

УДК 547:546.98:541.128

ОПРЕДЕЛЕНИЕ ДЕСКРИПТОРОВ КАТАЛИТИЧЕСКИХ СИСТЕМ В МАШИННОМ ОБУЧЕНИИ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ КИНЕТИЧЕСКИХ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

© 2024 г. А. Ф. Шмидт^а, *, Н. А. Сидоров^а, А. А. Курохтина^а, Е. В. Ларина^а, Н. А. Лагода^а

^аФГБОУ ВО Иркутский государственный университет, химический факультет,
ул. К. Маркса, 1, Иркутск, 664033 Россия

*e-mail: aschmidt@chem.isu.ru

Поступила в редакцию 20.08.2024 г.

После доработки 12.09.2024 г.

Принята к публикации 12.09.2024 г.

Проблема выбора и определения значений дескрипторов свойств компонентов химических реакций в математических моделях химических процессов является одной из центральных при создании моделей машинного обучения (МО), используемых для описания и предсказания особенностей функционирования химических систем. Сложившая практика в основном предполагает применение в качестве дескрипторов заранее найденных экспериментально или расчетным путем физико-химических характеристик компонентов реакционных систем (ионных радиусов, длин связей, энергий и иных параметров, связанных со строением и свойствами конкретных молекул или частиц). В настоящей работе представлены результаты предсказания вида интегральных кинетических зависимостей, а также подходов к определению значений дескрипторов для характеристики свойств ряда простейших палладиевых предшественников катализатора при их использовании в реакции Сузуки–Мияуры. Поставленная задача решалась путем создания моделей МО, учитывающих экспериментальные кинетические данные. Полученные в результате тренировки моделей дескрипторы позволяют удовлетворительно описывать кинетические закономерности реакции Сузуки–Мияуры с арилхлоридами в т.н. “безлигандных” каталитических условиях, характерной особенностью которых является повышенная чувствительность реакции к малым изменениям условий проведения.

Ключевые слова: катализ, машинное обучение, реакция Сузуки–Мияуры, палладий, кинетика

DOI: 10.31857/S0453881124030075, **EDN:** RVOAXS

ВВЕДЕНИЕ

Применение технологий машинного обучения (МО) в решении задач предсказания количественных параметров химических (включая каталитические) процессов, чаще всего, выходов продуктов реакции, на основании данных о составе и свойствах компонентов реакционных систем, а также об условиях проведения процесса в последние годы развивается чрезвычайно высокими темпами. Можно выделить два принципиально противоположных подхода к созданию таких моделей в части выбора исходных данных

для обучения – экспериментальных [1–3] или расчетных (т.н. *in silico*) [4–6]. Вне зависимости от источника данных (полученных в ходе реального химического эксперимента или путем его симулирования), в подавляющем большинстве случаев для машинного обучения используются данные интегрального типа, чаще всего величины выходов продуктов или интегральной селективности реакции по тому или иному продукту за конечный промежуток времени. Тем не менее, существуют немногочисленные примеры исследований каталитических реакций с привлечением технологий МО, в которых в числе прочих учитываются кинетические данные о закономерностях протекания реакций во времени [2, 7–9]. Авторы практически всех без исключения работ отмечают важность и при этом неоднозначность выбора входных параметров моделей МО, а именно дескрипторов свойств компонентов реакционных систем [3, 10–14]. Решение за-

Сокращения и обозначения: ДМФА – *N,N*-диметилформамид, МО – машинное обучение, МП – модели с прямым предсказанием концентраций, МС – модели с предсказанием аппроксимационной сигмоиды, НС – нейронная сеть, NMP – *N*-метилпирролидон, ТОН – число оборотов катализатора, ДИП – пламенно-ионизационный детектор, МАЕ – средний модуль отклонения (mean absolute error).

дачи определения типа дескриптора и его количественной характеристики в публикуемых работах решается различными, иногда прямо противоположными способами. Так в качестве дескрипторов применяются физико-химические характеристики молекул, определяющие, согласно фундаментальным теоретическим представлениям, в той или иной степени реакционную способность компонента в исследуемой химической реакции, например, стерические параметры металлокомплексных предшественников катализатора [15] или энергии ионизации молекул [16]. С другой стороны, существуют примеры использования в качестве дескрипторов параметров, с теоретической точки зрения неочевидно связанных с превращениями вещества в изучаемых условиях, таких, например, как данные о кристаллических структурах органических молекул для предсказания выходов продуктов реакций, проводимых в растворе [17]. Кроме того, возрастающий интерес к применению методов МО в химических исследованиях имеет следствием создание и использование при построении и тренировке моделей комплексных молекулярных дескрипторов, одновременно учитывающих информацию о составе, структурных, электронных и/или каких-либо иных свойствах молекул [12, 18–23]. Тем не менее, ни один из указанных подходов не может претендовать на универсальность. В настоящей работе нами была предпринята попытка качественного и количественного определения дескрипторов для характеристики ряда простейших растворимых палладиевых предшественников катализатора при их использовании в реакции Сузуки–Мияуры [24] в так называемых “безлигандных” условиях (схема 1). При этом ключевой особенностью создаваемых моделей МО являлся учет кинетических данных, полученных в ходе реальных каталитических экспериментов.

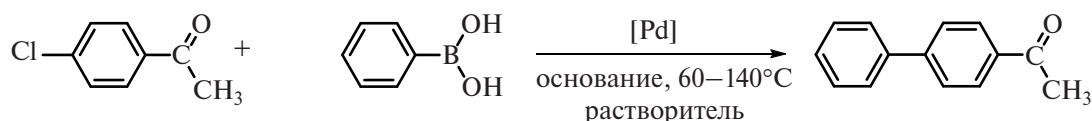
ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Все эксперименты, результаты которых использовались для тренировки создаваемых моделей МО, проводили без применения инертной атмосферы. Периодически отбираемые пробы реакционной смеси анализировали на газожидкостном хроматографе Кристалл 5000.2 (“Хроматэк”, Россия, ДИП, колонка НР-5 15 м) с программированным нагревом от 110 до 250°C. Значения аналитического выхода продукта в отбираемых пробах реакционного раствора находили методом внутреннего стандарта (нафталин) с применением фактора отклика, определяемого по аутентичному образцу. Материальный баланс реакции рассчитывали в каждой пробе реакционной смеси с учетом количества израсходовавшихся субстратов и образовавшихся в результате реакции продуктов. Отклонения баланса от теоретического значения не превышали 5%.

Для оценки воспроизводимости каждый эксперимент проводили 2 раза.

Каталитические эксперименты

В реакции Сузуки–Мияуры 4-хлорацетофенон (1.25–57.5 ммоль) и нафталин в качестве внутреннего стандарта для хроматографии (1 ммоль) растворяли при комнатной температуре в 5 мл растворителя. Полученный раствор вводили в стеклянный реактор, снабженный резиновой мембраной и магнитным мешальником, содержащий фенилборную кислоту (1.25–10 ммоль), основание (0.41–6.5 ммоль) и палладиевый предшественник катализатора (0.8×10^{-3} –0.16 ммоль). Реакцию начинали, помещая реактор в предварительно нагретую до температуры реакции (60–140°C) масляную баню при перемешивании (480 об/мин). Пробы реакционной смеси периодически отбирали из реактора с помощью шприца с металлической иглой. Для хроматографического анализа 100 мкл пробы реакционного раствора



[Pd] = PdCl₂ (порошок, раствор), PdBr₂, Pd(OAc)₂, Pd(AcAc)₂;
 основание = Na₂CO₃, NaOAc, NaOH;
 растворитель : ДМФА : H₂O (в разных соотношениях), NMP : H₂O (4 : 1),
 этанол : H₂O (4 : 1).

Схема 1. Реакция Сузуки–Мияуры.

экстрагировали 100 мкл хлороформа. Продолжительность реакции до полного прекращения конверсии исходных веществ составляла 3–7 ч.

В качестве растворителей использовали *N,N*-диметилформамид (ДМФА), *N*-метилпирролидон (NMP), этанол и их смеси с водой в различных соотношениях, в качестве оснований – Na_2CO_3 , NaOAc , NaOH .

В качестве палладиевого предшественника катализатора применяли PdCl_2 , $\text{Pd}(\text{Acac})_2$, $\text{Pd}(\text{OAc})_2$, PdBr_2 (все перечисленные в виде порошка), а также предварительно сформированный раствор PdCl_2 в ДМФА и его смесях с водой как растворителем.

Обработка результатов, построение и обучение моделей МО

Для тренировки моделей МО нами был создан датасет, содержащий экспериментальные данные о временных зависимостях концентрации 4-ацетилдифенила, образующегося в реакции Сузуки–Мияуры (схема 1) при различных исходных молярных концентрациях сочетающихся 4-хлорацетифенона и фенилборной кислоты, природы и концентрации основания (3 типа основания), палладиевого предшественника катализатора (5 типов катализатора), а также природы растворителя (3 типа растворителей плюс 10 типов их смесей с водой) и температуры процесса (4 варианта температур). Итоговый датасет содержал в общей сложности данные 60 кинетических экспериментов с 637 парами значений “время–концентрация продукта” (5–18 пар для каждого из 60 экспериментов).

Анализ методом главных компонент датасета (через матрицу парных корреляций) осуществляли с использованием расширения для Excel ExStatRC, версия 1.2 на базе статистического пакета “R” [25].

При создании предсказательных моделей для описания кинетических закономерностей реакции Сузуки–Мияуры по данным об исходном составе реакционной смеси, условиях проведения процесса и изменениях концентрации компонентов во времени нами применялись два основных подхода, описанных ниже. Программный код для всех обсуждаемых в работе моделей был написан в Python (версия 3.8.5) посредством приложения Jupiter Notebook (свободно распространяемый пакет программ Anaconda, версия 2023.09-0).

Первый подход. Для тренировки модели использовали данные о начальном состоянии реакционной смеси (компонентный состав, включая растворитель, и температура), а также значения концентрации продукта реакции Сузуки–Мияуры (4-ацетилдифенила, схема 1) в различные моменты времени, определяемые в ходе кинетических экспериментов. Для построения модели применяли ансамбли решающих деревьев (RandomForestRegressor, AdaBoostRegressor и их комбинацию, StackingRegressor) из пакета sklearn.ensemble свободно распространяемой библиотеки Sklearn для языка Python [26].

Для каждого из вариантов архитектуры модели ее гиперпараметры подбирали путем их полного варьирования (Grid Search из sklearn). Метрика качества при тренировке – средний модуль отклонения MAE (mean absolute error) предсказываемых значений концентрации продукта от определенных экспериментально для всех моментов времени в экспериментах, используемых при обучении. Наилучший результат показал метод случайного леса (MAE = 0.099) с 500 деревьями (рост качества с дальнейшим увеличением количества деревьев был пренебрежимо мал) без ограничений по размерам деревьев.

Результатом работы модели предполагалось предсказание концентрации продукта реакции в заданный момент времени, т.е. фактически определение вида интегральной кинетической кривой накопления продукта реакции Сузуки–Мияуры по данным об исходном состоянии реакционной системы. Далее при обсуждении моделей такого типа будем обозначать их как модели с прямым предсказанием концентраций (МП).

Второй подход. Для тренировки модели применяли предварительно обработанные экспериментальные данные. В этом случае интегральные кривые, характеризующие изменения концентрации продукта реакции Сузуки–Мияуры во времени, были предварительно аппроксимированы функцией типа сигмоида, описываемой следующим уравнением:

$$P(t) = \frac{K}{1 + e^{-At+B}}, \quad (1)$$

где $P(t)$ – концентрация продукта в текущий момент времени t (мин); A , B и K – параметры сигмоиды.

Выбор уравнения сигмоиды в качестве функции, аппроксимирующей экспериментальные кинетические кривые накопления продукта ка-

талитической реакции, был обусловлен ее более универсальной способностью описывать распространенные типы интегральных кинетических кривых каталитических реакций, включающих этапы автоускорения (в случае протекания процессов формирования/разработки катализатора) и окончания реакции при низких степенях превращения (в случае протекания процессов дезактивации/отравления катализатора или интенсивных обратных процессов). Такие типы кривых часто относят к так называемым s-образным кривым. Это выгодно отличает сигмоиду от, например, функции гиперболического тангенса, недавно использованного в работе [9] и способного описать лишь кривые с насыщением без начального этапа ускорения реакции. Параметры A , B , K уравнения (1) для каждого кинетического эксперимента были подобраны с помощью метода наименьших квадратов, реализованного в программном пакете `scipy.optimize`. Подобранные значения указанных параметров совместно с данными о начальном состоянии реакционной смеси применяли для тренировки моделей, построенных с применением методов `RandomForestRegressor`, `StackingRegressor`, `AdaBoostRegressor` из пакета `sklearn.ensemble` библиотеки `Sklearn` [26]. Гиперпараметры, необходимые для тренировки моделей с использованием данных, аппроксимированных уравнением сигмоиды (1), для каждого из указанных методов подбирали, оценивая качество модели с помощью следующей комплексной метрики:

$$m = 0.05|A_s - A_p| + 0.05|B_s - B_p| + 0.9|K_s - K_p|, \quad (2)$$

где A_s , B_s , K_s — параметры сигмоиды (1), найденные путем аппроксимации, A_p , B_p , K_p (от англ. “predicted”) — аналогичные параметры, предсказываемые моделью.

Величины вкладов слагаемых в метрику (2) были выбраны исходя из того, что значение параметра K (асимптота сигмоиды (1), ассоциируемая в нашем случае с величиной концентрации продукта реакции к моменту ее остановки, определяющей выход продукта) имело наименьшие относительные значения среди искоемых параметров уравнения (1) и колебалось в наиболее узких пределах.

Тренировку моделей, использующих уравнение сигмоиды (1) и ансамблевые методы, проводили по алгоритму, аналогичному МП. Наилучший результат также показал метод случайного леса ($MAE = 0.115$).

Результатом работы модели в этом случае являлось предсказание параметров A , B , K уравнения (1) по данным о начальном состоянии реакционной смеси. Полученные таким образом параметры применяли для расчета интегральных кинетических кривых накопления продукта реакции Сузуки—Мияуры $P(t)$ (1). Далее при обсуждении моделей такого типа будем обозначать их как модели с предсказанием аппроксимационной сигмоиды (МС).

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

При обращении к методам МО для создания предсказательных моделей каталитических процессов для тренировки моделей исследователями в основном используются данные об интегральных величинах (выходах продуктов или интегральных селективностях реакции) при заданных начальных условиях процесса [1, 2, 7, 13, 21–23, 27]. Такой подход к применению технологий МО, безусловно, не является неожиданным, поскольку выход продукта за конечный промежуток времени — зачастую единственный измеряемый параметр реакции в экспериментальных исследованиях каталитических процессов. Не случайно примеры учета кинетических данных в работах, связанных с применением технологий МО, весьма ограничены, что, безусловно, в первую очередь связано с высокой трудоемкостью кинетических исследований, требующих многократных измерений концентраций реагирующих веществ в ходе каждого эксперимента [28]. Тем не менее, поскольку катализ представляет собой кинетическое явление [29], анализ закономерностей развития реакции во времени способен дать уникальную информацию, необходимую для рационального поиска эффективных экспериментальных протоколов, что, в том числе, является целью создания предсказательных моделей с привлечением методов МО. Нами была предпринята попытка построения моделей МО, позволяющих предсказывать кинетику накопления продуктов реакции Сузуки—Мияуры при использовании наиболее привлекательной с точки зрения практического применения комбинации наиболее доступных в ряду арилгалогенидов арилхлоридов и простейших “безлигандных” палладиевых каталитических систем. Под “безлигандными” системами в данном случае понимаются каталитические системы, не содержащие в своем составе фосфиновых, аминовых, карбеновых или любых других сильных органических лигандов, добавки которых приводят

к существенному удорожанию и усложнению экспериментальных и технологических процедур. Однако в отсутствие сильных лигандов, способных стабилизировать соединения $\text{Pd}(0)$, несмотря на высокие пиковые значения каталитической активности резко ухудшается стабильность работы каталитических систем, что ведет к значительному падению значений TON (суммарного числа оборотов) катализаторов и, соответственно, снижению выходов продуктов. Обычно это связывают с интенсификацией процессов агломерации палладия с образованием металлической фазы, которые считаются главным направлением дезактивации каталитических систем реакций кросс-сочетания арилгалогенидов [28, 30]. Особенно большие проблемы со стабильностью работы катализаторов возникают именно в случае использованных в настоящем исследовании арилхлоридов — наиболее доступных, но при этом наименее реакционно-способных в ряду арилгалогенидов субстратов.

Применяемый нами для тренировки моделей датасет содержал достаточно большой с точки зрения экспериментальных исследований массив данных (60 кинетических экспериментов). Тем не менее, с позиций технологий МО такой датасет является ограниченным (small/limited/low data) [22, 31], что представляет собой существенное препятствие при обучении моделей, способных давать удовлетворительные предсказательные результаты.

Следует также отметить, что накопленный экспериментальный материал по реакциям кросс-сочетания арилгалогенидов в целом и реакции Сузуки—Мияуры в частности [30, 32–34] свидетельствует о большой вероятности того, что в каталитических системах этих реакций протекают нелинейные с точки зрения химической кинетики процессы. Например, к ним относится уже упоминавшийся выше процесс дезактивации катализатора в результате агломерации соединений $\text{Pd}(0)$. Одна из особенностей таких процессов — их способность к нелинейным эффектам (бифуркациям), что проявляется в повышенной чувствительности реакций к незначительным варьированиям условий реакций. Возникновение такого рода эффектов может поставить под сомнение способность эмпирических математических моделей, к которым, по сути, относятся модели МО, адекватно описывать экспериментальные данные.

В качестве нового аргумента в пользу значимого влияния нелинейных кинетических эф-

фектов на реакцию Сузуки—Мияуры можно рассматривать результаты анализа использованного нами экспериментального датасета методом главных компонент (principal component analysis) (рис. 1). Понижение размерности пространства параметров реакции до двумерного приводит к возникновению трех областей на поверхности главных компонент, соответствующих трем использованным диапазонам температур реакции (серые области на рис. 1), и трем областям, соответствующим типу растворителя (красные области на рис. 1). Это указывает на особую значимость температуры и типа растворителя в ряду параметров реакции. При этом, однако, на поверхности главных компонент не было зафиксировано локализаций экспериментов по такому важнейшему с практической точки зрения параметру реакции, как выход продукта (диапазоны наблюдаемых выходов продукта на рис. 1 отмечены цветом маркеров). Выходы продукта варьировались в широких пределах по всей поверхности главных компонент, в том числе внутри всех вариантов неперекрывающихся областей, за исключением области “60°—этанол”, в которой они были стабильно низкими. Таким образом, локализовать перспективные с точки зрения значений выходов продукта области пространства параметров не представлялось возможным, что, на наш взгляд, является следствием повышенной чувствительности реакции к условиям ее проведения и связано с реализацией нелинейных кинетических эффектов.

Проблема ограниченного набора данных, применяемых для обучения моделей, всегда возникает при использовании для обучения реальных экспериментально измеренных величин, в отличие от датасетов, содержащих исключительно *in silico* данные, полученные расчетным путем [7, 31]. В случае МО в такой ситуации логичным является выбор методов, требующих для своей работы значительно меньшего объема данных. Кроме того, этот выбор должен был учитывать способность методов описать бифуркационные (качественные) изменения режима протекания реакции при малом варьировании входных параметров. Одно из лучших семейств таких методов — ансамбли решающих деревьев (ensembles of decision trees), поскольку они дают возможность построения моделей с достаточно хорошей предсказательной точностью, не требуя для обучения большого количества данных. В связи с ограниченным размером датасета нами был использован наиболее жесткий с точки зрения оценки метрик качества обучения подход,

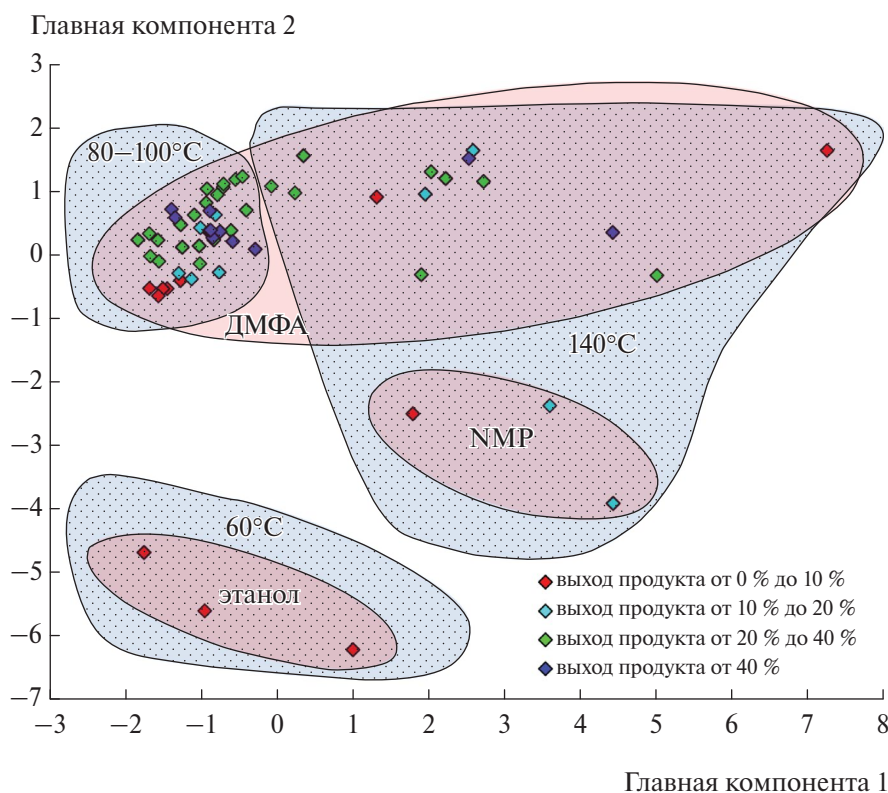


Рис. 1. Результаты анализа экспериментального датасета (60 экспериментов, 15 параметров реакции) методом главных компонент. Серые области соответствуют экспериментам при разных температурах, красные области — экспериментам с разными типами растворителей. Красные точки соответствуют экспериментам с выходами продукта от 0 до 10%, бирюзовые — от 10 до 20%, зеленые — от 20 до 40%, синие — от 40% и выше.

при котором для тренировки моделей применяли 59 различных наборов экспериментов, при этом каждый раз 1 эксперимент выбирали в качестве тестового. Соответственно, процедура тренировки включала 60 итераций типа “тренировка—оценка качества модели”.

Для сравнения итоговых моделей МП и МС были выбраны два типа метрик качества. В качестве первой использовали среднее отклонение предсказанного значения концентрации продукта по окончании реакции от экспериментального значения. Эта метрика предполагала оценку качества описания выхода продукта реакции. Значение метрики для МП составило 0.073, для МС — 0.089. Вторая метрика предусматривала расчет среднего отклонения предсказанной концентрации продукта от экспериментальной в каждом моменте времени для всех имеющихся в датасете экспериментов. В этом случае значение метрики для МП составило 0.051, для МС — 0.063. Найденные значения предлагаемых метрик указывают на близкое качество моделей МС и МП с некоторым превосходством последней. Зависимости концентраций продукта реак-

ции Сузуки—Мияуры в каждый момент времени для включенных в датасет экспериментов, предсказанных с использованием полученных моделей и измеренных экспериментально, представлены на рис. 2.

Поскольку модели МП и МС, построенные с помощью решающих деревьев, представляют собой кусочно-постоянные функции, монотонно изменяющиеся в пространстве переменных параметров датасета, результаты предсказания с их применением в промежуточных областях пространства между точками с известными экспериментальными значениями переменных (задача интерполяции) могут оказаться достаточно грубыми. Альтернативным методом МО, обеспечивающим непрерывность предсказательной способности в используемом для обучения диапазоне переменных входных параметров (в нашем случае, концентраций компонентов реакционной смеси и температуры), может быть применение искусственных нейронных сетей (НС), также способных воспроизводить резкие изменения выходных параметров при незначительном варьировании входных.

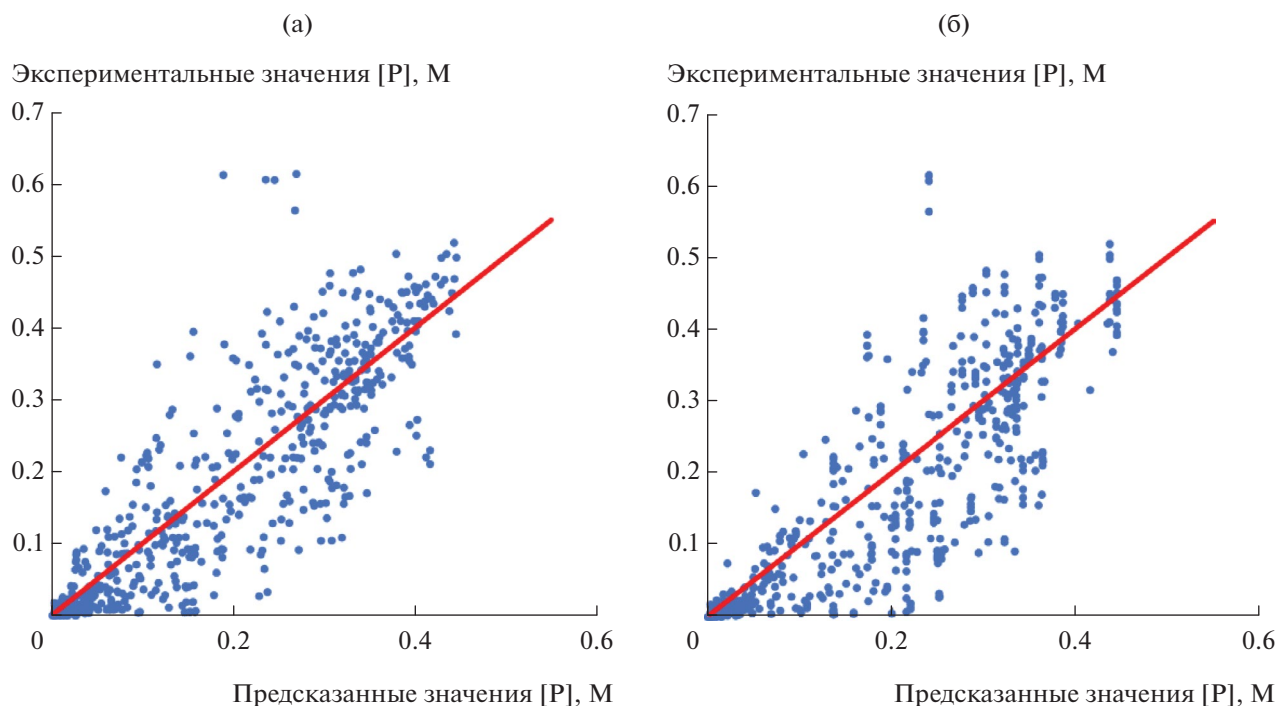


Рис. 2. Зависимости включенных в датасет экспериментально определенных значений концентрации продукта реакции Сузуки–Мияуры (схема 1) от предсказанных с использованием модели МП (а) и рассчитанных по уравнению (1) с применением параметров, предсказанных МС (б).

Однако в силу ограниченности экспериментального датасета прямое применение этого подхода для получения адекватной предсказательной модели оказывается невозможным вследствие нехватки данных для ее тренировки. Для преодоления обозначенных сложностей для обучения полносвязных НС на ограниченных датасетах нами был использован гибридный подход, базирующийся на комбинации ансамблевого метода и НС. Обучение последних проводилось на большом объеме *in silico* данных, полученных с помощью предварительно обученного на реальных экспериментальных данных ансамбля решающих деревьев.

Для этой цели мы выбрали модель МС, результаты тренировки которой представлены на рис. 2б. Для получения датасета, имеющего достаточный для обучения НС размер, из имеющихся в исходном экспериментальном датасете значений входных параметров генерировали их случайные комбинации. 10 000 таких случайных комбинаций (далее строк) из ~350 000 возможных применяли в качестве входных данных для расчета выходных параметров с использованием модели МС, обученной на экспериментальных данных в соответствии с описанной в Экспериментальной части процедурой (рис. 3). Посколь-

ку предсказательная способность модели МС для датасета с экспериментальными данными была удовлетворительной, можно полагать, что предсказательная способность НС, обученной с помощью симулированного датасета, содержащего в качестве входных параметров те же экспериментальные данные в случайных комбинациях, не будет принципиально отличаться от модели МС, обученной методом решающих деревьев. При этом качество описания будет превосходить способности ансамбля, применяемого в МС, для промежуточных областей пространства параметров между экспериментальными точками за счет непрерывности и дифференцируемости НС.

Архитектуру НС выбирали подбором количества слоев, количества нейронов в каждом слое, функций активации, а также подбирая тип оптимизатора (SGD, RMSprop, Adagrad, Adadelat, Adam, Adamax, Nadam) и его шаг. В качестве метрики по аналогии с МП использовали MAE. Количество изменяемых параметров модели варьировалось в диапазоне от 5 000 до 10 000, т.е. таким образом, число таких параметров не превышало числа строк датасета. При этом при добавлении новых слоев нейронов количество нейронов в существующих слоях уменьшалось

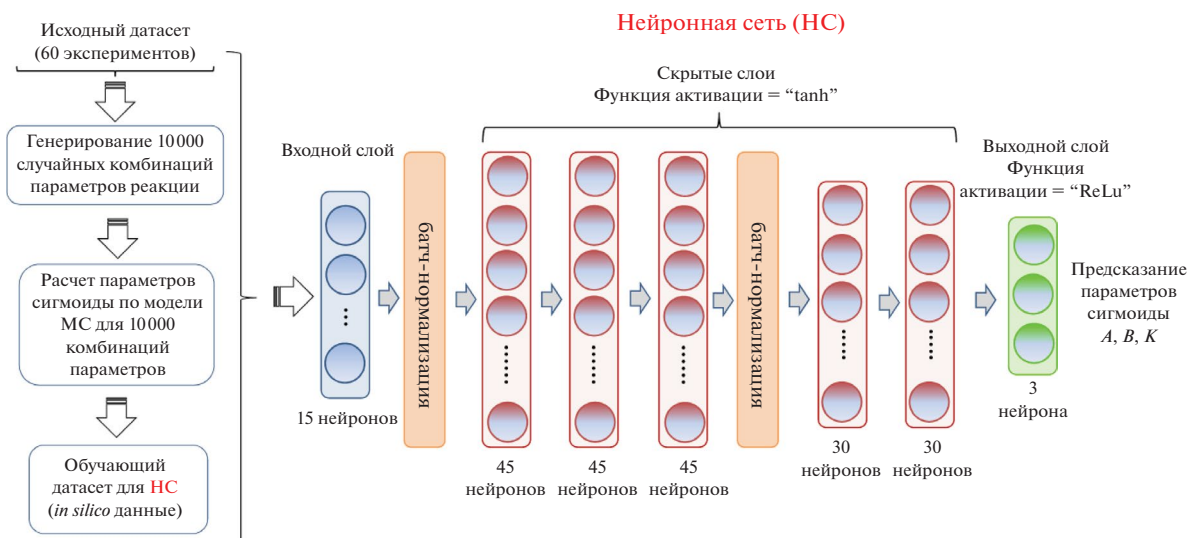


Рис. 3. Подготовка данных для обучения НС с использованием модели МС (слева) и архитектура полносвязной НС, определенная по результатам подбора гиперпараметров.

таким образом, чтобы число параметров находилось в указанных пределах, и количество нейронов в каждом последующем слое было не меньше, чем в следующем. Добавление слоев прекращали при отсутствии улучшения результата тренировки. Для заданного числа слоев количество нейронов варьировалось в соответствии с указанными выше особенностями. Далее добавляли слои батч-нормализации, перебирая их количество и место в последовательности слоев. В конечном варианте использовался оптимизатор Adam (learning rate 0.003). Итоговая модель (рис. 3) содержала 5 скрытых слоев (последовательность из 3 слоев по 45 нейронов и 2 слоев из 30 нейронов), функция активации — гиперболический тангенс, и решающий слой из 3 нейронов с функцией активации ReLu. Также сеть содержала 2 слоя для нормирования входных и выходных данных (батч-нормализация).

В результате обучения НС с применением указанных гиперпараметров достигалось удовлетворительное отклонение выходных параметров, предсказанных НС, от их значений в симулированном датасете (MAE по сумме параметров A , B , K составляло 0.1845 при тестовой выборке 2000 строк, 20% данных). Зависимость значений концентраций продукта реакции Сузуки—Мияуры, рассчитанных по уравнению (1) с параметрами A , B , K , определенными в результате работы обученной НС, от наблюдаемых в каждый момент времени для всех экспериментов, входящих в исходный датасет, представлена на рис. 4.

Таким образом, в результате применения ансамблевых методов в двух различных вариантах — с аппроксимацией интегральных кинетических зависимостей математической функцией (сигмоидой) и непосредственным использованием интегральных кинетических данных о зависимостях концентраций продукта реакции Сузуки—Мияуры при варьировании начальных условий ее проведения — нам удавалось получить модели МО, обладающие удовлетворительной предсказатель-

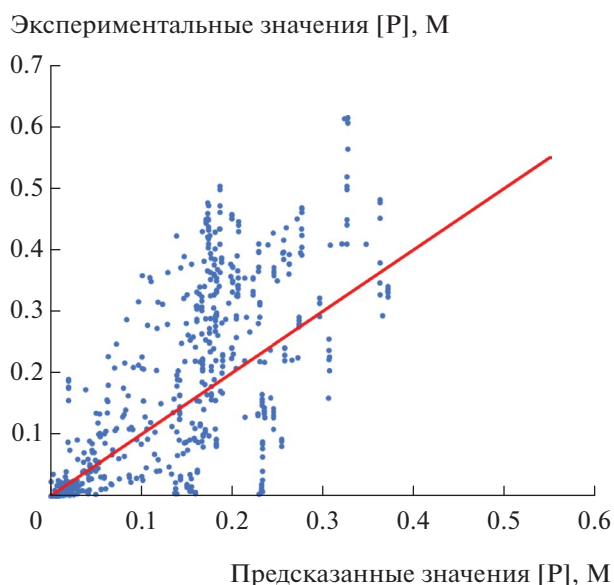


Рис. 4. Зависимость экспериментально определенных значений концентрации продукта реакции Сузуки—Мияуры (схема 1) от рассчитанных по уравнению (1) с использованием параметров уравнения, предсказанных НС.

ной способностью. Гибридный подход на основе комбинирования ансамблевого метода случайного леса и полносвязной НС позволил решить проблему создания модели МО для непрерывного диапазона условий проведения процесса в условиях ограниченного набора исходных данных.

Для обучения всех описанных выше моделей использовались датасеты, содержащие исключительно количественные параметры без включения категориальных переменных. В части компонентов реакционной системы это означает, что входные параметры датасета для каждого эксперимента были представлены полными наборами концентраций компонентов, включая нулевые значения количеств компонентов, которые отсутствовали в данном конкретном эксперименте. Такой подход перестает быть продуктивным, если иметь в виду большие датасеты (Big Data), подразумевающие учет гораздо большего числа компонентов химических систем, например, разных типов предшественников катализаторов, субстратов, добавок и т.п. В этом случае гораздо более эффективным видится применение подходов, которые вообще не накладывают ограничений на число типов компонентов, описываемых той или иной математической моделью. В этом случае неизбежен переход к используемым для МО датасетам, содержащим категориальные параметры. По сути, речь идет о привычных химикам названиях/свойствах веществ, закодированных определенным образом специальными численными идентификаторами в применяемом для МО датасете. Такие идентификаторы могут быть включены в датасет в качестве дескрипторов тех или иных свойств веществ в случае, если они оказываются значимыми для закономерностей протекания химического процесса. Как было отмечено выше, определение подходящего набора дескрипторов свойств компонентов, в том числе катализатора, представляет собой отдельную сложную задачу. Нами была предпринята попытка ее решения путем прямого включения поиска численных значений дескрипторов на этапе обучения модели.

В качестве предшественников катализатора в экспериментальном датасете использовался набор солей двухвалентного палладия. Необходимо отметить, что датасет с точки зрения набора предшественников катализатора являлся несбалансированным, поскольку из 60 включенных в него экспериментов 47 были проведены с применением одного и того же типа предшественника — PdCl_2 (в виде порошка). Соответственно,

в датасете присутствовали данные только 13 экспериментов, проведенных с 4 другими типами предшественников — $\text{Pd}(\text{Acac})_2$, $\text{Pd}(\text{OAc})_2$, PdBr_2 (все перечисленные в виде порошка) и предварительно приготовленного раствора PdCl_2 . Категоризация раствора и порошка дихлорида палладия как различных типов предшественника катализатора обусловлена существенными различиями экспериментально наблюдаемой кинетики реакции Сузуки—Мияуры в их присутствии — в случае предварительно растворенного PdCl_2 начальная скорость реакции оказывалась более высокой, что, однако, сопровождалось низкой стабильностью катализатора. Потеря каталитической активности происходила намного быстрее в сравнении с экспериментом в абсолютно аналогичных условиях, но с введением предшественника в виде порошка. Наблюдаемые закономерности подтверждают предположения о ключевой роли процессов дезактивации истинного катализатора в функционировании каталитических систем в реакции с арилхлоридами [35, 36].

С целью балансировки датасета на первоначальном этапе тренировки модели МС данные для 13-ти экспериментов (13 строк с наборами значений параметров реакции) с отличными от порошка PdCl_2 предшественниками катализатора были искусственно продублированы 14 раз для увеличения их значимости в процессе обучения модели. Применение модифицированного таким образом датасета для обучения приводило к уменьшению потерь на строках с этими предшественниками на ~10% относительно варианта обучения без дублирования строк, а также к уменьшению суммарного значения функции потерь $\text{MAE} = 0.113$ в сравнении с $\text{MAE} = 0.115$ без дублирования строк. После этого вместо 5 столбцов в исходном датасете с экспериментальными данными, содержащих значения концентраций каждого из 5 типов предшественников катализатора, вставляли столбец с концентрацией катализатора и столбец со значениями его дескриптора, выполняющего функцию закодированного категориального параметра. Тип катализатора обозначали числами от 1 до 5. Затем перебирали все комбинации наборов пар катализатор—дескриптор (всего 120 комбинаций) и для каждой комбинации обучали 60 моделей по описанному выше для МС алгоритму (60 моделей с ансамблем деревьев, для которых 59 экспериментов использовали в качестве обучающего множества и 1 — в качестве тестового). Критерием качества в этом случае вновь являлся минимум MAE по

Таблица 1. Численные значения и интервалы значимости дескрипторов* палладиевых предшественников катализаторов, определенные в результате обучения полносвязной НС

Предшественник катализатора	Численное значение дескриптора	Интервал значимости дескриптора
PdCl_2 (порошок)	0.481	0.474–0.487
PdCl_2 (раствор)	0.382	0.369–0.393
$\text{Pd}(\text{Acac})_2$	0.356	0.352–0.365
$\text{Pd}(\text{OAc})_2$	0.381	0.361–0.392
PdBr_2	0.368	0.347–0.375

*См. пояснения в тексте.

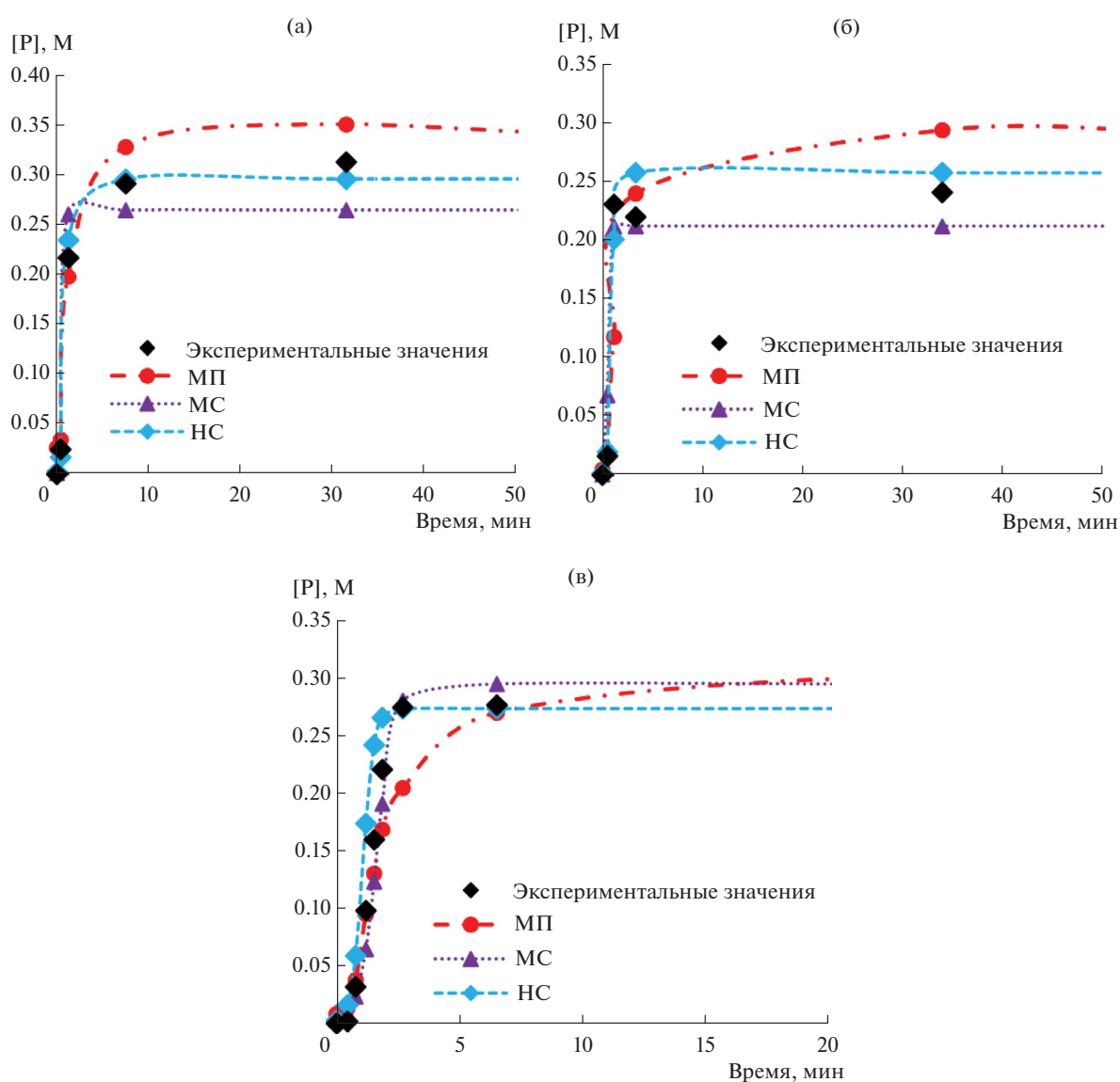


Рис. 5. Примеры экспериментальных и предсказанных интегральных кинетических данных (а–в) о накоплении продукта реакции Сузуки–Мияуры (схема 1) в различных начальных условиях (природа и концентрация основания, субстрата, катализатора). Расчетные значения получены с помощью модели МП, а также моделей МС и НС, использующих дескрипторы катализаторов.

значениям концентраций продукта по окончании реакции. Минимальный MAE был получен на модели, обученной со следующим порядком дескрипторов предшественников катализатора: PdCl_2 (раствор) > PdCl_2 (порошок) > PdBr_2 > $\text{Pd}(\text{Acac})_2$ > $\text{Pd}(\text{OAc})_2$ (от 5 до 1 соответственно). Многократное повторение указанной последовательности действий при обучении приводило к аналогичной последовательности значений дескрипторов, свидетельствуя об устойчивости найденного решения.

Определенный в результате работы модели МС порядок дескрипторов был включен в модель, которая далее была применена для генерирования искусственных данных для обучения НС аналогично описанной выше процедуре — 10 000 случайных комбинаций исходных параметров, составленных из встречающихся в датасете с экспериментальными данными значений, использовались для предсказаний величин концентраций продукта реакции Сузуки—Мияуры с применением модели МС, включающей набор дескрипторов. Полученные данные, в том числе численные значения дескрипторов для каждого типа предшественника катализатора, использовали в качестве обучающего датасета для полносвязной НС. Процедуру обучения проводили многократно с перебором значений дескрипторов с шагом 0.001, отслеживая после каждого цикла величину метрики качества (MAE). Итоговый набор значений дескрипторов, представленный в табл. 1, был определен по наименьшей величине метрики (MAE = 0.0372).

Для примерной оценки точности расчета значений дескрипторов нейронной сетью нами была оценена чувствительность MAE к их возможному изменению, которая оказалась достаточно высокой — величина MAE увеличивалась на 0.01 при варьировании величин дескрипторов в диапазонах, указанных в табл. 1. Определенные таким образом интервалы значений дескрипторов оказывались довольно узкими, позволяя сделать заключение об их значимости при расчете MAE, а, следовательно, о точности их определения в процессе обучения НС в рамках используемой модели. Таким образом, определенные в результате тренировки НС значения дескрипторов позволяют удовлетворительно описывать интегральные кинетические кривые накопления продукта реакции Сузуки—Мияуры, в том числе с применением предшественников катализатора, присутствующих в ограниченном количестве экспериментов в исходном датасете,

используемом для первичного обучения всех описанных в работе моделей. На рис. 5 даны примеры экспериментальных и предсказанных интегральных кинетических данных о накоплении продукта реакции Сузуки—Мияуры, найденных в результате расчета ансамблевыми методами по обученным на кинетических данных моделям МП и МС, а также по модели НС, полученной при использовании гибридного подхода (решающие деревья, нейронная сеть).

ЗАКЛЮЧЕНИЕ

Представленные в работе результаты показывают, что интегральные кинетические данные о закономерностях протекания реакции во времени могут быть эффективно использованы для создания моделей МО, позволяющих, в отличие от большинства описанных в литературе моделей МО для каталитических реакций, не только удовлетворительно предсказывать величины итоговых значений выходов продуктов по окончании реакции, но и кинетические профили реакций, включая такие ключевые для любого каталитического процесса характеристики, как величина максимальной скорости, протяженность периода индукции/автоускорения, стабильность работы катализатора и, в конечном итоге, время реакции. Эти характеристики являются принципиально важными для управления химическими процессами, в том числе, оптимизации их технологических параметров (например, производительностью реакторов), а также экологических параметров “зеленой” химии (E-factor [37] и иных производных метрик [38]). Разработка моделей МО на основе гибридного подхода, использующего ансамблевые методы для описания пространства ограниченного множества реальных экспериментальных данных, и нейронные сети, обученные на сгенерированных ансамблевыми моделями *in silico* данных большого объема, позволила предсказывать характер кинетических кривых накопления продукта реакции Сузуки—Мияуры с арилхлоридом в условиях применения простейших “безлигандных” каталитических систем на основе солей двухвалентного палладия, для которых характерна повышенная чувствительность кинетики реакции к малым изменениям условий процесса. Кроме того, с использованием описанных подходов удалось решить принципиальную задачу определения дескрипторов предшественников катализатора посредством включения их значений в число тренируемых параметров мо-

делей МО на этапе их обучения. Применение разработанных моделей МО, предсказывающих кинетическое поведение каталитической системы, позволит существенно сократить временные и материальные затраты при планировании экспериментов, направленных на оптимизацию экспериментальных протоколов для повышения эффективности каталитических процессов.

ФИНАНСИРОВАНИЕ

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 24-23-00382) с использованием оборудования Центра коллективного пользования аналитическим оборудованием ИГУ (<http://ckp-rf.ru/ckp/3264/>).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов, требующего раскрытия в данной статье.

СПИСОК ЛИТЕРАТУРЫ

1. *Fitzner M., Wuitschik G., Koller R., Adam J.-M., Schindler T.* // ACS Omega. 2023. V. 8 № 3. P. 3017.
2. *Davies J.C., Pattison D., Hirst J.D.* // J. Mol. Graph. Model. 2023. V. 118. P. 108356.
3. *Zahrt A.F., Henle J.J., Denmark S.E.* // ACS Comb. Sci. 2020. V. 22. № 11. P. 586.
4. *Burés J., Larrosa I.* // Nature. 2023. V. 613. № 7945. P. 689.
5. *Freeze J.G., Kelly H.R., Batista V.S.* // Chem. Rev. 2019. V. 119. № 11. P. 6595.
6. *Orlandi M., Escudero-Casao M., Licini G.* // J. Org. Chem. 2021. V. 86. № 4. P. 3555.
7. *Adebar N., Keupp J., Emenike V.N., Kühlborn J., Vom Dahl L., Möckel R., Smiatek J.* // J. Phys. Chem. A. 2024. V. 128. № 5. P. 929.
8. *Clarke G.E., Firth J.D., Ledingham L.A., Horbaczewskyj C.S., Bourne R.A., Bray J.T.W., Martin P.L., Eastwood J.B., Campbell R., Pagett A., MacQuarrie D.J., Slattery J.M., Lynam J.M., Whitwood A.C., Milani J., Hart S., Wilson J., Fairlamb I.J.S.* // Nat. Commun. 2024. V. 15. № 1. Art. 3968.
9. *Yada A., Matsumura T., Ando Y., Nagata K., Ichinoseki S., Sato K.* // Synlett. 2021. V. 32. № 18. P. 1843.
10. *Gensch T., Dos Passos Gomes G., Friederich P., Peters E., Gaudin T., Pollice R., Jorner K., Nigam A., Lindner-D'Addario M., Sigman M. S., Aspuru-Guzik A.* // J. Am. Chem. Soc. 2022. V. 144. № 3. P. 1205.
11. *Mou L.-H., Han T., Smith P.E.S., Sharman E., Jiang J.* // Adv. Sci. 2023. V. 10. № 22. Art. 2301020.
12. *Skoraczynski G., Dittwald P., Miasojedow B., Szymkuć S., Gajewska E.P., Grzybowski, B.A., Gambin A.* // Sci. Rep. 2017. V. 7. № 1. P. 3582.
13. *Raghavan P., Haas B.C., Ruos M.E., Schleinitz J., Doyle A.G., Reisman S.E., Sigman M.S., Coley C.W.* // ACS Cent. Sci. 2023. V. 9. № 12. P. 2196.
14. *Kalikadien A.V., Mirza A., Hossaini A.N., Sreenithya A., Pidko E.A.* // ChemPlusChem. 2024. V. 89. № 7. Art. e202300702.
15. *Chan K., Ta L.T., Huang Y., Su H., Lin Z.* // Molecules. 2023. V. 28. № 12. Art. 4730.
16. *Jorner K., Brinck T., Norrby P.-O., Buttar D.* // Chem. Sci. 2021. V. 12. № 3. P. 1163.
17. *Gladstone J.H.* // Sci. Am. 2023. V. 16. № 408 supp. P. 6511.
18. *Cammarota R.C., Liu W., Bacsá J., Davies H.M.L., Sigman M.S.* // J. Am. Chem. Soc. 2022. V. 144. № 4. P. 1881.
19. *Lustosa D.M., Milo A.* // ACS Catal. 2022. V. 12. № 13. P. 7886.
20. *Durand D.J., Fe N.* // Chem. Rev. 2019. V. 119. № 11. P. 6561.
21. *Caldeweyher E., Elkin M., Gheibi G., Johansson M., Sköld C., Norrby P.-O., Hartwig J.F.* // J. Am. Chem. Soc. 2023. V. 145. № 31. P. 17367.
22. *Taniike T., Fujiwara A., Nakanowatari S., García-Escobar F., Takahashi K.* // Commun. Chem. 2024. V. 7. № 1. P. 11.
23. *Ebi T., Sen A., Dhital R.N., Yamada Y.M.A., Kaneko H.* // ACS Omega. 2021. V. 6. № 41. P. 27578.
24. *Suzuki A.* // Angew. Chem. Int. Ed. 2011. V. 50. P. 6722.
25. *Новаковский А.Б.* // Вестник ИБ Коми НЦ УрО РАН. 2018. С. 26. (Novakovskiy A.B. // Vestnik of Institute of Biology Komi Scientific Centre of the Ural Branch of RAS. 2018. P. 26.)
26. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* // J. Mach. Learn. Res. 2011. V. 12. № 85. P. 2825.
27. *Grebennikov N.O., Boiko D.A., Prima D.O., Madiyeva M., Minyaev M.E., Ananikov V.P.* // J. Catal. 2024. V. 429. Art. 115240.
28. *Biffis A., Centomo P., Del Zotto A., Zecca M.* // Chem. Rev. 2018. V. 118. № 4. P. 2249.
29. *Widegren J.A., Finke R.G.* // J. Mol. Catal. A: Chem. 2003. V. 198. № 1–2. P. 317.
30. *Schmidt A.F., Al Halaiqa A., Smirnov V.V.* // Synlett. 2006. № 18. P. 2861.
31. *Shim E., Tewari A., Cernak T., Zimmerman P.M.* // J. Chem. Inf. Model. 2023. V. 63. № 12. P. 3659.
32. *Schmidt A.F., Al-Halaiqa A., Smirnov V.V.* // J. Mol. Catal. A: Chem. 2006. V. 250. № 1–2. P. 131.

33. Schmidt A.F., Kurokhtina A.A., Larina E.V., Vidyaeva E.V., Lagoda N.A. // Mol. Catal. 2021. V. 499. Art. 111321.
34. Курохтина А.А., Ларина Е.В., Лагода Н.А., Шмидт А.Ф. // Кинетика и катализ. 2022. Т. 63. № 5. С. 614. (Kurokhtina A.A., Larina E.V., Lagoda N.A., Schmidt A.F. // Kinet. Catal. 2022. V. 63. № 5. P. 543.)
35. Galushko A.S., Prima D.O., Burykina J.V., Ananikov V.P. // Inorg. Chem. Front. 2021. V. 8. № 3. P. 620.
36. Schmidt A.F., Kurokhtina A.A., Larina E.V., Lagoda N.A. // Organometallics. 2023. V. 42. № 24. P. 3442.
37. Sheldon R.A. // ACS Sustain. Chem. Eng. 2018. V. 6. № 1. P. 32.
38. Mercer S.M., Andraos J., Jessop P.G. // J. Chem. Educ. 2012. V. 89. № 2. P. 215.

The Determination of Descriptors for Catalytic Systems in Machine Learning Models Using Kinetic Experimental Data

A. F. Schmidt^{1, *}, N. A. Sidorov¹, A. A. Kurokhtina¹, E. V. Larina¹, N. A. Lagoda¹

¹Irkutsk State University, Chemical Department,
K. Marx str., 1, Irkutsk, 664003 Russia

*e-mail: aschmidt@chem.isu.ru

The problem of selection and determining the values of descriptors for the properties of chemical reactions components in mathematical models for chemical processes is one of the essential ones when creating machine learning (ML) models used to describe and predict the functioning patterns of chemical systems. Current practice in the field mainly involves the use as the descriptors physical and chemical characteristics of the components of reaction systems (ionic radii, bond lengths, energies, and other parameters related to the structure and properties of specific molecules or particles) determined experimentally or by calculation. This work presents the results of the predicting of the integral kinetic dependences, as well as approaches to determine the values of descriptors for characterizing the properties of a set of simple palladium catalyst precursors when used in the Suzuki–Miyaura reaction. The problem stated has been solved by creating the ML models that take into account experimental kinetic data. The descriptors obtained as a result of training the models make it possible to satisfactorily describe the kinetic patterns of the Suzuki–Miyaura reaction with aryl chlorides under the so-called “ligand-free” catalytic conditions possessing higher sensitivity of the reaction to small changes in the conditions.

Keywords: catalysis, machine learning, Suzuki–Miyaura reaction, palladium, kinetics