://doi.org/10.1021/ci100050t
17. *Xiong J., Li Z., Wang G. et al.* // Bioinformatics / Ed. by Z. Lu. 2022. V. 38. № 3. P. 792. https://doi.org/10.1093/bioinformatics/btab714
18. *Pan X., Wang H., Li C. et al.* // J. Chem. Inf. Model. 2021. V. 61. № 7. P. 3159. https://doi.org/10.1021/acs.jcim.1c00075
19. *Reza Ghiasi, Zamani A., Shamami M.K.* // Russ. J. Phys. Chem. A. 2019. V. 93. № 8. P. 1537. https://doi.org/10.1134/S0036024419080247
20. *Prasad S., Huang J., Zeng Q. et al.* // J. Comput. Aided Mol. Des. 2018. V. 32. № 10. P. 1191. https://doi.org/10.1007/s10822-018-0167-1
21. *Pracht P., Wilcken R., Udvarhelyi A. et al.* // J. Comput. Aided Mol. Des. 2018. V. 32. № 10. P. 1139. https://doi.org/10.1007/s10822-018-0145-7
22. *Pedregosa F., Varoquaux G., Gramfort A. et al.* Scikit-learn: Machine Learning in Python: arXiv:1201.0490. arXiv, 2018. https://arxiv.org/abs/1201.0490
23. *Bento A.P., Hersey A., Félix E. et al.* // J. Cheminform. 2020. V. 12. № 1. P. 51. https://doi.org/10.1186/s13321-020-00456-1
24. *Chang C.-C., Lin C.-J.* // ACM Trans. Intell. Syst. Technol. 2011. V. 2. № 3. P. 1. https://doi.org/10.1145/1961189.1961199
25. *Willighagen E.L., Mayfield J.W., Alvarsson J. et al.* // J. Cheminform. 2017. V. 9. № 1. P. 33. https://doi.org/10.1186/s13321-017-0220-4
26. https://github.com/czodrowskilab/Machine-learning-meets-pKa
27. *Heller S., McNaught A., Stein S. et al.* // J. Cheminform. 2013. V. 5. № 1. P. 7. https://doi.org/10.1186/1758-2946-5-7
28. *Matyushin D.D., Buryak A.K.* // IEEE Access. 2020. V. 8. P. 223140. https://doi.org/10.1109/ACCESS.2020.3045047