

Федеральное государственное бюджетное учреждение науки
Институт общей и неорганической химии им. Н.С. Курнакова
Российской академии наук
(ИОНХ РАН)

На правах рукописи

Краснов Лев Валерьевич

**Создание баз данных и разработка моделей машинного обучения для
прогнозирования люминесцентных и противоопухолевых свойств
комплексов переходных металлов и растворимости органических
соединений**

1.4.5 - Хемоинформатика

Диссертация на соискание ученой степени
кандидата химических наук

Научный руководитель:
кандидат химических наук
Беззубов Станислав Игоревич

Москва - 2026 г.

Оглавление

1. Введение	5
Обозначения и сокращения.....	12
2. Обзор литературы	13
2.1. Данные как узкое место хемоинформатики	13
2.1.1. Смена узкого места: от алгоритмов к данным	13
2.1.2. Источники экспериментальных данных и способы их извлечения.....	14
2.1.3. Типовые дефекты химических датасетов	16
2.1.4. Требования к данным: FAIR, машиночитаемость, воспроизводимость..	19
2.1.5. Обзор существующих ресурсов химических данных	21
2.2. Прогнозирование растворимости органических соединений.....	23
2.2.1. Задача: величины, единицы, методы измерения	23
2.2.2. Растворимость в воде.....	29
2.2.3. Растворимость в органических растворителях	34
2.2.4. Бинарные смеси растворителей.....	38
2.2.5. Физико-химические методы как альтернатива	41
2.2.6. Выводы по разделу	43
2.3. Координационные соединения переходных металлов: представление и наборы данных	46
2.3.1. Машиночитаемое представление координационных соединений.....	46
2.3.2. Наборы данных по комплексам переходных металлов.....	49
2.4. Люминесцентные свойства циклометаллированных комплексов иридия(III)	54
2.4.1. Фотофизические свойства и их настройка структурой лигандов	54
2.4.2. Прогнозирование фотофизических свойств.....	56
2.5. Цитотоксичность комплексов переходных металлов.....	58
2.5.1. Противоопухолевая активность комплексов	58
2.5.2. Прогнозирование цитотоксичности	64
2.6. Заключительные замечания и постановка задачи	66
3. Материалы и методы	68
3.1. Сбор и стандартизация экспериментальных данных.....	68
3.1.1. Отбор публикаций и критерии включения.....	69
3.1.2. Извлечение данных из первичных источников	71
3.1.3. Стандартизация структур, идентификаторов и единиц	72

3.1.4.	Верификация и контроль качества.....	74
3.1.5.	Структура наборов и депонирование.....	75
3.2.	Представление соединений	76
3.2.1.	Органические соединения и растворители.....	76
3.2.2.	Комплексы иридия(III) по структурным формулам лигандов	77
3.3.	Модели машинного обучения	78
3.3.1.	Алгоритмы и подбор гиперпараметров	78
3.3.2.	Схемы разбиения выборки и валидация.....	79
3.3.3.	Программные средства и вычислительные ресурсы	81
3.4.	Синтез и физико-химические измерения.....	81
3.4.1.	Исходные вещества.....	81
3.4.2.	Методы исследования	82
3.4.3.	Синтез комплексов иридия(III).....	83
3.4.4.	Измерение спектров люминесценции.....	89
3.4.5.	Квантово-химические расчеты	90
3.5.	Интерактивные онлайн-инструменты	90
4.	Обсуждение результатов	91
4.1.	Базы данных растворимости органических соединений.....	91
4.1.1.	BigSolDB 2.0.....	91
4.1.2.	MixtureSolDB.....	95
4.1.3.	Температура как полноправная переменная	100
4.1.4.	Внутренняя согласованность собранных данных.....	102
4.1.5.	Сведения о методике измерения и твердой фазе	104
4.2.	Люминесцентные свойства комплексов иридия(III).....	109
4.2.1.	База данных IrLumDB.....	109
4.2.2.	Модели прогнозирования длины волны максимума испускания и квантового выхода	114
4.2.3.	Экспериментальная проверка на синтезированных комплексах	119
4.3.	Цитотоксичность комплексов переходных металлов.....	123
4.3.1.	Базы данных IrCytoToxDB и MetalCytoToxDB.....	123
4.3.2.	Модели классификации по уровню цитотоксичности	141
4.3.3.	Представленность комплексов металлов в ресурсах данных о биологической активности.....	153

4.4.	Публикация наборов данных и интерактивные онлайн-инструменты	154
4.5.	Использование созданных наборов данных сторонними исследователями	157
5.	Выводы.....	163
6.	Список цитируемой литературы.....	165
7.	Список публикаций, в которых изложены основные результаты диссертации .	180
8.	Благодарности	180

1. Введение

Актуальность работы и степень разработанности темы исследования.

Растворимость является одним из ключевых физико-химических свойств органических соединений, определяющих их применимость в синтетической, медицинской и промышленной химии. Ввиду трудоемкости экспериментального определения растворимости для больших библиотек соединений разработка методов ее прогнозирования по молекулярной структуре является важной задачей современной хемоинформатики. Наибольшее число исследований посвящено растворимости в воде, благодаря наличию больших и аннотированных датасетов, таких как AqSolDB. Прогнозирование растворимости в органических растворителях развивается менее интенсивно, а ключевым ограничением остается нехватка крупных и разнородных наборов данных. В еще большей мере это справедливо для растворимости в смесях растворителей, которые широко используются в технологических процессах: существующие публичные базы данных в этой области невелики по объему и охватывают ограниченное химическое пространство.

Наряду с задачами прогнозирования свойств органических соединений, актуальным направлением хемоинформатики является разработка методов машинного обучения (ML) для комплексов переходных металлов. Циклометаллированные комплексы иридия(III) широко применяются в качестве эмиттеров в фосфоресцентных органических светодиодах и агентов противоопухолевой терапии. Разработка новых эффективных иридиевых комплексов с необходимыми свойствами требует значительных экспериментальных усилий, тогда как прогностическая точность квантово-химических расчетов для таких соединений остается недостаточной для задач рационального молекулярного дизайна. Применение методов машинного обучения для прогнозирования фотофизических свойств комплексов иридия ограничивалось либо узкими классами соединений, либо использованием расчетных, а не экспериментальных данных в качестве целевых переменных. Систематической экспериментальной базы данных люминесцентных свойств структурно разнообразных бис-циклометаллированных комплексов иридия(III) в машиночитаемом формате до настоящего времени создано не было.

Другим важным применением комплексов переходных металлов является противоопухолевая активность. Соединения рутения, иридия, осмия, родия и рения активно исследуются как альтернатива препаратам платины, а ряд комплексов рутения, таких как NAMI-A, KP1019/KP1339, достиг стадии клинических испытаний. Цитотоксичность является одной из ключевых количественных характеристик при первичном скрининге потенциальных металлолекарств, однако ее экспериментальное определение для больших библиотек соединений остается трудоемким и дорогостоящим процессом. Применение методов машинного обучения для прогнозирования цитотоксичности хорошо развито для органических молекул, однако комплексы переходных металлов традиционно остаются недостаточно представлены в этой области вследствие трудностей в их машиночитаемом представлении и отсутствия крупных систематизированных баз данных.

Цель работы: восполнение дефицита структурированных экспериментальных данных в областях прогнозирования фотофизических и противоопухолевых свойств комплексов переходных металлов и растворимости органических соединений в неводных растворителях путем создания баз данных и разработки моделей машинного обучения на их основе.

Для достижения указанной цели поставлены следующие **задачи**:

1. Систематический сбор, стандартизация и верификация экспериментальных данных о люминесцентных свойствах бис-циклометаллированных комплексов иридия(III) из рецензируемых научных публикаций.
2. Обучение и валидация моделей машинного обучения для прогнозирования люминесцентных свойств (длины волны максимума и квантового выхода) комплексов иридия(III).
3. Систематический сбор, стандартизация и верификация экспериментальных данных о цитотоксичности комплексов переходных металлов из рецензируемых научных публикаций.

4. Обучение и валидация моделей машинного обучения для классификации комплексов переходных металлов по уровню цитотоксичности против раковых клеточных линий.
5. Систематический сбор, стандартизация и верификация экспериментальных данных о растворимости органических соединений (неэлектролиты) в индивидуальных органических растворителях и бинарных смесях растворителей (вода-органический растворитель или пара органических растворителей) из рецензируемых научных публикаций.
6. Создание интерактивных онлайн-инструментов для визуализации и навигации по созданным базам данных.

Научная новизна работы.

Впервые создана база данных люминесцентных свойств бис-циклометаллированных комплексов иридия(III) – IrLumDB, состоящая из 1287 соединений с разнообразным лигандным окружением и на ее основе разработаны ML-модели прогнозирования длины волны максимума эмиссии (λ_{max}) и квантового выхода (PLQY), требующие только SMILES лигандов. Прогноз длины волны максимума эмиссии проверен экспериментально на выборке из 33 спектров комплексов иридия(III), двенадцать из которых впервые получены в настоящей работе.

Впервые созданы открытые базы данных цитотоксичности комплексов переходных металлов: IrCytoToxDB, содержащая 2694 значения IC_{50} для 803 комплексов иридия(III) против 127 клеточных линий, и расширенная MetalCytoToxDB, охватывающая 26500 значений IC_{50} для 7050 комплексов Ru, Ir, Os, Re и Rh против 754 клеточных линий из 1921 статьи. На основе MetalCytoToxDB впервые разработаны ML-модели для разделения комплексов переходных металлов на высоко- и низкоцитотоксичные против раковых клеточных линий (HeLa, A549, MCF-7 и др.), обеспечивающие прогнозы для металлов с ограниченным объемом экспериментальных данных за счет переноса знаний между подмножествами комплексов различных металлов.

Создана крупнейшая на момент публикации открытая база данных растворимости индивидуальных органических соединений (неэлектролиты) в неводных растворителях BigSolDB 2.0, содержащая свыше 100 тысяч экспериментальных значений для 1448 соединений в 213 растворителях из 1595 статей, восполняющая дефицит данных для разработки моделей машинного обучения для прогнозирования растворимости.

Создана крупнейшая на момент публикации открытая база данных растворимости органических соединений (неэлектролиты) в бинарных смесях растворителей MixtureSolDB, содержащая свыше 175 тысяч экспериментальных значений для 810 соединений в 750 уникальных смесях (вода-органический растворитель или пара органических растворителей) из 1115 статей в стандартизованном машиночитаемом формате, пригодном для разработки моделей машинного обучения.

Теоретическая и практическая значимость работы.

Базы данных BigSolDB 2.0 и MixtureSolDB формируют наиболее полную на сегодняшний день основу для разработки моделей машинного обучения для прогнозирования растворимости органических соединений в индивидуальных органических растворителях и их бинарных смесях (в том числе с водой), что имеет прямое значение для оптимизации технологических процессов в фармацевтической и химической промышленности. Обе базы данных размещены в открытом доступе и опубликованы в журнале Scientific Data. За период с момента публикации датасет BigSolDB набрал более 16000 просмотров статьи и 4700 скачиваний с Zenodo, а также использовался различными научными группами и компаниями (MIT, Los Alamos National Laboratory, University of Strathclyde, Loughborough University, Case Western Reserve University, Zhejiang University, Imperial College London, Sanofi, GSK и др.) в ряде работ по прогнозированию растворимости в органических растворителях.

Созданные базы данных IrLumDB, IrCytoToxDB и MetalCytoToxDB представляют собой систематизированные ресурсы экспериментальных данных о свойствах комплексов переходных металлов, открывающие возможности для количественного анализа зависимостей структура-свойство в этом классе соединений.

Разработанная на основе IrLumDB модель прогнозирования длины волны максимума эмиссии прошла экспериментальную проверку на 33 спектрах комплексов иридия(III), включая двенадцать вновь синтезированных, и позволяет проводить быстрый высокопроизводительный скрининг потенциальных люминесцентных эмиттеров для OLED-устройств без проведения квантово-химических расчетов.

Разработанные ML-модели на основе MetalCytoToxDB обеспечивают предварительную оценку цитотоксичности новых комплексов переходных металлов в отношении раковых клеточных линий, что может ускорить отбор перспективных противоопухолевых агентов на ранних этапах молекулярного дизайна.

Положения, выносимые на защиту.

1. Унифицированная методология систематического сбора, стандартизации и верификации экспериментальных данных из рецензируемых научных публикаций, апробированная на массиве более 4000 источников и применяемая как к органическим соединениям, так и к комплексам переходных металлов.
2. Базы данных растворимости индивидуальных органических соединений (неэлектролиты) BigSolDB 2.0 и MixtureSolDB в стандартизованном машиночитаемом формате, содержащие свыше 100 тысяч и 175 тысяч экспериментальных значений для индивидуальных органических растворителей и бинарных смесей (вода-органический растворитель или пара органических растворителей), соответственно.
3. База данных люминесцентных свойств бис-циклометаллированных комплексов иридия(III) IrLumDB, содержащая экспериментальные данные для 1287 соединений с разнообразным лигандным окружением и ML-модели для прогнозирования длины волны максимума эмиссии и квантового выхода.
4. Базы данных цитотоксичности комплексов переходных металлов IrCytoToxDB и MetalCytoToxDB, содержащие 2694 и 26500 значений IC_{50} для комплексов иридия(III) и комплексов Ru, Ir, Os, Re, Rh, соответственно, а также ML-модели бинарной классификации комплексов по цитотоксичности против раковых клеточных линий с возможностью переноса знаний между комплексами разных металлов в пределах датасета.

Степень достоверности результатов, методы и методология диссертационного исследования.

Достоверность и обоснованность научных положений и выводов, сформулированных в диссертации, определяется использованием большого массива экспериментальных данных, извлеченных из рецензируемых научных публикаций по разработанной автором методологии. Для минимизации ошибок извлечения применялась процедура перекрестной проверки. Два независимых исполнителя раздельно собирали данные из первичных источников, после чего третий исполнитель сверял и объединял полученные массивы. Корректность молекулярных структур обеспечивалась стандартизацией SMILES-представлений средствами RDKit, верификацией источников через CrossRef API и проверкой физической допустимости значений. Достоверность разработанных моделей машинного обучения подтверждается применением кросс-валидации, внешней валидацией на независимых тестовых выборках, а также адекватностью полученных метрик качества. Для модели прогнозирования длины волны максимума эмиссии дополнительно проведена экспериментальная проверка на спектрах вновь синтезированных комплексов иридия(III). Состав и строение полученных соединений достоверно определены с помощью комплекса высокоточных и взаимодополняющих физико-химических методов анализа (рентгеноструктурный анализ, рентгенофазовый анализ, ^1H ЯМР-спектроскопия, масс-спектрометрия высокого разрешения), данные которых хорошо воспроизводятся и согласуются между собой. Сформулированные выводы и результаты научно обоснованы и соответствуют современным представлениям в области хемоинформатики и машинного обучения в химии.

Апробация результатов.

Результаты настоящей работы были представлены на XIV, XV Конференциях молодых ученых по общей и неорганической химии (г. Москва, ИОНХ РАН, 2024, 2025), Всероссийской научной конференции “Николаевские чтения 2025” (г. Новосибирск, 2025), 1-я Научно-практической конференции “Дисплейные материалы и технологии” (г. Москва, 2025), XXIX Международной Чугаевской конференции по координационной химии (г. Казань, 2025), II Научной конференции “Искусственный интеллект в химии и материаловедении” (г. Москва, 2025), Конференции Data Fusion (г. Москва, 2026) и AI4X - Accelerate Conference (Сингапур, 2026).

Работа выполнена в рамках реализации проекта Российского научного фонда № 24-73-10232, а также в соответствии с государственным заданием ИОНХ РАН.

Публикации по теме исследования.

Основные результаты работы опубликованы в 5 статьях в научных журналах, рекомендованных ВАК РФ, индексируемых в базах данных Web of Science и Scopus, а также в виде 6 тезисов докладов на научных конференциях.

Личный вклад автора.

Автором выполнены: обзор литературы, постановка цели и задач исследования, разработка методологии сбора и стандартизации данных, разработка схем баз данных, написание программного кода для обработки, очистки и постобработки данных, разработка и обучение моделей машинного обучения, анализ важности признаков методом SHAP, разработка онлайн-инструментов для визуализации и навигации по базам данных, анализ и интерпретация результатов, формулировка выводов; синтез комплексов иридия(III) для экспериментальной валидации ML-моделей. Извлечение экспериментальных данных из первичных источников проводилось совместно с соавторами по разработанной автором методологии. Подготовка публикаций по теме диссертационной работы проводилась совместно с научным руководителем и соавторами.

Соответствие паспорту научной специальности.

Диссертационная работа соответствует пп. 2, 3, 9 паспорта специальности 1.4.5 - Хемоинформатика (химические науки).

Объем и структура диссертационной работы.

Диссертация изложена на 180 страницах и содержит 56 рисунков и 15 таблиц. Она состоит из введения, литературного обзора, экспериментальной части, обсуждения результатов, выводов, списка цитируемой литературы (204 наименования), списка публикаций, в которых изложены основные результаты диссертации, и благодарностей.

Обозначения и сокращения

ВЗМО — высшая занятая молекулярная орбиталь

ВЭЖХ — высокоэффективная жидкостная хроматография

НСМО — низшая свободная молекулярная орбиталь

ЯМР — ядерный магнитный резонанс

CAS — регистрационный номер вещества в реестре Chemical Abstracts Service

COSMO-RS — conductor-like screening model for real solvents, модель расчета термодинамических свойств растворов

DOI — digital object identifier, цифровой идентификатор объекта

ESI — electrospray ionization, ионизация электрораспылением

FAIR — findable, accessible, interoperable, reusable; находимость, доступность, совместимость и пригодность к повторному использованию

FDA — Food and Drug Administration, Управление по санитарному надзору за качеством пищевых продуктов и медикаментов США

GSE — general solubility equation, обобщенное уравнение растворимости

HRMS — high-resolution mass spectrometry, масс-спектрометрия высокого разрешения

IC₅₀ — концентрация вещества, вызывающая полумаксимальное ингибирование

InChI — international chemical identifier, международный химический идентификатор

LogP — десятичный логарифм коэффициента распределения в системе н-октанол-вода

LogS — десятичный логарифм растворимости

MAE — mean absolute error, средняя абсолютная ошибка

ML — machine learning, машинное обучение

pIC₅₀ — величина 6 - lg IC₅₀ при IC₅₀, выраженной в мкмоль/л

PLQY — photoluminescence quantum yield, квантовый выход фотолюминесценции

RMSE — root mean square error, среднеквадратичная ошибка

ROC-AUC — area under the receiver operating characteristic curve, площадь под ROC-кривой

SHAP — shapley additive explanations, метод оценки вклада признаков в прогноз

SMILES — simplified molecular input line entry system, линейная форма записи структуры

TPSA — topological polar surface area, площадь топологической полярной поверхности

2. Обзор литературы

2.1. Данные как узкое место хемоинформатики

2.1.1. Смена узкого места: от алгоритмов к данным

Прогнозирование свойств соединений по их структуре развивается с 1960-х годов и прошло путь от линейных соотношений с несколькими физико-химическими дескрипторами до современных методов машинного обучения, обзор которого дан в работе[1]. За последние два десятилетия это развитие сопровождалось быстрым усложнением архитектур - от ансамблей решающих деревьев к нейронным сетям, затем к графовым сетям с передачей сообщений, наиболее употребительной реализацией которых является Chemprop[2], и далее к моделям на основе механизма внимания. Каждое поколение сопровождалось сообщениями об улучшении результатов на стандартных наборах, вследствие чего именно выбор архитектуры длительное время считался основным фактором качества прогнозирования, а статистические методы стали неотъемлемой частью исследовательской практики в химии[3].

Накопленный опыт, однако, эту картину не подтверждает. Систематические сопоставления методов на единых данных показывают, что различия между архитектурами сопоставимы по величине с разбросом от выбора схемы разбиения выборки, а простейшие подходы нередко не уступают современным. Так, при сопоставлении методов прогнозирования активности соединений в отношении биологических мишеней простой анализ ближайших соседей достиг точности сложных моделей, включая глубокое обучение[4], что было подтверждено затем на более широком наборе архитектур[4], а в задаче прогнозирования растворимости органических соединений в органических растворителях восемь различных алгоритмов дали практически совпадающие результаты на уровне шума обучающих данных[5]. Наиболее масштабная проверка выполнена на материале 26 задач, охватывающих свыше 165 тысяч экспериментальных значений, при трех схемах разбиения выборки и 156 попарных сравнениях. Наибольшую долю наилучших результатов показали классические методы машинного обучения (47.4 %), далее следуют предобученные последовательностные модели (28.8 %), графовые нейронные сети (21.8 %) и подходы на основе больших языковых моделей (1.9 %)[6].

Показательно, что это ограничение признается и самими разработчиками сложных моделей. Дообученная большая языковая модель решает задачи прогнозирования свойств сопоставимо со специализированными моделями, а превосходит их преимущественно при малом объеме обучающих данных, причем и само развитие методов, встраивающих химическое знание в архитектуру, авторы связывают именно с малыми размерами характерных для химии наборов данных[7]. К тому же выводу приходят и разработчики моделей глубокого обучения: констатируя неспособность глубокого обучения превзойти классические подходы на наборах ограниченного объема, они предлагают предварительное обучение на вычисляемых молекулярных дескрипторах, и построенная таким образом базовая модель CheMeleon позволяет графовой сети с передачей сообщений превзойти классические методы в области малых выборок[8]. Недостаток экспериментальных данных восполняется здесь расчетными величинами - прием, который встретится далее неоднократно.

Одновременно выяснилось, что работа с данными дает прирост, превышающий эффект от смены модели. Тщательное курирование обучающей выборки улучшает качество прогнозирования сильнее, чем переход к более сложной архитектуре[9], причем улучшение достигается даже при сокращении объема выборки за счет удаления недостоверных записей[10]. Приведение значений к единым условиям измерения способно снизить межлабораторный разброс в несколько раз без проведения каких-либо новых экспериментов[11].

Таким образом, узкое место сместилось. Определяющим фактором стали объем, однородность и достоверность экспериментальных данных, на которых обучаются модели. Это смещение и определяет постановку задач настоящей работы, в которой основное содержание составляет создание баз данных, а разработка моделей выступает средством проверки их пригодности.

2.1.2. Источники экспериментальных данных и способы их извлечения

Экспериментальные данные, пригодные для обучения моделей, могут быть получены тремя принципиально различными путями, каждый из которых характеризуется своим соотношением между объемом, точностью и прослеживаемостью.

Первый путь состоит в агрегации уже существующих электронных ресурсов. Он наиболее производителен. Объединение нескольких открытых наборов позволяет за короткое время получить массив в десятки тысяч записей. Его недостатки, однако, систематичны и хорошо документированы. Объединяемые ресурсы, как правило, сами являются компиляциями, вследствие чего одно и то же экспериментальное значение, восходящее к единственному первоисточнику, попадает в итоговый массив многократно[10]. Сведения об условиях измерения, если они и присутствовали в первоисточнике, к моменту второй или третьей агрегации оказываются утраченными, что приводит к объединению в одном массиве величин, различающихся по физическому смыслу или полученных при несопоставимых условиях[11,12]. Наконец, ошибки в структурах и идентификаторах, допущенные на промежуточных этапах, наследуются производными наборами без возможности их обнаружения[13]. Отчасти эти затруднения преодолеваются присвоением объединяемым источникам весовых коэффициентов достоверности[9], однако такой прием предполагает, что качество источника известно, тогда как при последовательной агрегации оно как раз и утрачивается.

Второй путь состоит в автоматическом извлечении данных из текстов публикаций. Разработанные для этого инструменты позволяют распознавать химические названия, структуры и числовые значения свойств непосредственно в тексте и таблицах статей[14]. В последние годы для этой задачи все шире применяются большие языковые модели, обеспечивающие извлечение структурированной информации из научного текста без предварительной разметки[15]. Производительность такого подхода несопоставимо выше ручной обработки, однако он наследует все особенности исходных текстов и добавляет собственные. Числовые значения могут быть соотнесены не с тем веществом или не с той температурой, единицы измерения — интерпретированы неверно, а сведения, содержащиеся только в подписях к рисункам или в сопроводительных материалах, — пропущены. Полученный таким образом массив требует выборочной проверки, объем которой определяет фактическую трудоемкость метода.

Третий путь состоит в ручном курировании первичных публикаций - извлечении сведений непосредственно из статей с их проверкой и приведением к единому виду. Он наименее производителен, но обеспечивает то, чего не дают первые два. Каждое значение сопровождается указанием на работу, в которой оно получено, и потому может быть проверено независимо, а условия эксперимента извлекаются в том объеме, в каком они приведены авторами. Тем самым устраняются оба перечисленных выше источника систематических дефектов. Дублирование, возникающее при заимствовании из промежуточных компиляций, становится обнаружимым, а сведения об условиях измерения не утрачиваются. Ценой этого является производительность. Масштаб, при котором подход остается осуществимым, составляет несколько тысяч публикаций. Именно этот путь избран в настоящей работе.

Этот подход имеет длительную традицию. Критически оцененные компиляции экспериментальных данных издаются с 1960-х годов, причем каждое приводимое значение сопровождается заключением специалиста о его достоверности. Современная оцифрованная версия одной из таких компиляций, посвященной константам кислотной диссоциации кислот, содержит 24222 значения для 10564 соединений и сохраняет наряду с ними температуру, метод измерения и присвоенную оценку надежности[16]. В базе AQUASOL, описанной в 1989 году, каждое извлеченное из литературы значение оценивалось по пяти критериям с присвоением весового коэффициента[17]. Ревизия более чем восьмисот публикаций, посвященных определению равновесной растворимости, позволила установить пределы воспроизводимости литературных данных и предложить процедуру их нормировки[11]. Крупнейший набор экспериментальных значений коэффициента распределения октанол/вода (LogP) восходит к обзору литературы, охватившему свыше трех тысяч источников[18]. Аналогичным образом собраны и компиляции термодинамических характеристик плавления двухкомпонентных кристаллов, охватывающие публикации начиная с начала XX века[19]. Методические основания такой работы изложены в руководствах по курированию химических данных[20,21].

2.1.3. Типовые дефекты химических датасетов

Дефекты, характерные для наборов химических данных, систематически описаны в руководствах по курированию[20,21], где предложены и процедуры их выявления и исправления. Целесообразно разделить их на несколько классов, поскольку они по-разному влияют на обучаемые модели и требуют различных средств устранения.

Первый класс образуют дубликаты. Помимо очевидного случая повторного внесения одной записи, существует менее заметная разновидность. Одно и то же экспериментальное значение, восходящее к единственному первоисточнику, попадает в объединенный массив несколько раз, будучи независимо заимствованным разными компиляциями[10]. Без разбора до первоисточника такое дублирование неразличимо, а его последствия выходят за рамки простого искажения статистики. При случайном разбиении выборки идентичные записи оказываются одновременно в обучающей и тестовой частях, что напрямую завышает оценку качества модели. Этот дефект не является умозрительным. При подготовке набора данных о константах кислотной диссоциации кислот было установлено, что ранее опубликованные модели прогнозирования этой величины в большинстве своем не проходили надежной проверки именно вследствие перекрывания обучающей и тестовой выборок, в связи с чем потребовалось построение теста, свободного от такого наложения массивов[16].

Второй класс составляет объединение величин, полученных несопоставимыми способами. Наиболее выражен этот дефект в данных о биологической активности, где значения, измеренные в различных экспериментальных условиях, объединяются в единый массив под общим обозначением. Количественная оценка вносимого при этом шума получена сопоставлением измерений, выполненных для одних и тех же соединений в разных условиях. При минимальном отборе почти 65 % пар значений различаются более чем на 0.3 логарифмической единицы, 27 % — более чем на порядок, а ранговая корреляция между постановками составляет лишь 0.51. Требование совпадения основных характеристик эксперимента улучшает согласие (48 % и 13 % соответственно при корреляции 0.71), но существенно сокращает объем пригодных данных[12].

Третий класс образуют ошибки в химических идентификаторах и структурах: неверные регистрационные номера, расхождения между названием, номером и структурной формулой, нарушения валентности. Ревизия PHYSPROP, одной из ранних компиляций, послужившей основой для многих последующих наборов, выявила такие ошибки в качестве систематических и потребовала разработки отдельной процедуры автоматического курирования[13]. Особенность этого класса состоит в том, что ошибка относится не к измеренной величине, а к ее отнесению, и потому не может быть обнаружена анализом распределения значений. Устранение таких дефектов требует отдельной последовательности операций — проверки и нормализации структур, удаления противоионов, солей и смесей, приведения таутомерных форм и ароматических связей к единому виду, — выполняемых до какого-либо анализа данных[20].

Четвертый класс связан с отсутствием сведений об условиях эксперимента. Измеренное значение физико-химического или биологического свойства определено лишь совместно с условиями, при которых оно получено. При их утрате запись перестает соответствовать определенной физической величине. Проявления этого дефекта многообразны. Два крупнейших свободно доступных ресурса по растворимости в воде представляют собой неопределенные смеси величин, различающихся по определению[10]. Значительная часть первичных публикаций сообщает вместо измеренной температуры лишь о комнатной температуре, которая различается между лабораториями и сезонно, вследствие чего объединение таких данных само по себе увеличивает дисперсию[11]. А для данных о биологической активности требование совпадения основных характеристик эксперимента сокращает долю существенно расходящихся значений почти вдвое[12].

Наконец, пятый класс образует публикационное смещение. Состав набора отражает распределение интересов и возможностей исследователей, опубликовавших исходные работы. Показательным примером служит выбор объектов исследования. При построении моделей растворимости в органических растворителях круг рассматриваемых сред был ограничен тремя, причем один из них включен с прямой оговоркой, что в современной практике он употребляется редко и выбран именно

потому, что для него имелось достаточно данных[5]. Сходным образом объем накопленных сведений оказывается несопоставимо большим для тех сред, применение которых обусловлено технологическими потребностями.

Совокупным следствием перечисленных дефектов является систематическое завышение публикуемых оценок качества моделей. Наиболее отчетливо это проявляется при ужесточении условий проверки. По мере того, как тестовые соединения удаляются от обучающего множества, точность моделей всех рассмотренных типов быстро падает вплоть до уровня случайного угадывания, причем этот эффект сохраняется и при добавлении трехмерных геометрических дескрипторов[22]. Существенно, что получаемая оценка зависит от способа формирования тестовой выборки. Разбиение по времени публикации данных, воспроизводящее условия действительного прогнозирования, дает результат, систематически отличающийся от получаемого при случайном разбиении[23].

2.1.4. Требования к данным: FAIR, машиночитаемость, воспроизводимость

Общепринятая система требований к научным данным сформулирована в виде принципов FAIR: данные должны быть находимыми (Findable), доступными (Accessible), совместимыми (Interoperable) и пригодными к повторному использованию (Reusable)[24]. Существенно, что эти принципы с самого начала формулировались применительно не только к данным, но и к описывающим их метаданным, и что их центральной целью объявлена машинная обрабатываемость — возможность использования данных вычислительными средствами при минимальном участии человека. Именно этим FAIR отличается от более ранних представлений об открытости данных, для которых достаточным считался свободный доступ к файлу.

Находимость требует, чтобы данные и метаданные имели глобально уникальный и постоянный идентификатор (DOI и т.д.), были снабжены содержательными метаданными и зарегистрированы в системе, допускающей поиск (Zenodo, Figshare и т.д.). Применительно к наборам химических данных это означает наличие собственного идентификатора у набора как такового, а не только у описывающей его публикации.

Доступность предполагает возможность получения данных по их идентификатору посредством открытого и общедоступного протокола. Отдельного упоминания заслуживает требование, согласно которому метаданные должны оставаться доступными даже в том случае, когда сами данные более недоступны. Сведения о том, что и при каких условиях было измерено, сохраняют ценность независимо от судьбы конкретного файла.

Совместимость требует использования формального, общепринятого языка представления сведений и словарей, которые сами удовлетворяют принципам FAIR, а также наличия явных ссылок на связанные данные. Для химических наборов это означает представление структур в стандартной машиночитаемой нотации и обязательное указание единиц измерения. Набор, опубликованный в виде таблиц или графиков в тексте статьи либо в сопроводительных материалах в формате, не допускающем автоматического чтения, этому требованию не удовлетворяет независимо от качества содержащихся в нем данных.

Наконец, пригодность к повторному использованию раскрывается через четыре положения: подробное описание набора совокупностью значимых атрибутов, наличие явной лицензии на использование, сопровождение данных детальными сведениями об их происхождении и соответствии принятым в предметной области стандартам. Третье из них (требование прослеживаемости) имеет для настоящей работы определяющее значение: именно оно обеспечивает возможность независимой проверки каждого значения и, как показано выше, служит единственным средством обнаружения дублирования, возникающего при последовательной агрегации наборов. Практически это означает, что каждая запись должна нести указание на публикацию или справочник, в которой соответствующее измерение было выполнено, а не на промежуточную компиляцию, из которой она заимствована. Примером набора, удовлетворяющего перечисленным требованиям, может служить упомянутая выше оцифрованная компиляция констант кислотной диссоциации кислот: она размещена в общедоступном репозитории под постоянным идентификатором и открытой лицензией, содержит машиночитаемые представления структур в нотациях SMILES и

InChI, а каждое значение сопровождается указанием температуры, метода измерения и оценки надежности[16].

Перевод перечисленных общих положений на язык конкретных требований к химическому эксперименту составляет предмет разрабатываемых в настоящее время минимальных стандартов информации, определяющих, какие именно сведения должны сопровождать результат измерения, чтобы он оставался пригодным к повторному использованию[25]. Создаются и исследовательские инфраструктуры, обеспечивающие соблюдение этих требований непосредственно на этапе получения данных, что позволяет фиксировать условия эксперимента в момент его проведения, а не восстанавливать их впоследствии по тексту публикации[26]. Формы представления данных, этим требованиям не отвечающие, приведены на рис. 1.



Рисунок 1. Формы представления результатов эксперимента, не удовлетворяющие принципам FAIR[25].

2.1.5. Обзор существующих ресурсов химических данных

Существующие ресурсы химических данных целесообразно разделить на несколько групп по характеру содержащихся в них сведений.

Наиболее крупными являются агрегаторы общего назначения. Ресурс PubChem объединяет сведения, поступающие более чем от тысячи организаций, и содержит по состоянию на сентябрь 2024 года 119 миллионов соединений, 322 миллиона депонированных записей о веществах и 295 миллионов значений биологической активности[27]. База ChEMBL специализируется на данных о биологической активности, извлекаемых из первичной литературы, и в 33-м выпуске содержала 2.4 миллиона соединений и свыше 20 миллионов измерений, извлеченных из 88 тысяч публикаций и патентов[28]. Кембриджская структурная база данных объединяет результаты рентгеноструктурных исследований и служит основным источником сведений о кристаллическом строении органических и металлоорганических соединений[29]. Особое положение занимают каталоги доступных для приобретения структур: база ZINC-22 содержит свыше 37 миллиардов перечисленных соединений, из которых более 4.5 миллиарда представлены в трехмерном виде, пригодном для молекулярного докинга[30]. Ресурсы этого рода задают верхнюю границу пространства, доступного для перебора, но экспериментально измеренных свойств не содержат.

Вторую группу образуют наборы расчетных величин. Наиболее известный из них содержит квантово-химические характеристики для 134 тысяч малых органических молекул[31], а недавно опубликованный набор QuantumPioneer - характеристики 348 тысяч соединений и свыше ста миллионов рассчитанных значений энергии и энтальпии сольватации для 295 растворителей[32]. Эти ресурсы обеспечивают беспрецедентный объем и полную внутреннюю согласованность, но описывают результат расчета, а не измерения, вследствие чего пригодны главным образом для предварительного обучения моделей.

Третью группу составляют наборы, посвященные отдельным свойствам и создаваемые в связи с конкретной задачей прогнозирования. Объем их измеряется сотнями, тысячами, реже десятками тысяч записей. Именно к этой группе принадлежат ресурсы, рассматриваемые в последующих разделах применительно к растворимости, люминесцентным свойствам и цитотоксичности. Общей их особенностью является

существенно меньший объем по сравнению с агрегаторами и, как следствие, определяющее влияние качества курирования на достигаемые результаты.

Наконец, необходимо отметить неравномерность покрытия химического пространства перечисленными ресурсами. Координационные соединения переходных металлов представлены в них несопоставимо беднее органических, причем это положение не является следствием одной лишь меньшей изученности. Общепринятые процедуры подготовки химических данных предписывают удаление неорганических и металлоорганических соединений в качестве одного из первых шагов, поскольку для них не удастся вычислить стандартные дескрипторы, а невозможность моделирования соединений этого класса прямо признается ограничением используемого аппарата[20]. Эта практика сохраняется и в настоящее время: при создании набора SangsterLogP, объединившего более 23 тысяч экспериментальных значений LogP из трех тысяч источников, металлосодержащие и неорганические соединения были исключены отдельными пунктами процедуры фильтрации с прямым указанием на невозможность их обработки общепринятыми вычислительными методами[18]. Комплексы, таким образом, не просто реже попадают в наборы данных - они систематически исключаются из них на этапе курирования.

Масштаб этой неравномерности поддается прямой оценке и приведен в разделе 2.3.2: сведения о строении комплексов накоплены в изобилии, расчетные характеристики для них получены в сопоставимом объеме, а экспериментально измеренные свойства остаются на три порядка более редкими. Этим же обстоятельством объясняется и распространенность подхода, при котором недостающие экспериментальные величины замещаются расчетными. Модель в таком случае обучается воспроизводить результат расчета, а не измерения.

2.2. Прогнозирование растворимости органических соединений

2.2.1. Задача: величины, единицы, методы измерения

Растворимость определяет применимость органического соединения в синтетической практике, фармацевтической разработке и промышленной технологии: от нее зависят выбор среды для проведения реакции[33], условия перекристаллизации[34], а для лекарственных веществ - биодоступность при

пероральном приеме[35]. Потребность в прогнозировании возникает поэтому на самых ранних стадиях разработки, когда экспериментальное определение для всей библиотеки соединений непозволительно трудоемко.

Прежде чем обсуждать методы прогнозирования, необходимо уточнить, что именно понимается под растворимостью, поскольку под одним термином в литературе объединены величины, различающиеся как по физическому смыслу, так и по способу определения. Равновесная, или термодинамическая, растворимость (equilibrium, thermodynamic solubility) - это концентрация насыщенного раствора, находящегося в равновесии с избытком нерастворившегося вещества в его устойчивой кристаллической форме[36]. Кинетическая растворимость (kinetic solubility) определяется как максимальная концентрация наиболее быстро осаждающейся формы вещества и обычно измеряется при разбавлении концентрированного раствора в диметилсульфоксиде водной средой. Выпадающий при этом осадок, как правило, не характеризуется и может представлять собой аморфную фазу, соль или сокристалл[36]. Наконец, собственная растворимость (intrinsic solubility) — это равновесная растворимость неионизованной формы ионизируемого соединения при заданной температуре. Именно она наиболее строго определена и используется в качестве целевой величины в слепых испытаниях прогностических методов[37].

Различие между этими величинами не является терминологической тонкостью. Вследствие пересыщения, возникающего при разбавлении раствора в диметилсульфоксиде, кинетическая растворимость систематически превышает равновесную, причем расхождение достигает нескольких порядков. Для амиодарона кинетическая растворимость, измеренная через десять минут после разбавления, превышает собственную в 3860 раз, для глибенкламида - в 1414, для тамоксифена - в 443 раза[36]. Увеличение времени выдержки до двадцати часов сближает величины, но не устраняет расхождения (рис. 2). Смещение этих величин при объединении литературных данных приводит, таким образом, не к росту случайного разброса, а к появлению систематически завышенных значений.

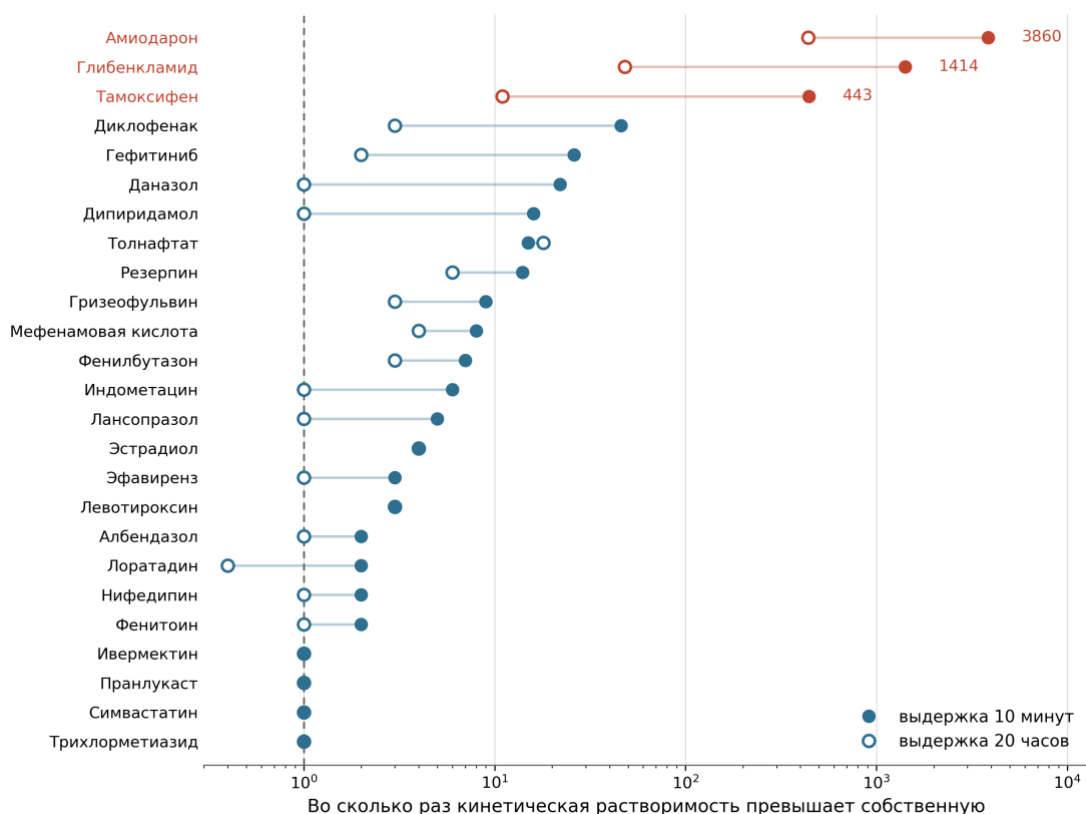


Рисунок 2. Отношение кинетической растворимости к собственной для двадцати пяти лекарственных веществ при двух временах выдержки. Штриховая линия отвечает совпадению величин. Для резерпина, толнафтата и трихлорметиазида одна из величин лежала ниже предела обнаружения, приведенные для них отношения представляют собой нижние оценки. Построено по данным[36].

Величина растворимости относится, кроме того, к определенной кристаллической фазе. Одно и то же соединение может существовать в виде нескольких полиморфных модификаций, а также гидратов и сольватов, различающихся энергией кристаллической решетки и, следовательно, растворимостью. Это явление распространено шире, чем принято считать: по данным Кембриджского банка структурных данных полиморфизм обнаружен у каждого третьего соединения, а при систематическом скрининге твердых форм в фармацевтических лабораториях - у каждого второго[38]. Величина различия при этом умеренна. Обобщение литературных данных по 55 соединениям, охватывающее 81 отношение растворимостей, показало, что растворимости полиморфов различаются обычно менее чем вдвое, тогда как для пар ангидрат-гидрат разброс шире[39]. Оценка

с другой стороны, по вычисленным энергиям кристаллической решетки для 508 полиморфных соединений, дает согласующуюся картину: более половины пар различаются менее чем на 2 кДж/моль, что при комнатной температуре отвечает отношению растворимостей около двух, и лишь у пяти процентов пар разность превышает 7.2 кДж/моль[40].

Практическое затруднение состоит, однако, не в величине эффекта, а в том, что твердая фаза в первичных публикациях указывается редко. Растворимость метастабильной модификации к тому же плохо поддается измерению: в ходе установления равновесия она превращается в устойчивую форму, вследствие чего измеренное значение зависит от продолжительности опыта и способа регистрации[41]. Объединение литературных данных без сведений о твердой фазе означает, таким образом, смешение величин, относящихся к разным фазам, причем оценить долю таких случаев по опубликованным наборам невозможно.

Дополнительная неоднозначность связана с формой представления результата. В публикациях растворимость сообщается в мольных долях, в единицах массовой концентрации, в молях на литр раствора или в виде десятичного логарифма молярной растворимости ($\log S$), которая наиболее распространена в машинном обучении. Переход между мольной долей и молярностью требует знания плотности насыщенного раствора, которая приводится далеко не всегда, вследствие чего пересчет вносит собственную погрешность. По этой причине при формировании баз данных существенно фиксировать, в каких единицах величина была приведена в первоисточнике, и какие из представленных в базе форм являются производными.

Совокупность используемых экспериментальных методов также неоднородна. Классическим и наиболее надежным остается метод изотермического насыщения (shake-flask), в котором избыток вещества выдерживают в растворителе до установления равновесия с последующим анализом надосадочной жидкости[42]. Недостатки его - длительность, трудоемкость и значительный расход вещества. Потенциометрические методы, прежде всего CheqSol[43] и титрование по шаблону растворения (ДТТ)[44], проводят систему через последовательность пересыщенных и ненасыщенных состояний, что позволяет достичь равновесия за десятки минут и

попутно обнаружить изменение полиморфной формы осадка, но применимы лишь к ионизируемым соединениям в водной среде. Синтетический (лазерный) метод фиксирует момент полного растворения по восстановлению интенсивности проходящего света и естественным образом дает температурную зависимость, к нему примыкают турбидиметрические методики. Испарительные методики, такие как LYSA, и высокопроизводительные анализы на основе разбавления растворов в диметилсульфоксиде производительны, но дают лишь оценку порядка величины и склонны к образованию метастабильных твердых фаз. В органических растворителях потенциометрические подходы неприменимы, и измерения проводят гравиметрическим вариантом изотермического насыщения, хроматографическим или спектрофотометрическим определением концентрации, а также синтетическим методом[45].

Отдельного внимания заслуживает температурная зависимость. В работах, посвященных растворимости в органических растворителях, экспериментальные значения обычно измеряют в интервале порядка 40 К с шагом 5 К и описывают полуэмпирическими соотношениями — уравнением Вант-Гоффа, модифицированным уравнением Апельблата[46] или моделью λh [47], — параметры которых приводятся наряду с самими значениями[45]. Для задач машинного обучения существенно, что температура является полноправной переменной, а не условием эксперимента, которым можно пренебречь. В пределах указанного интервала растворимость одного и того же вещества может изменяться в разы.

Наконец, необходимо охарактеризовать структуру погрешностей экспериментальных данных, поскольку именно она определяет достижимый предел точности прогнозирования. Принято выделять три класса ошибок[37]. К первому относятся грубые ошибки: смещение растворимости неионизованной и ионизованной форм, подмена равновесной величины кинетической вследствие недостаточного времени эксперимента, неучтенное протекание химической реакции и опечатки при переносе значений из первоисточника. Второй класс образуют систематические расхождения между лабораториями и между методиками. Третий составляют случайные ошибки при повторении опыта в одних и тех же условиях, для метода

CheqSol оцениваемые величиной порядка 0.05 логарифмической единицы. Дополнительным источником искажений служит образование коллоидных агрегатов, характерное для малорастворимых органических молекул[48].

Последовательное применение критического разбора первоисточников с учетом условий эксперимента, чистоты и кристалличности образцов позволило подготовить для второго слепого испытания набор из ста соединений с межлабораторным стандартным отклонением около 0.17 логарифмической единицы, тогда как сопутствующий набор из тридцати двух соединений, для которых такого согласования достичь не удалось, характеризуется величиной около 0.62[49]. Обычно цитируемые оценки межлабораторного разброса лежат в диапазоне 0.5-0.7 логарифмической единицы[37].

Второе из указанных обстоятельств имеет прямое отношение к структуре создаваемых баз данных. Объединение в одной обучающей выборке значений, измеренных при различных температурах без пересчета, само по себе увеличивает дисперсию примерно на 0.13 логарифмической единицы, при том, что значительная часть первоисточников сообщает лишь о «комнатной температуре», которая различается как между лабораториями, так и сезонно[11]. Температура должна поэтому фиксироваться для каждого значения как обязательный атрибут, а не рассматриваться как второстепенное условие эксперимента.

Показательно соотношение этих величин. Систематический разбор более чем восьмисот публикаций, посвященных определению равновесной растворимости методом изотермического насыщения, показал, что ожидаемая межлабораторная воспроизводимость порядка 0.7 логарифмической единицы может быть снижена приблизительно до 0.15 за счет двух нормировок: приведения измеренной величины к собственной растворимости и пересчета значений, полученных в интервале 20-40 °C, к стандартной температуре 25 °C[11]. Воспроизводимость литературных данных может быть, иными словами, улучшена в четыре-пять раз без проведения новых измерений - исключительно за счет критического разбора условий каждого эксперимента и приведения значений к единым условиям.

2.2.2. Растворимость в воде

Растворимость в воде занимает в задачах прогнозирования особое положение. Это одно из немногих свойств такого рода, для которых накоплен достаточный по объему и разнообразию массив экспериментальных данных, поскольку растворимость в воде во многом определяет биодоступность лекарственного вещества и фармацевтическая промышленность десятилетиями финансировала как измерения, так и разработку моделей. Пример водной растворимости служит поэтому естественной точкой отсчета. На нем видно, чего удастся достичь в наиболее благоприятных условиях.

Первые количественные подходы восходят к линейным соотношениям структура-свойство 1960-1970-х годов, связывавшим растворимость с коэффициентом распределения октанол/вода ($\log P$). Их обобщением стала работа Ялковского и Валвани[50], в которой растворение кристаллического вещества разложено на два вклада: плавление гипотетического переохлажденного расплава и последующее смешение с растворителем. В окончательной редакции[51] обобщенное уравнение растворимости (general solubility equation, GSE) записывается как

$$\log S = 0.5 - 0.01(T_{\text{пл}} - 25) - \log P$$

где S - молярная растворимость, $T_{\text{пл}}$ - температура плавления в градусах Цельсия, а $\log P$ - коэффициент распределения октанол/вода. Для веществ, жидких при комнатной температуре, член с температурой плавления обнуляется. Коэффициент 0.01 не является подгоночным. Он получается из усредненной энтропии плавления органических кристаллов по правилу Вальдена. Существенно, что это соотношение не содержит параметров, полученных регрессией, и использует всего два дескриптора. На выборке из 410 органических неэлектролитов, растворимость которых охватывает шестнадцать порядков, средняя абсолютная ошибка составляет 0.42 логарифмической единицы. Эта величина далее служит естественным ориентиром при оценке всех последующих методов.

Систематическая проверка применимости этого соотношения была предпринята существенно позднее[52]. На выборке из 2740 лекарство подобных соединений, для

которых и температура плавления, и коэффициент распределения измерены экспериментально, оценка признавалась удовлетворительной, если она отклоняется от опыта не более чем на логарифмическую единицу, то есть в пределах погрешности самого измерения. Условию этому отвечают 64 % соединений, тогда как для остальной трети уравнение дает существенно неверный ответ. Разделение проходит главным образом по размеру и конформационной подвижности молекулы: соотношение хорошо работает для небольших соединений с жесткой структурой и оказывается неточным для крупных, более гидрофобных и с большей площадью полярной поверхности. На этом основании построена классификационная модель GSESolver, прогнозирующая по одной лишь строке SMILES, окажется ли оценка достаточно точной.

Существенно, что задача поставлена как определение границ применимости расчета. Авторы прямо указывают, что главное затруднение состоит в отсутствии способа заранее установить, попадает ли в эту область рассматриваемое соединение. Отсюда вытекает и преимущество перед моделями, обучаемыми непосредственно на растворимости: точность последних падает за пределами структурного пространства обучающей выборки, тогда как само общее уравнение растворимости обучающей выборки не имеет вовсе, а область его применимости теперь определяется явно. Этот сдвиг показателен: коль скоро точность прогнозирования уперлась в предел, задаваемый воспроизводимостью самих измерений, указывать границы применимости расчета оказалось полезнее, чем добиваться дальнейшего уменьшения ошибки. Предложен и обходной путь, снимающий необходимость располагать экспериментальной температурой плавления: последняя прогнозируется моделью, после чего уравнение применяется к прогнозируемому значению[53].

В 1990-2000-е годы развитие шло по пути расширения набора дескрипторов и усложнения регрессионных моделей - от топологических и электротопологических индексов[54,55] до предельно экономных соотношений, из которых наибольшую известность получило уравнение ESOL[56]. Заявляемая точность лежала в диапазоне 0.6-1.0 логарифмической единицы, однако сопоставление между работами было практически невозможным. Каждая группа использовала собственную выборку и собственное разбиение, а независимых тестовых наборов не существовало.

Отдельно следует отметить, как была устроена работа с данными на раннем этапе. В базе AQUASOL, описанной в 1989 году[17], каждое извлеченное из литературы значение оценивалось по пяти критериям - температуре, чистоте вещества, полноте установления равновесия, методу анализа и точности приводимых данных, - на основании которых вычислялся весовой коэффициент, а все найденные для соединения значения сохранялись по отдельности со ссылкой на первоисточник. Практика фиксации условий эксперимента и оценки надежности каждой записи была, таким образом, выработана уже на этом этапе, но при переходе к укрупненным компиляциям оказалась в значительной мере утрачена. В наборах, легших в основу моделей 2010-х годов, каждому соединению в базе данных соответствует единственное значение без указания источника и условий измерения. Показательна и судьба второй из ранних компиляций. Дефекты набора PHYSPROP были отмечены уже в современной ему рецензии: из примерно 11000 записей около 4000 не содержали ни структуры, ни названия, названия встречались с опечатками и подменялись формульными обозначениями, а сама растворимость в воде была приведена лишь для 880 веществ[57]. Тем не менее эта компиляция наряду с AQUASOL легли в основу набора Huuskonen, а через него - многих последующих[10].

Ответом на эту трудность стала серия слепых испытаний Solubility Challenge. В первом из них[58] участникам предоставлялся набор из ста надежных измерений и предлагалось спрогнозировать растворимость тридцати двух соединений, значения для которых не публиковались. Результаты оказались весьма отрезвляющими. Ошибки лучших методов составили порядка 0.6 логарифмической единицы, значительная часть присланных прогнозов была существенно хуже, а сложные модели не имели систематического преимущества перед простыми. Повторное испытание, проведенное спустя десятилетие[49] с двумя тестовыми наборами, различающимися воспроизводимостью эталонных значений (стандартное отклонение 0.17 и 0.62 логарифмической единицы), и его развернутый анализ[59] показали, что за прошедшие годы, несмотря на радикальное развитие методов машинного обучения, достигаемая точность практически не изменилась.

Параллельно область осваивала методы глубокого обучения - от рекурсивных архитектур, работающих непосредственно с молекулярным графом[60], до моделей передачи сообщений и архитектур на основе механизма внимания, в частности SolTranNet[61]. Систематическое сравнение их на единых данных[62] выявило, однако, что различия между архитектурами сопоставимы по величине с разбросом, вносимым выбором схемы разбиения выборки, то есть не являются определяющими.

К концу 2010-х годов фокус усилий начал смещаться от архитектур к данным. Была создана курированная база AqSolDB[63], объединившая девять открытых источников и снабженная рассчитанными дескрипторами. Было показано, что тщательная очистка обучающей выборки дает прирост качества, превышающий эффект от смены модели[9]. В последней работе семь наборов, включая AqSolDB, были сведены воедино с отдельной обработкой термодинамических и кинетических данных, причем каждому источнику присваивался весовой коэффициент, отражающий его достоверность. Наименьший среди термодинамических наборов вес получил именно AqSolDB, а при фильтрации из набора OCHEM было исключено около трети записей.

Наиболее полный критический анализ сложившейся практики был проведен в работе[10], охватившей публикации наборов данных и моделей растворимости в воде за более чем двадцать лет. Установлено, что широко используемые наборы существенно перекрываются между собой, причем одно и то же экспериментальное значение, восходящее к одному первоисточнику, попадает в объединенный массив многократно, будучи заимствованным разными компиляциями, и без разбора до первоисточника такое дублирование неразличимо. Существеннее, однако, замечание о содержании самих наборов: показано, что AqSolDB[63] и OCHEM[64], два крупнейших свободно доступных ресурса, представляют собой неопределенные смеси значений собственной, кажущейся и равновесной растворимости, то есть величин, различающихся по своему физическому смыслу. Обучение на таком массиве означает, что модель воспроизводит смесь физических величин, состав которой определяется происхождением данных.

Количественная оценка последствий получена прямым экспериментом. Очистка AqSolDB по предложенной процедуре сократила набор с 9982 до 8047 записей и при этом систематически улучшила среднеквадратичную ошибку моделей приблизительно на 0.10 логарифмической единицы. Удаление пятой части данных, таким образом, повысило качество прогнозирования - результат, прямо противоречащий распространенному представлению о том, что увеличение объема обучающей выборки всегда предпочтительно. Внешняя же проверка опубликованных моделей на независимом наборе показала, что заявляемые в литературе величины ошибки, лежащие в диапазоне от 0.35 до 1.71 логарифмической единицы, при переносе на данные, не пересекающиеся с обучающими, существенно ухудшаются, а нейросетевые архитектуры не обнаруживают при этом прорывного превосходства.

По итогам анализа предложена формализованная процедура курирования данных о растворимости в виде дерева решений: проверка структуры по регистрационному номеру CAS, оценка экспериментального протокола и межлабораторного отклонения, исключение значений ниже предела обнаружения, учет чистоты вещества и ограничение выборки однокомпонентными веществами, растворимость которых измерена при комнатной температуре и нейтральном pH. Эта работа представляет собой наиболее последовательный из опубликованных подходов к курированию данных о растворимости и служит методическим ориентиром для сходных задач.

Вопрос о природе наблюдаемого предела точности прогнозирования был поставлен прямо в работе[65], где показано, что погрешность лучших моделей приблизилась к воспроизводимости самого эксперимента. Расхождения между значениями растворимости, измеренными в разных лабораториях для одних и тех же соединений, имеют тот же порядок величины. Современные оценки алеаторного предела для доступных наборов данных составляют 0.5-1 логарифмическую единицу[66], то есть по порядку величины совпадают с погрешностью, достигнутой лучшими моделями.

Таким образом, за сорок пять лет — от двухпараметрического соотношения без обучаемых коэффициентов до графовых нейронных сетей и трансформеров с

миллионами параметров — достигаемая точность прогнозирования растворимости в воде изменилась несопоставимо мало по сравнению с ростом сложности используемых моделей. Этот факт имеет принципиальное значение для дальнейшего изложения. Он показывает, что ограничивающим фактором выступает качество и воспроизводимость экспериментальных данных, на которых обучаются модели. Соответственно, продвижение в задачах, рассматриваемых в следующих разделах, следует ожидать в первую очередь от расширения и стандартизации экспериментальной базы, а не от усложнения архитектур.

2.2.3. Растворимость в органических растворителях

Переход от растворимости в воде к растворимости в органических растворителях не сопровождается сменой методического аппарата. Используются те же дескрипторы, те же архитектуры моделей и те же принципы валидации, что описаны выше. Различие состоит прежде всего в обеспеченности задачи экспериментальными данными, и именно оно определяет наблюдаемое отставание примерно на полтора десятилетия.

Постановка задачи, однако, усложняется в двух отношениях. Во-первых, растворитель становится полноправным входом модели. Прогнозировать требуется свойство пары «растворенное вещество — растворитель», вследствие чего необходимо машиночитаемое представление обоих компонентов и способ описания их взаимодействия. Во-вторых, температура перестает быть фиксированным условием и превращается в непрерывную переменную. Различие здесь не в самих измерениях: в первоисточниках растворимость и в воде, и в органических растворителях определяют, как правило, серией при нескольких температурах. Оно заключено в устройстве наборов данных - компиляции по водной растворимости температуру не сохраняют вовсе, в AqSolDB соответствующего поля нет, - тогда как наборы для органических растворителей приводят ее для каждой записи.

Одной из первых систематических попыток перенести аппарат машинного обучения на неводные среды стала работа[5], в которой физико-химические и квантово-химические дескрипторы сочетались с восемью алгоритмами машинного обучения. Данные для нее собирались из базы Reaxys и открытого набора Open

Notebook Science Challenge, причем отбирались только нейтральные органические растворенные вещества в индивидуальных растворителях. Показательно, каким образом нехватка данных определила постановку задачи: поиск по Reaxys дал столь ограниченный массив для органических растворителей, что исследование пришлось свести к трем средам — этанолу, бензолу и ацетону, — выбранным так, чтобы одновременно перекрыть диапазон полярности и обеспечить достаточный для обучения объем выборки. Авторы прямо отмечают, что бензол не относится к числу употребительных в современной практике растворителей и включен в рассмотрение именно потому, что для него имелось достаточно данных. Тем самым химический охват исследования диктовался доступностью литературных значений.

Существенны и два методических вывода этой работы. Во-первых, точность прогнозирования оказалась практически независимой от примененного алгоритма. Восемь различных методов дали сопоставимые результаты, а достигнутый уровень соответствовал ожидаемому уровню шума в обучающих данных, порядка 0.7 логарифмической единицы. Во-вторых, авторы отказались от использования коэффициента детерминации и среднеквадратичной ошибки в качестве основных метрик, поскольку при характерной для литературных данных экспериментальной погрешности 0.5-0.7 логарифмической единицы эти величины теряют информативность, и ввели вместо них долю прогнозов, попадающих в интервал ± 0.7 и ± 1.0 логарифмической единицы от экспериментального значения[5].

Дальнейшее развитие направления шло по нескольким направлениям. Сопоставлялись молекулярные отпечатки и физико-химические дескрипторы с графовыми сверточными сетями[67,68]. Существенным шагом стало использование термодинамического цикла, при котором вместо прямого прогнозирования растворимости модель обучалась на энергиях сольватации и растворения, что позволило охватить широкий круг растворителей и температур[69]. В рамках этой работы была собрана и опубликована коллекция баз данных SolProp, включающая как экспериментальные наборы — CombiSolu-Exp с 4953 значениями растворимости 115 твердых органических веществ в 97 органических растворителях в интервале 243-364 К, отдельный набор для повышенных температур вплоть до 593 К и 11804 значения

растворимости в воде, — так и рассчитанные методом COSMO-RS массивы энергий сольватации объемом до миллиона записей, использованные для предварительного обучения моделей. Впоследствии был предложен ряд единых архитектур, охватывающих различные растворители в рамках одной модели[70].

Ограничивающим фактором на протяжении всего этого периода оставался объем доступных данных. Используемые наборы содержали, как правило, от нескольких сотен до нескольких тысяч значений для единиц или десятков растворителей, что несопоставимо с масштабом ресурсов, доступных для растворимости в воде (табл. 1). Такой объем не позволяет ни обучать модели высокой емкости, ни корректно оценивать их обобщающую способность. При малом числе растворителей и растворенных веществ любое разбиение выборки оставляет в обучении близкие аналоги тестовых систем, вследствие чего заявляемые метрики характеризуют скорее интерполяцию внутри заполненной матрицы, чем прогнозирование для новых объектов.

Особняком стоит случай диметилсульфоксида. Поскольку библиотеки соединений, находящиеся в фармацевтической разработке, хранятся в виде растворов в этом растворителе, данные о растворимости в нем накоплены в объемах, существенно превышающих объемы данных для остальных органических растворителей. Модели строились на выборке из 163 тысяч молекул, объединяющей коммерческую коллекцию и два корпоративных собрания[71]. Однако решаемая при этом задача иная — бинарная классификация соединений на растворимые и нерастворимые по порогу $10 \text{ ммоль} \cdot \text{л}^{-1}$ при доле нерастворимых от 1.7 до 5.8 %, — а измеряемая величина представляет собой кинетическую растворимость, определяемую в условиях скрининга. Кроме того, данные такого происхождения в большинстве своем не публикуются. Этот случай показывает, что объем экспериментальных данных определяется наличием отчетливого промышленного запроса, но накопленные таким образом сведения не восполняют дефицита значений равновесной растворимости, необходимых для количественного прогнозирования. Сходным образом устроено и положение с растворимостью в сверхкритическом диоксиде углерода, применяемом в экстракционных и осадительных процессах. Для этой среды собран массив почти в 32

тысячи экспериментальных значений, на котором построены модели, использующие наряду со структурными дескрипторами термодинамические характеристики фазового перехода — температуру плавления, критические параметры, энтальпию испарения и энергию Гиббса сольватации[72].

Вместе с тем достигнутый уровень не следует переоценивать. Обобщение на растворители, отсутствовавшие в обучающей выборке, остается слабым местом. Химическое пространство растворителей в имеющихся наборах покрыто крайне неравномерно, поскольку определяется тем, какие среды представляли практический интерес для авторов первичных публикаций, на что указывают и авторы моделей FastSolv[66]. В этом проявляется публикационное смещение, характерное для любых данных, извлекаемых из литературы. Кроме того, сопоставление результатов различных работ затруднено отсутствием общепринятого тестового набора. В отличие от растворимости в воде, для которой такую роль играли слепые испытания Solubility Challenge[49,58], для органических растворителей общепризнанного аналога до настоящего времени не сложилось, вследствие чего каждая работа оценивается на собственном разбиении собственных данных, а заявляемые метрики оказываются несопоставимыми между публикациями.

Практическая востребованность этого направления подтверждается его внедрением в промышленную практику. Модели прогнозирования растворимости в органических растворителях используются, в частности, при оценке эффективности удаления мутагенных примесей в фармацевтическом производстве[73], а также при подборе сред и сорастворителей[74].

Существующие агрегированные наборы объединяют данные, полученные различными методами и при различных температурах, причем сама температура либо приводится единственным значением, либо утрачивается при агрегации наряду с прочими сведениями об условиях измерения[10,11]. Восполнение этого пробела — создание максимально полного и стандартизованного массива экспериментальных данных о растворимости органических соединений в индивидуальных органических растворителях с обязательной фиксацией температуры и прослеживаемостью каждого значения до первоисточника — составляет одну из задач настоящей работы.

2.2.4. Бинарные смеси растворителей

Смеси растворителей широко применяются в химической технологии и фармацевтическом производстве. Подбор бинарной среды позволяет управлять растворимостью в широких пределах, что используется при антирастворительной кристаллизации, экстракции и разработке лекарственных форм. Практический интерес к таким системам обусловлен не только возможностью плавно варьировать растворяющую способность изменением состава смеси растворителей, но и наличием синергетических эффектов, при которых растворимость в смеси превышает растворимость в каждом из чистых компонентов[75].

С точки зрения прогнозирования переход к смесям означает качественное усложнение задачи. Входными данными модели становятся структура растворенного вещества, структуры двух растворителей, состав смеси и температура, причем зависимость растворимости от состава существенно нелинейна и не выводится из свойств чистых компонентов. Возникает и самостоятельное затруднение при построении представления. Описание смеси должно быть инвариантным относительно перестановки компонентов и допускать произвольное их число, тогда как простое объединение признаков отдельных растворителей этому требованию не удовлетворяет. Предложены различные способы его преодоления — функция агрегации MolPool, позволяющая графовой сети работать с произвольным числом молекул растворителя[76], и подграфовые архитектуры, описывающие взаимодействия внутри смеси явно[77]. Одновременно резко возрастает объем химического пространства, подлежащего покрытию. Число возможных бинарных комбинаций растет квадратично с числом растворителей, а каждая комбинация требует измерений при нескольких составах и нескольких температурах.

Традиционно для описания таких систем применяются полуэмпирические корреляционные соотношения, из которых наиболее известна модель Джойбана-Акри[78], связывающая растворимость в смеси с растворимостью в чистых компонентах и составом смеси. Принципиальное ограничение этих моделей состоит в том, что они являются интерполяционными. Их параметры определяются подгонкой по экспериментальным данным для той же самой системы, и для новой пары

растворителей или нового вещества модель неприменима. Предпринимались попытки преодолеть это ограничение путем объединения корреляционных соотношений с нейросетевыми регрессорами[79], однако потребность в экспериментальных данных по рассматриваемой системе при этом сохраняется.

Собственно машинное обучение пришло в эту область поздно. Подавляющее большинство работ приходится на период с 2024 года. Были предложены подходы, сочетающие высокопроизводительный эксперимент с обучением моделей[80], что отражает более общую тенденцию к автоматизации исследований по кристаллизации исследований, рассматриваемую как средство ускоренного получения экспериментальных данных[81], модели на основе градиентного бустинга[82,83], графовые сверточные архитектуры[84], а также подходы, опирающиеся на термодинамический цикл и прогнозирование энергий сольватации в смешанных средах[76].

Наиболее полно постановка задачи проработана в работе[83], где был собран набор из 28098 экспериментальных значений растворимости в бинарных смесях при различных температурах, обучены модели градиентного бустинга, достигшие средней абсолютной ошибки 0.33 логарифмической единицы на отложенной выборке, и, что существенно, проведена экспериментальная валидация. Растворимость четырех лекарственных веществ была сначала спрогнозирована, а затем измерена экспериментально. Такая проверка встречается в литературе редко и представляет собой наиболее убедительный способ оценки модели.

Однако анализ состава этого набора показывает, насколько узким остается химическое пространство задачи: 28098 значений относятся всего к 125 растворенным веществам, 44 растворителям и 110 бинарным смесям, что дает лишь 373 уникальные системы «вещество-бинарная смесь» (рис. 3). Показательно, что последующие работы, сообщающие об улучшении точности за счет более сложных архитектур — например, о снижении средней абсолютной ошибки с 0.33 до 0.28 логарифмической единицы[84], — используют тот же самый набор данных. Прогресс в области, таким образом, измеряется сотыми долями логарифмической единицы на одной и той же выборке из

125 соединений, тогда как вопрос о поведении моделей за ее пределами остается открытым.

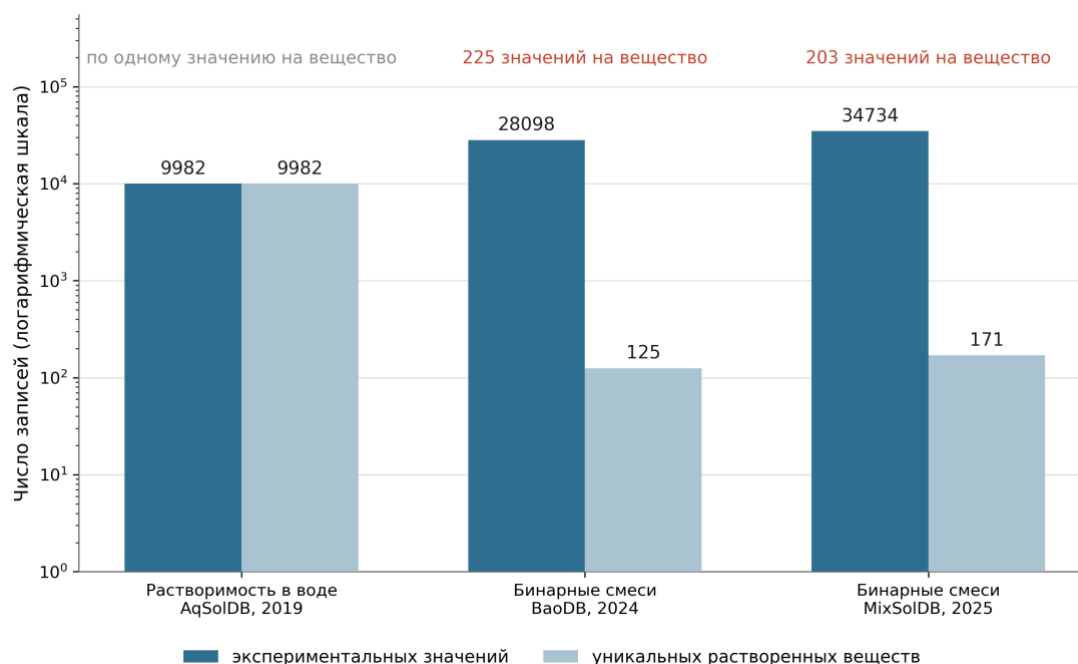


Рисунок 3. Объем и разнообразие наборов данных о растворимости. Для наборов по бинарным смесям число экспериментальных значений более чем на два порядка превышает число уникальных растворенных веществ. Величины для BaoDB и MixSolDB получены прямым пересчетом опубликованных файлов и приведены в табл. 5. Построено по данным[63,77,83].

Попытка расширить охват была предпринята при создании набора MixSolDB[77], объединившего 56789 значений для 11796 растворенных веществ и 1445 растворителей в интервале температур 252-368 К. Распределение этих данных, однако, крайне неравномерно: 25909 значений относятся к индивидуальным растворителям, 29158 — к бинарным смесям и лишь 1722 — к тройным. При этом число уникальных растворенных веществ, представленных в бинарных системах, составляет 174, а в тройных — три. Иными словами, заявленное разнообразие набора обеспечивается преимущественно его частью, посвященной растворимости в индивидуальных растворителях. Для компенсации нехватки экспериментальных данных авторы привлекли расчетные величины COSMO-RS, введя их в обучение через схему полуконтролируемой дистилляции «учитель — ученик». Приведенные

величины даны по сообщению авторов набора. Прямой пересчет опубликованного файла дает для смесевой части 34734 значения и 171 растворенное вещество.

Последнее обстоятельство заслуживает отдельного внимания как характерный признак областей с дефицитом данных. Недостающие экспериментальные значения замещаются расчетными, вследствие чего модель обучается воспроизводить результат квантово-химического или термодинамического расчета, а не эксперимент. Такой прием расширяет покрытие химического пространства, но привносит в модель систематические погрешности используемого расчетного метода, и достигаемая точность уже не может быть непосредственно сопоставлена с экспериментальной.

Таким образом, прогнозирование растворимости органических соединений в смесях растворителей находится в состоянии, которое область прогнозирования растворимости в воде прошла полтора-два десятилетия назад. Имеющиеся наборы данных немногочисленны, охватывают узкий круг соединений и растворителей, общепринятого способа сопоставления результатов не существует, а объем доступных экспериментальных данных не позволяет ни обучать модели высокой емкости, ни надежно оценивать их применимость к новым системам. Создание крупного стандартизованного массива экспериментальных данных о растворимости органических соединений в бинарных смесях растворителей с фиксацией состава смеси и температуры составляет вторую задачу настоящей работы.

2.2.5. Физико-химические методы как альтернатива

Прогнозирование растворимости методами, опирающимися на физическую модель процесса растворения, развивается параллельно подходам машинного обучения и представляет собой их естественную альтернативу. Подробный обзор этого направления дан в работе[37], здесь же достаточно сопоставить его возможности с рассмотренными выше методами.

Условно можно выделить две группы методов. К первой относятся эмпирические соотношения, использующие физически осмысленные параметры: обобщенное уравнение растворимости, рассмотренное выше, параметры растворимости Хансена[85], модель сольватационных дескрипторов Абрахама[86] и

методы групповых вкладов для расчета коэффициентов активности, прежде всего UNIFAC[87]. Эти подходы вычислительно необременительны и широко применяются при подборе растворителей[45], однако параметры их определяются подгонкой по экспериментальным данным, вследствие чего они разделяют с моделями машинного обучения зависимость от объема и качества этих данных.

Применимость подходов этой группы определяется, кроме того, доступностью термодинамических характеристик твердой фазы. Идеальная растворимость - та, которая отвечала бы раствору, где взаимодействие растворенного вещества с растворителем не отличается от взаимодействия его молекул между собой, - служит опорной величиной при разделении вкладов энергий кристаллической решетки и сольватации и вычисляется по температуре и энтальпии плавления, а отношение к ней экспериментальной растворимости дает коэффициент активности, характеризующий уже собственно взаимодействие с растворителем. Систематический сбор этих величин из литературы был предпринят для двухкомпонентных молекулярных кристаллов - сокристаллов и солей, - в которых состав твердой фазы целенаправленно изменяется для управления растворимостью. Соответствующая база данных, первоначально включавшая температуры плавления 1175 двухкомпонентных кристаллов[19], последовательно расширялась до 1947[88] и 2170 записей с добавлением энтальпий плавления и оценкой идеальной растворимости[89], а в последних версиях достигла 3032 записей и включает анализ влияния полиморфизма и стехиометрии на термодинамику сокристаллизации[90,91]. Наличие таких ресурсов существенно расширяет применимость методов, опирающихся на термодинамический цикл растворения, однако охватывают они лишь ту часть химического пространства, для которой измерения были выполнены, и приводятся, как правило, в составе публикаций в формате, не допускающем автоматического считывания.

Вторую группу образуют методы, позволяющие вычислить растворимость из первых принципов - прямым моделированием сосуществования фаз, расчетом химических потенциалов либо через термодинамический цикл сублимации и сольватации[92,93]. Промежуточное положение занимают континуальные модели сольватации, прежде всего COSMO-RS[94].

Принципиальное преимущество этих методов состоит в том, что они не требуют параметризации по экспериментальным значениям растворимости и потому не ограничены составом обучающей выборки, а, кроме того, дают недоступные статистическим моделям энтальпийный и энтропийный вклады в растворение.

Практические ограничения, однако, существенны. Физические методы применяются заметно реже подходов, основанных на данных; выполненные исследования охватывают небольшое число веществ и узкий диапазон условий, а достигаемая точность оказывается сопоставимой с точностью статистических моделей при вычислительной стоимости, на порядки большей, что делает такие расчеты непригодными для высокопроизводительного скрининга. Существенно и то, что рост вычислительных затрат не обеспечивает пропорционального выигрыша. При сравнении четырех способов построения конформеров для расчетов COSMO-RS - от простых процедур до основанных на теории функционала плотности - все они дали среднюю абсолютную ошибку около одной логарифмической единицы[95].

Независимость от экспериментальных данных является к тому же лишь частичной. Подтверждение достоверности расчета требует сопоставления с измерениями, а следовательно, тех же массивов надежных экспериментальных значений. Наблюдаемое в последние годы направление развития состоит в гибридизации подходов - использовании физически обоснованных величин, прежде всего энергий сольватации и характеристик плавления, в качестве промежуточных целевых переменных для моделей машинного обучения[69,96,97]. Физические методы, таким образом, не устраняют потребности в экспериментальных данных, а изменяют ее характер.

2.2.6. Выводы по разделу

Три рассмотренные задачи - прогнозирование растворимости в воде, в индивидуальных органических растворителях и в их бинарных смесях - опираются на единый методический аппарат и различаются почти исключительно объемом и однородностью доступных экспериментальных данных. Наглядное представление об этом различии дает сводка опубликованных наборов данных (табл. 1).

Для растворимости в воде за три с половиной десятилетия создано более десяти общедоступных ресурсов, причем большинство из них восходят к двум компиляциям конца 1980-х — начала 1990-х годов и представляет собой их последовательные расширения и пересборки. Для индивидуальных органических растворителей публичные наборы немногочисленны и на порядок меньше по объему: открытый краудсорсинговый проект и коллекция, собранная в связи с разработкой моделей на основе термодинамического цикла, содержат менее десяти тысяч измерений каждый, охватывая около сотни растворителей и порядка ста растворенных веществ, тогда как остальные данные собирались под конкретное исследование и отдельно не публиковались. Для бинарных смесей известны два набора, оба опубликованы в последние два года, причем заявленный объем в значительной мере обеспечивается их частью, связанной с растворимостью в индивидуальных растворителях, а не смесях: в собственно смесевых системах представлено 123 и 174 растворенных вещества, соответственно.

Таблица 1. Наборы экспериментальных данных о растворимости органических соединений, использовавшиеся для разработки моделей машинного обучения. Наборы, созданные в настоящей работе, в таблицу не включены и рассматриваются в разделе 4.1.

Набор	Год	Записей	Соединений	Растворителей
Растворимость в воде				
AQUASOL[17]	1989	~20000	~6000	1
PHYSPROP[57]	1994	~880	~880	1
Huuskonen[54]	2000	1297	1297	1
ESOL[56]	2004	2874	2874	1
Wang et al.[98]	2007	1708	1708	1
Solubility Challenge 1[58]	2008	132	132	1
Wiki-pS0[99]	2009	6355	3014	1
OCHEM[64]	2011 ^a	51602	~15000	1
AqSolDB[63]	2019	9982	9982	1
Solubility Challenge 2[49]	2019	132	132	1
Meng et al.[9]	2022	46014	—	1

Набор	Год	Записей	Соединений	Растворителей
AqueousSolu-Exp (SolProp)[69]	2022	11804 ^г	11804 ^г	1
AqSolDBc[10]	2024	8047	8047	1
Органические растворители				
ONSC [100]	2010	9730	3312 ^б	182 ^б
Boobier et al.[5]	2020	1611	—	3
CombiSolu-Exp (SolProp)[69]	2022	4953	115	97
Lee et al.[68]	2022	17536	12849	179
Смеси растворителей				
Bao et al.[83]	2024	28098	125	44
MixSolDB[77]	2025	34734 ^б	171 ^б	95 ^б

а - объем приведен по редакции 2022 года. б - часть набора, относящаяся к бинарным смесям, подсчитанная так же, как в табл. 5. в - подсчитано по канонизованным структурам, в источнике не приводится. г - одно значение на соединение при 298 К. Прочерк означает, что величина в источнике не приводится.

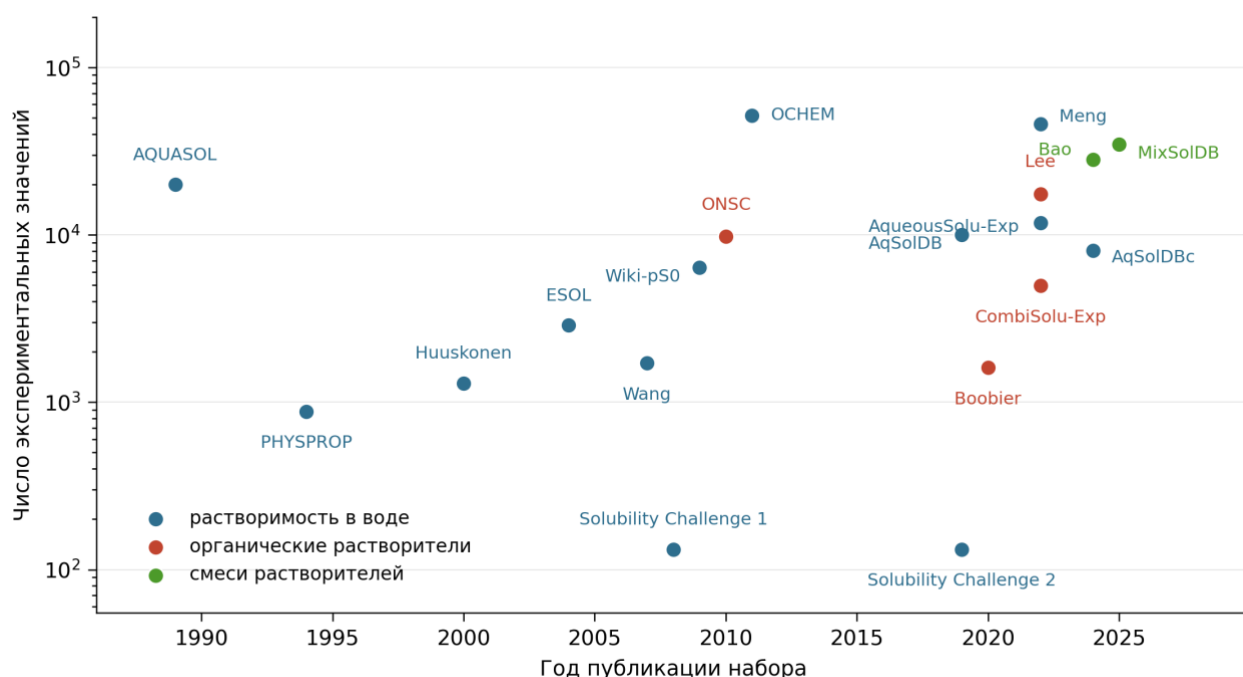


Рисунок 4. Наборы экспериментальных данных о растворимости, опубликованные с 1989 по 2025 год. Приведены наборы, перечисленные в табл. 1. Число экспериментальных значений отложено в логарифмическом масштабе.

Этому распределению данных (рис. 4) отвечает и достигнутый уровень результатов. Для органических растворителей последовательность методов воспроизводит путь, пройденный для растворимости в воде, с отставанием примерно в полтора десятилетия, причем химический охват исследований определяется доступностью литературных значений. Для смесей растворителей область находится в начальной стадии. Работы последних лет опираются на считанные наборы, один из которых многократно используется несколькими группами, а нехватка экспериментальных значений компенсируется расчетными. Предпринимаются и первые попытки распространить прогнозирование растворимости на комплексы переходных металлов[101], где к перечисленным затруднениям добавляется проблема машиночитаемого представления самих координационных соединений, рассматриваемая ниже.

2.3. Координационные соединения переходных металлов: представление и наборы данных

2.3.1. Машиночитаемое представление координационных соединений

Аппарат, рассмотренный выше, создавался применительно к органическим молекулам, строение которых полностью описывается набором атомов и связей между ними с указанием их кратности. Координационные соединения этому описанию не подчиняются, и это различие принципиальное.

Прежде всего, связь металл-лиганд является, как правило, донорно-акцепторной, и способ ее представления в линейных нотациях не определен однозначно. Она может быть записана как ковалентная, как ионная или не записана вовсе, причем каждый вариант приводит к иной формуле и иным вычисляемым дескрипторам (рис. 5). Не определены и степень окисления металла с распределением формального заряда между центральным атомом и лигандами. Далее, существенными характеристиками комплекса являются дентатность лигандов и природа донорных атомов, тогда как в линейной записи они не выражены. Наконец, координационная геометрия и связанная с нею стереохимия — например, различие геометрических изомеров октаэдрического комплекса (цис-/транс-), положение лигандов друг относительно друга — в нотации SMILES не передаются вовсе. Неполнота такого описания применительно к

комплексам переходных металлов констатируется прямо. Строка SMILES не передает ни стереоизомерии, ни донорно-акцепторного характера связывания[102].

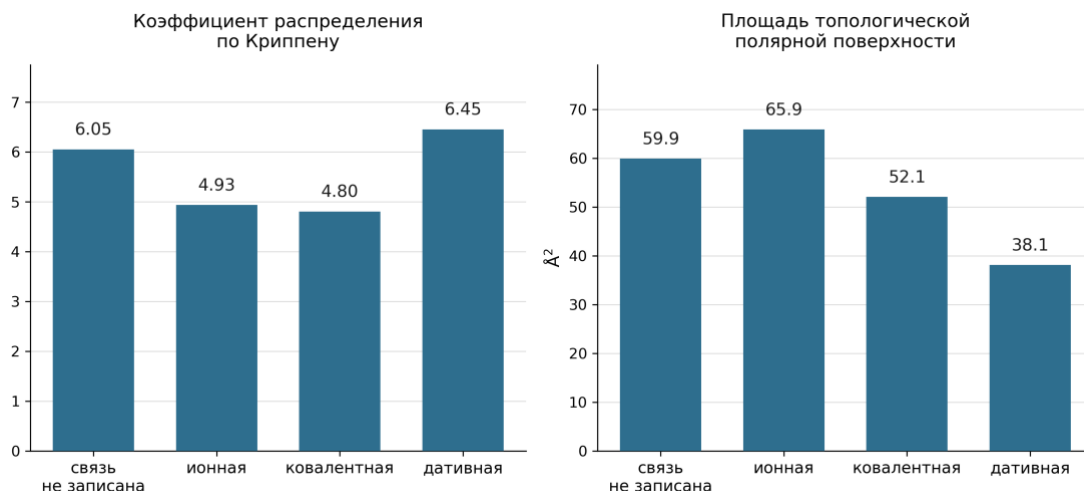


Рисунок 5. Зависимость вычисляемых дескрипторов от способа записи координационной связи на примере комплекса $[\text{Ir}(\text{ppy})_2(\text{acac})]$, где ppy - 2-фенилпиридин, acac - ацетилацетонат. Одно и то же соединение записано четырьмя употребительными способами, величины вычислены средствами RDKit. Расхождение возникает в том числе потому, что схемы расчета коэффициента распределения по Криппену и площади топологической полярной поверхности для переходных металлов не определены.

Эти затруднения не сводятся к выбору условных обозначений: сами величины, подлежащие записи, во многих случаях не определены однозначно. Так, распределение формального заряда между металлом и лигандами в кристаллографических данных приводится редко и непоследовательно. При построении набора зарядов лигандов BOS-Lig на основе почти 127 тысяч моноядерных комплексов из структурной базы данных удалось надежно приписать заряд лишь 66810 лигандам из 94581 выявленной уникальной структуры, причем для этого потребовалась итерационная процедура согласования зарядов комплекса, степеней окисления металла и совпадающих кристаллографических наблюдений[103]. Не менее сложным оказывается и определение способа координации: отдельные лиганды допускают десятки, а иногда и сотни вариантов связывания — монодентатного, бидентатного или полидентатного, — и выбор правильного затруднителен даже для опытного химика[104]. Наконец,

распространенное допущение о том, что данной паре металл-лиганд отвечает единственный способ связывания, нарушается для гемилабильных лигандов, изменяющих координацию в ходе процесса[105].

Следствия этой неопределенности описаны выше. Стандартные дескрипторы для таких соединений не вычисляются, и общепринятые процедуры подготовки данных предписывают их удаление[20]. Затруднение проявляется и на уровне инфраструктуры. В крупнейшем курируемом ресурсе данных о биологической активности ChEMBL металлосодержащие соединения исключены из процедуры стандартизации структур, поскольку принятый формат их хранения не выражает координационную связь[106]. Положение усугубляется тем, что комплексы переходных металлов отстают от органических соединений одновременно по всем трем условиям, определяющим успех работы с данными, — по наличию представительных наборов данных, по четкости постановки задачи и по наличию пригодного машиночитаемого представления[107].

Предпринимаемые попытки преодолеть это затруднение можно разделить на несколько направлений. Первое состоит в том, чтобы все же получить корректную запись SMILES, восстанавливая связность и заряды лигандов из пространственного строения комплекса. Соответствующая процедура построена как расширение известного алгоритма xyz2mol, преобразующего координаты в структурную формулу, и требует только координат атомов и суммарного заряда[108]. Второе направление отказывается от универсальной нотации в пользу специализированных дескрипторов металлоцентра. Предложен, в частности, отпечаток ELECTRUM, основанный на электронной конфигурации металла и применимый к соединению любого строения[107]. Третье направление предполагает разработку линейной нотации, изначально рассчитанной на координационные соединения. Нотация T-REX строится из разделяемых блоков, в заголовке которых указан центральный атом, а взаимное расположение лигандов кодируется через отношения транс-пар, что позволяет передать координационную геометрию, стереохимию и топологию донорных атомов[109].

Наконец, ряд подходов отказывается от линейной записи вовсе. Графовые нейронные сети способны работать со строением комплекса непосредственно, и для них разработаны архитектуры, учитывающие особое положение металлического центра. Предложены модели, прогнозирующие по такому представлению координационное окружение металла[110], а также схемы с явным механизмом внимания к металлу, применяемые для высокопроизводительного прогнозирования свойств[111]. Отдельное направление связано с тем, что ни линейная запись, ни молекулярный граф не передают пространственного строения: разработаны программные средства построения трехмерных структур комплексов — Architector, охватывающий произвольные металлы периодической системы[112], и MetalloGen[113], — а также методы достраивания полной геометрии комплекса по ее фрагменту[114].

О том, насколько далеко область находится от единого решения, свидетельствует прямое сопоставление. Для одного и того же подмножества комплексов записи SMILES были получены тремя независимыми способами — восстановлением из координат, анализом натуральных связевых орбиталей по результатам квантово-химического расчета и исправлением строк, приведенных в структурной базе данных, — и согласие между любыми двумя из этих способов составило немногим более 70 %[108]. Иными словами, почти для трети соединений три существующие процедуры дают различные машиночитаемые описания одного и того же вещества.

Существенно и то, что все перечисленные подходы предложены в последние два-три года и ни один из них не получил положения общепринятого. Практическое следствие для настоящей работы состоит в том, что при создании баз данных о свойствах комплексов приходится принимать самостоятельное решение о способе их представления, а сопоставимость с чужими наборами данных этим решением ограничивается.

2.3.2. Наборы данных по комплексам переходных металлов

Обеспеченность координационных соединений данными устроена своеобразно. Сведения о строении накоплены в изобилии, расчетные характеристики получены в сопоставимом объеме, а экспериментально измеренные свойства остаются редкими.

Структурная составляющая наиболее полна. Кембриджская структурная база данных содержала на середину 2024 года свыше 260 тысяч уникальных мооядерных комплексов переходных металлов. Значительная их часть послужила основой для расчетных наборов. Около 86 тысяч соединений снабжены квантово-химическими характеристиками основного состояния в наборе tmQM[115], а для 74 тысяч рассчитаны характеристики возбужденных состояний в наборе tmQMg*[116].

Расчетные наборы продолжают расширяться и уточняться. В наборе BOS-TMC для 159 тысяч экспериментально охарактеризованных мооядерных комплексов, отобранных из структурной базы данных, характеристики рассчитаны в нескольких спиновых состояниях, что дало в совокупности около 344 тысяч сочетаний «комплекс — спиновое состояние». При этом экспериментальные координаты тяжелых атомов сохранялись, а суммарный заряд комплекса приписывался по итерационной процедуре[117].

Наборы же экспериментальных значений на несколько порядков меньше (рис. 6) и, как правило, ограничены отдельным свойством и узким кругом соединений. Их объем измеряется сотнями, реже тысячами записей[102]. Прямое следствие этого - вынужденный переход от прогнозирования значения к прогнозированию класса. Разобран он в разделе 2.5.2 на примере, относящемся к биологической активности комплексов.

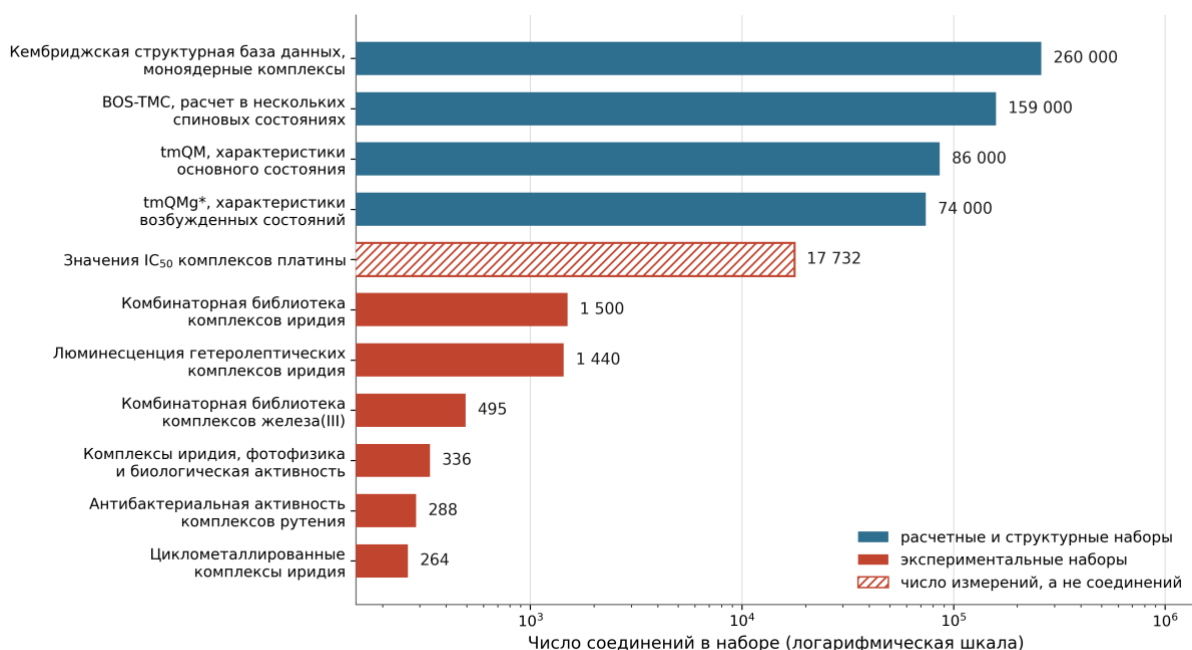


Рисунок 6. Объемы наборов данных по комплексам переходных металлов. Расчетные и структурные наборы превосходят экспериментальные на два-три порядка. Заштрихованная полоса отвечает числу измерений, а не числу соединений. Данные приведены по источникам, указанным в тексте раздела.

Другой пример такого набора дают комплексы редкоземельных элементов и трехвалентных актинидов. Сведения о константах устойчивости были собраны из базы данных NIST46 и первичной литературы, причем отбирались лишь комплексы состава 1:1, измеренные в водной среде и снабженные указанием температуры и ионной силы, а при наличии нескольких значений для одних и тех же условий предпочтение отдавалось потенциометрическому титрованию. Небольшой объем данных был отчасти возмещен устройством модели. Графовая нейронная сеть получала на вход одновременно граф лиганда и вектор, описывающий металл, к которому присоединялись условия опыта, что позволило обучать единую модель на всех металлах сразу. Представление металла строилось методами, заимствованными из обработки естественного языка, и было дополнительно обучено на кристаллографических данных, поскольку исходные векторы охватывали элементы лишь до урана. Работоспособность подхода проверялась исключением из обучения всех соединений одного металла с последующим прогнозированием для него, что показало перенос закономерностей на не встречавшийся при обучении катион. Авторы

отмечают при этом, что близкие по строению лиганды неизбежно попадают одновременно в обучающую и в проверочную выборки, поскольку одни и те же хелатирующие остовы применяются с разными металлами, а полное разделение по лигандам не оставило бы достаточного объема для обучения[118].

Следует отметить и то, что модели, построенные на материале органической химии, не переносятся на комплексы автоматически. При расширении эталонного набора реакций органических соединений комплексами переходных металлов обнаруживается принципиальная неспособность генеративных моделей обобщаться на элементы, не встречавшиеся при обучении[119]. Соотношение объемов имеет прямое следствие для методологии. Поскольку расчетных характеристик доступно на порядки больше, чем измеренных, модели для комплексов переходных металлов зачастую обучаются воспроизводить результат квантово-химического расчета, а не эксперимента. Показателен пример прогнозирования светопоглощения комплексов платины, иридия, рутения и родия в терапевтическом окне 600-850 нм. Методы машинного обучения применяются здесь именно как замена расчету по нестационарной теории функционала плотности, вычислительно затратному и медленному[120]. Проявления этого подхода рассматриваются в последующих разделах применительно к люминесцентным свойствам и цитотоксичности.

Наряду с этим в последние годы оформился и противоположный подход. Вместо извлечения данных из литературы исследовательская группа получает их самостоятельно, синтезируя библиотеку комплексов комбинаторным способом и подвергая ее сплошному испытанию[121]. Комбинаторный синтез в этом случае проводится, как правило, в одну стадию и без выделения индивидуальных соединений, а полученные растворы сразу передаются на скрининг.

Масштабы, достигаемые таким образом, сопоставимы с объемом наборов, собираемых из литературы. В отдельных исследованиях получены и испытаны библиотеки из 78[122], 336[123] и 495 соединений[124]. Для циклометаллированных комплексов иридия сообщается о библиотеках в 264 соединения, а при использовании одностадийного комбинаторного подхода — о выборках, достигающих полутора тысяч (рис. 7)[125,126]. Разработаны также полуавтоматизированные платформы синтеза и

идентификации таких библиотек[127] и приемы комбинаторной химии, приспособленные к получению металлокомплексов[128]. Тем же способом получают и физико-химические характеристики. Для 1440 гетеролептических комплексов иридия, синтезированных параллельным комбинаторным методом, измерены спектры испускания и времена жизни возбужденного состояния, причем люминесцентными оказались 72% соединений, максимумы испускания охватили диапазон от 459 до 652 нм, а времена жизни — от 0.1 до 14 мкс[129].

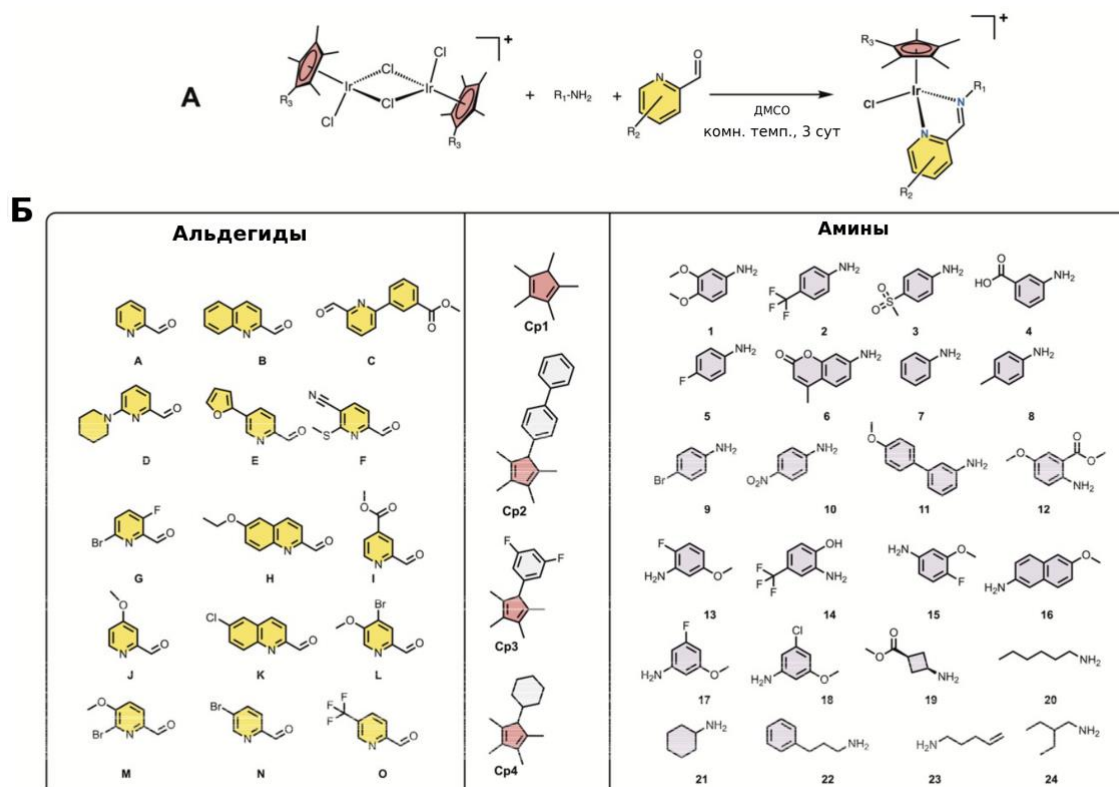


Рисунок 7. Одностадийный комбинаторный синтез 1440 комплексов иридия(III). А — общая схема реакции. Б — набор строительных блоков: 15 пиколинальдегидов, 24 амина и 4 производных цикlopentadiенильного лиганда[126].

Существенно, однако, что данные, получаемые этим способом, имеют иную природу, нежели извлекаемые из литературы. Они внутренне согласованы, поскольку получены в единых условиях одной группой, но охватывают узкую область химического пространства, определяемую выбранной синтетической схемой, и, как правило, содержат не количественные значения, а результат первичного скрининга.

2.4. Люминесцентные свойства циклометаллированных комплексов иридия(III)

2.4.1. Фотофизические свойства и их настройка структурой лигандов

Октаэдрические циклометаллированные комплексы иридия(III) занимают особое положение среди люминесцентных соединений благодаря сочетанию яркого испускания, широкой настраиваемости спектральных характеристик и высокой химической устойчивости.

Природа их люминесценции определяется тяжелым атомом металла. Сильное спин-орбитальное взаимодействие обеспечивает эффективную интеркомбинационную конверсию возбужденного синглетного состояния в триплетное, вследствие чего испускание происходит из триплетного состояния и представляет собой фосфоресценцию. Запрет по спину при этом частично снимается тем же спин-орбитальным взаимодействием[130]. Существенным обстоятельством является наличие конкурирующих металл-центрированных состояний, заселение которых не приводит к испусканию, а напротив, ослабляет связи металл-лиганд в возбужденном состоянии и снижает устойчивость соединения. Существенно и то, что излучательное состояние имеет смешанный характер и включает состояния с перенос заряда с металла на лиганд, вследствие чего энергия его зависит от полярности среды, а значения, измеренные в разных растворителях, между собой напрямую несопоставимы[131].

Практически важным свойством этих комплексов является то, что положение и эффективность испускания определяются строением лигандного окружения. Энергия триплетного состояния задается главным образом циклометаллированным лигандом, тогда как дополнительный лиганд позволяет более тонко настраивать фотофизические и окислительно-восстановительные свойства. Введение заместителей различной электронной природы, расширение сопряженной системы и изменение конформационной жесткости лиганда смещают максимум испускания и влияют на квантовый выход. Именно это обстоятельство делает осмысленной постановку задачи о прогнозировании фотофизических характеристик непосредственно по структурным формулам лигандов. Общие закономерности этого влияния приведены на рис. 8.

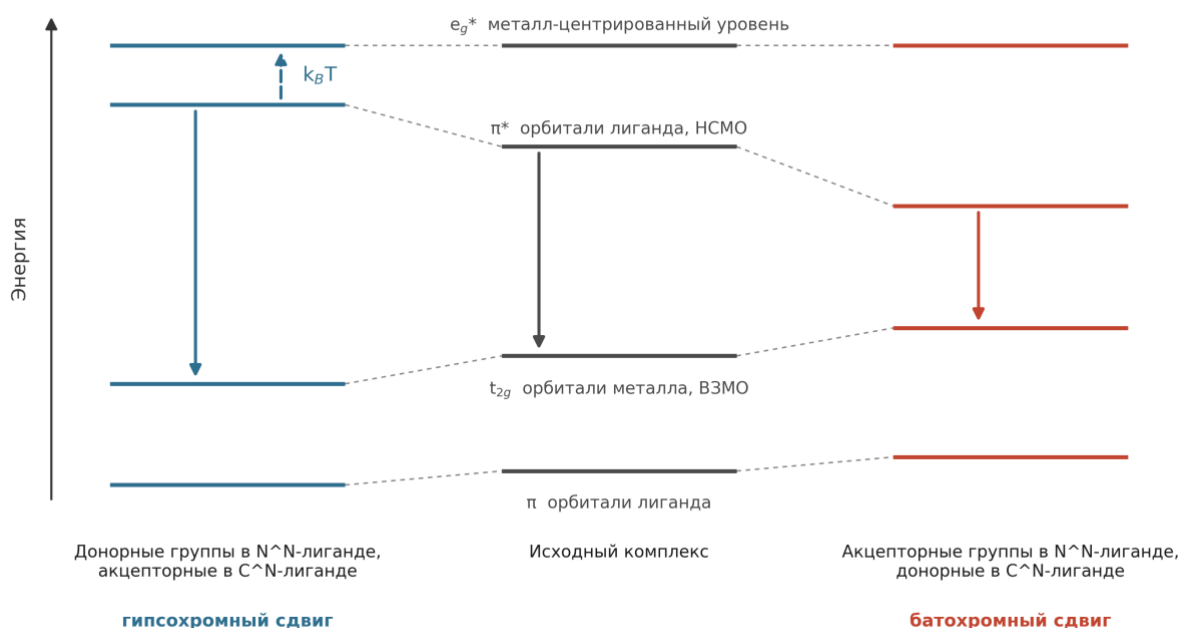


Рисунок 8. Влияние электронной природы заместителей на энергии граничных орбиталей бис-циклометаллированных комплексов иридия(III). Сокращение зазора до металл-центрированного уровня при гипсохромном сдвиге делает возможным его термическое заселение. Построено по данным[132].

Эта настройка, однако, не одинаково успешна во всех областях спектра. В зеленой и желтой областях получены эмиттеры с высокой внешней квантовой эффективностью, тогда как смещение к обоим краям видимого диапазона сопровождается существенным падением эффективности. В синей области становится заметным термическое заселение неизлучательных металл-центрированных состояний, что снижает как квантовый выход, так и долговременную устойчивость эмиттера. В красной и ближней инфракрасной областях квантовые выходы падают вследствие роста скорости безызлучательной релаксации при уменьшении энергетического зазора — закономерность, известная как закон энергетической щели[133].

Применения этих соединений определяются перечисленными свойствами. Они используются в качестве эмиттеров фосфоресцентных органических светодиодов, фотосенсибилизаторов в солнечных элементах на основе красителей[134], фотокатализаторов, а также агентов люминесцентной визуализации клеток и

фотодинамической терапии. Каждое из этих применений предъявляет собственные требования к длине волны испускания и квантовому выходу, вследствие чего разработка комплекса под конкретную задачу сводится к целенаправленному поиску в пространстве возможных лигандных окружений.

Этот поиск дорог. Получение и характеристика каждого нового комплекса требуют значительных синтетических усилий, а число гипотетических сочетаний циклометаллированного и дополнительного лигандов практически неограниченно. Именно это обстоятельство создает запрос на средства предварительной оценки свойств до синтеза.

2.4.2. Прогнозирование фотофизических свойств

Естественным средством такой оценки представляется квантово-химический расчет. Однако применительно к возбужденным состояниям комплексов иридия нестационарная теория функционала плотности оказывается затруднительной одновременно в двух отношениях — по достигаемой точности и по вычислительной стоимости, — что и делает сплошной виртуальный перебор вариантов практически неосуществимым[135]. Сходное соображение высказывается и применительно к оценке светопоглощения комплексов платины, иридия, рутения и родия, где методы машинного обучения предлагаются именно как замена ресурсоемкому расчету[120].

К примеру, для чисто органических красителей задача решается на несопоставимо большем объеме данных. На выборке из 25608 измерений молярного коэффициента поглощения для 20498 соединений пяти классов, извлеченных из литературы 1980-2021 годов, построена модель случайного леса, воспроизводящая логарифм коэффициента поглощения со среднеквадратичной ошибкой 0.26 логарифмической единицы при пятикратной перекрестной проверке[136].

Обращение к методам машинного обучения для прогнозирования свойств комплексов металлов представляется в этих условиях закономерным, и соответствующие работы появились. Разработаны ансамблевые модели прогнозирования фотофизических характеристик фосфоресцентных комплексов платины(II)[137], а также протокол, объединяющий порождение структур,

высокопроизводительный виртуальный скрининг и быструю оценку свойств. Для циклометаллированных комплексов платины(II) таким образом получено свыше 3600 структур, доступных для синтеза, на которых обучены модели прогнозирования константы излучательного распада и квантового выхода[138]. Для комплексов иридия(III) с испусканием в красной и ближней инфракрасной областях выполнено исследование, сочетающее машинное обучение с расчетом и экспериментом[139], а на ограниченном объеме экспериментальных данных о люминесценции гетеролептических комплексов иридия(III) построены модели, восстанавливающие полный спектр испускания[140].

Существующие работы, однако, обнаруживают две системные слабости. Первая состоит в узости используемых выборок. Наиболее обширный массив экспериментальных данных о фотофизике комплексов иридия получен параллельным комбинаторным синтезом и охватывает 1440 гетеролептических соединений, охарактеризованных единообразно[129]. Этот массив внутренне согласован, но ограничен единственным структурным типом, и именно он служит основой для последующих работ. На подмножестве из 1380 комплексов обучены модели, прогнозирующие среднюю энергию испускания, время жизни возбужденного состояния и интегральную интенсивность полосы с точностью, сопоставимой с расчетной или превосходящей ее[135]. Тот же массив положен в основу и других работ: на выборке из 1046 соединений, содержащих два одинаковых циклометаллированных лиганда и один вспомогательный, построены составные модели прогнозирования длины волны испускания[141]. Применимость таких моделей за пределами исходной серии не установлена.

Вторая слабость связана с природой используемых величин. Поскольку экспериментальных данных мало, а расчетных много, значительная часть работ опирается на результаты квантово-химического расчета — либо как на целевые величины, либо как на входные признаки. Так, при отборе перспективных эмиттеров иридия(III) предобученное представление дообучалось на рассчитанных энергиях граничных орбиталей и адиабатической энергии возбуждения триплетного состояния[142], а поиск новых красных фосфоров велся по дескрипторам

возбужденных состояний[143]. Масштаб доступных расчетных ресурсов этому способствует. Характеристики возбужденных состояний рассчитаны для 74 тысяч комплексов переходных металлов[116]. Существенно, что даже наиболее точные из построенных таким образом моделей требуют предварительного расчета для каждого нового соединения[135], вследствие чего выигрыш в вычислительной стоимости оказывается неполным, а сама модель наследует систематические погрешности использованного расчетного метода.

Таким образом, применительно к люминесцентным свойствам циклометаллированных комплексов иридия(III) складывается положение, при котором обширный массив экспериментальных данных рассеян по первичной литературе и не сведен в машиночитаемую форму, тогда как разрабатываемые модели опираются либо на узкие однородные серии, либо на расчетные величины. Восполнение этого пробела — систематический сбор и стандартизация экспериментальных данных о люминесцентных свойствах структурно разнообразных бис-циклометаллированных комплексов иридия(III) и построение на их основе моделей прогнозирования — составляет одну из задач настоящей работы.

2.5. Цитотоксичность комплексов переходных металлов

2.5.1. Противоопухолевая активность комплексов

Поиск новых противоопухолевых средств остается одной из наиболее востребованных задач медицинской химии. Подавляющее большинство испытываемых соединений отсеивается уже на стадии исследований на клеточных культурах или на животных, и лишь ничтожная часть доходит до клинических испытаний, вследствие чего ценность приобретают средства предварительного отбора, позволяющие сократить число заведомо бесперспективных экспериментов.

Комплексы переходных металлов занимают в этом поиске особое положение, поскольку только для них клиническая состоятельность подтверждена практикой в качестве цитотоксических средств. Три препарата платины(II) — цисплатин, карбоплатин и оксалиплатин — применяются повсеместно и входят в схемы лечения широкого круга злокачественных новообразований, еще три — недаплатин, лобоплатин и гептаплатин — разрешены к применению соответственно в Японии,

Китае и Республике Корея[144]. Отечественной разработкой этого ряда является циклоплатам - S-малато(аммин)(циклопентиламин)платина(II), синтезированная в Институте общей и неорганической химии имени Н. С. Курнакова в 1982 году[145]. По итогам клинических испытаний, показавших действенность при раке яичников, мезотелиоме плевры, полирезистентной множественной миеломе и гормонорезистентном раке предстательной железы, препарат рекомендован Фармакологическим комитетом Министерства здравоохранения Российской Федерации к применению в клинике[146]. Присущие им ограничения — узкий спектр действия, тяжелые побочные эффекты и развитие приобретенной устойчивости — и обусловили обращение к другим металлам.

Внимание к таким соединениям заметно возросло. Так, число ежегодно публикуемых работ, отвечающих в базе данных PubMed запросу «metal complexes anticancer», выросло с 53 в 2000 году до 594 в 2025 году, то есть более чем в десять раз (рис. 9). Эта величина зависит от выбора ключевых слов и полноты охвата предметной области не отражает, но об изменении интереса к предмету судить позволяет. Это продвижение оказалось, однако, гораздо более скромным, чем можно было бы заключить по объему публикаций. За четверть века стадии клинических испытаний достигли лишь единичные соединения, и все они содержат рутений.

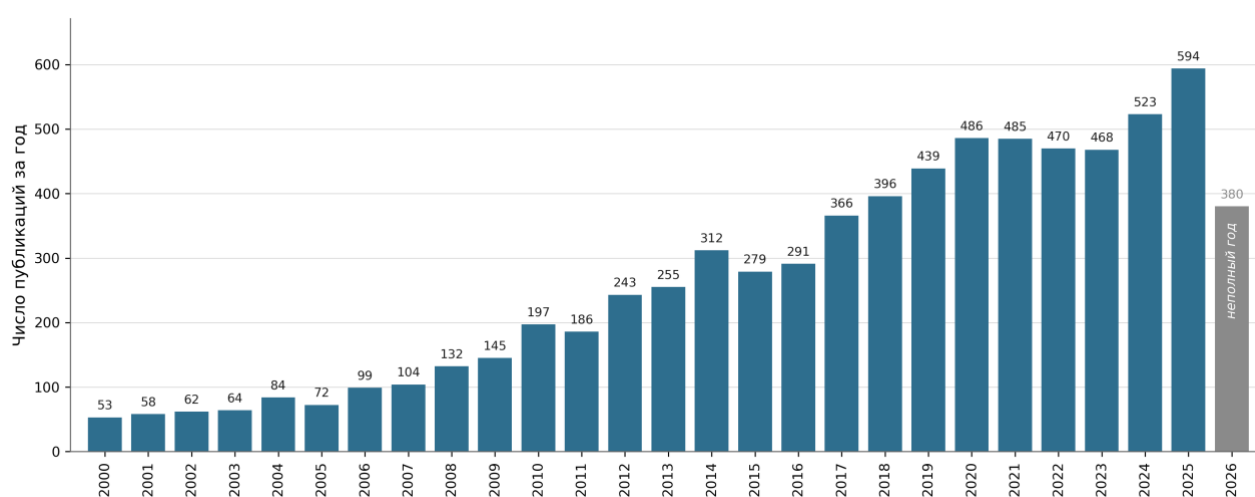


Рисунок 9. Число публикаций, отвечающих в базе данных PubMed запросу «metal complexes anticancer», по годам. Данные за 2026 год неполны, поиск проведен в сентябре.

Первым был NAMI-A, изучавшийся как ингибитор метастазирования. После испытаний первой фазы[147] соединение исследовали совместно с гемцитабином у больных немелкоклеточным раком легкого, получавших ранее иную терапию. Итог оказался отрицательным. Сочетание переносилось лишь умеренно и было менее действенным, чем один гемцитабин, вследствие чего разработка была прекращена[148]. Судьба второго кандидата сложилась иначе. Исходное соединение KP1019 оказалось слишком малорастворимым в воде, и на испытания была выведена его натриевая соль KP1339, известная также под обозначениями NKP-1339, IT-139 и BOLD-100, действие которой связывают с подавлением индукции белка GRP78 и нарушением ответа на накопление неправильно свернутых белков. В испытаниях первой фазы, охвативших 46 больных с распространенными солидными опухолями, препарат переносился удовлетворительно, однако противоопухолевое действие оказалось скромным. Продолжительный частичный ответ отмечен лишь у одного больного, еще у девяти - стабилизация заболевания[149].

Именно это соединение продвинулось дальше прочих. Под обозначением BOLD-100 оно испытывается в сочетании со схемой FOLFOX при распространенных опухолях желудочно-кишечного тракта, и по результатам, полученным у больных метастатическим колоректальным раком после неудачи не менее двух линий терапии, отмечены как противоопухолевая активность, так и приемлемая переносимость[150]. В настоящее время проводится рандомизированное испытание второй фазы, в котором сочетание BOLD-100 с FOLFOX сравнивается с одной лишь этой схемой[151]. Отдельно стоит фотосенсибилизатор TLD-1433 - первое соединение рутения(II), доведенное до клинических испытаний в качестве средства фотодинамической терапии. Это соединение представляет собой хлорид катионного комплекса, содержащего имидазо[4,5-f][1,10]фенантролиновый лиганд с присоединенным α -тертиенильным хромофором и два лиганда 4,4'-диметил-2,2'-бипиридина, и было отобрано из небольшой серии близких по строению соединений по совокупности фотофизических свойств и наблюдавшейся в опытах активности[152]. В испытании фазы Ib при немышечно-инвазивном раке мочевого пузыря, не отвечающем на вакцину БЦЖ, лечение переносилось удовлетворительно, а у двух из трех больных, получивших полную дозу, достигнут полный ответ, сохранявшийся к 18 месяцам[39]. Испытание

второй фазы, охватившее 75 больных в 11 центрах Канады и США, продолжается[153,154]. Строение перечисленных соединений приведено на рис. 10.

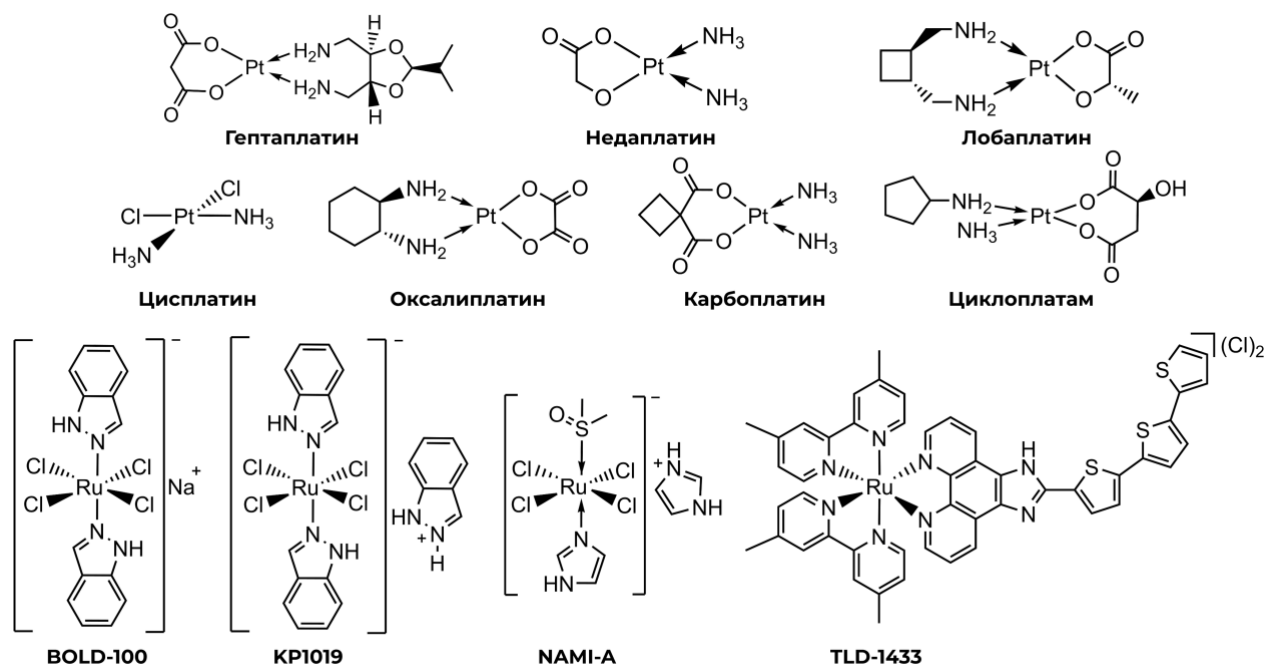


Рисунок 10. Строение комплексов металлов, разрешенных к применению и дошедших до клинических испытаний. В двух верхних рядах приведены препараты платины(II), в нижнем кандидаты на основе рутения.

Таким образом, среди цитотоксических средств ни одно соединение неплатиновых металлов до настоящего времени не разрешено к применению, а перечень дошедших хотя бы до испытаний на человеке исчерпывается несколькими названиями[155,156]. Разрешенные препараты других металлов существуют, однако действуют иначе. Ауринофин, комплекс золота(I), применяется при ревматоидном артрите, а средства на основе радия-223 и лютеция-177 действуют за счет излучения, тогда как лиганд служит для доставки к мишени. Хронология выхода в клинику приведена на рис. 11.

Разрешены к применению

комплексы платины(II)

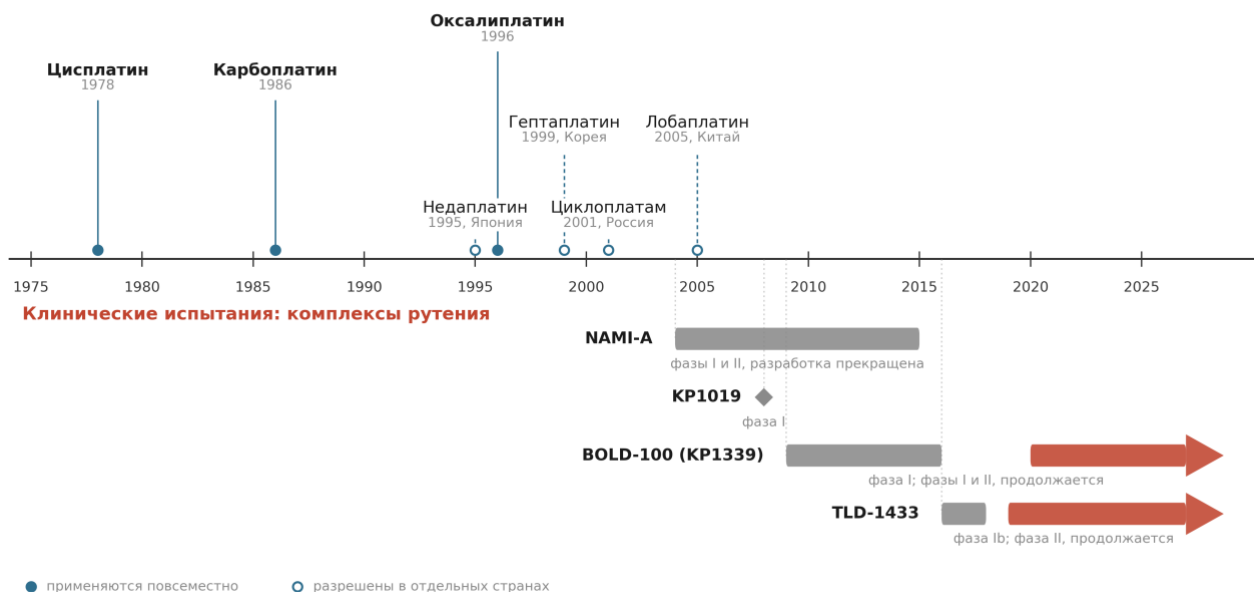


Рисунок 11. Комплексы металлов в клинической практике и в клинических испытаниях. Для разрешенных препаратов указан год выхода на рынок, для циклоплатина год публикации, сообщающей о рекомендации Фармакологического комитета. Для комплексов рутения отрезки отвечают испытаниям, зарегистрированным в ClinicalTrials.gov, а для NAMI-A и KP1019, испытания которых предшествовали обязательной регистрации, отмечены годы публикации отчетов.

Все остальное множество исследуемых соединений — комплексы иридия(III), рассматриваемые и для химиотерапии, и для фотодинамической терапии, соединения золота, меди, кобальта, а также существенно менее изученные производные осмия, родия и рения — остается на доклинической стадии[156,157]. Это соотношение существенно для настоящей работы. Именно на доклинической стадии, где отбор ведется по результатам испытаний на клеточных культурах, сосредоточен основной объем накопленных экспериментальных данных, и именно здесь расчетные средства предварительного отбора способны принести наибольшую пользу.

Среди остающихся на доклинической стадии соединений особое внимание привлекают комплексы иридия(III). Причины тому следующие. Такие соединения устойчивы к действию воздуха и влаги, сохраняют активность в отношении опухолей, приобретших устойчивость к препаратам платины, и допускают тонкую настройку как

биологических, так и фотофизических свойств путем изменения лигандного окружения, что открывает доступ к широкому многообразию структурных типов и механизмов действия, в том числе к избирательному накоплению в отдельных органеллах. Существенно и то, что одни и те же соединения способны выступать и как цитотоксические средства, и как фотосенсибилизаторы, и как средства визуализации. Яркость свечения и устойчивость к выцветанию позволяют наблюдать за поведением соединения в отдельной клетке, а образование активных форм кислорода при облучении - вызывать ее гибель, вследствие чего лечебное и диагностическое действия оказываются совмещены в одной молекуле[158,159]. Последнее обстоятельство связывает настоящий раздел с изложенным в разделе 2.4. Люминесцентные и противоопухолевые свойства комплексов иридия(III) определяются одним и тем же строением, и оба они составляют предмет настоящей работы.

Основной количественной характеристикой при первичном отборе служит цитотоксичность, выражаемая через концентрацию полумаксимального ингибирования IC_{50} — концентрацию, при которой подавление пролиферации клеток достигает половины предельного. Наряду с ней используется индекс селективности, определяемый как отношение IC_{50} на нормальных клетках к IC_{50} на опухолевых и характеризующий избирательность действия.

Следует оговорить, что представления о токсичности соединений металлов нередко опираются на сведения, относящиеся к самим металлам или их простым солям, тогда как биологическое действие координационного соединения определяется комплексом в целом и может существенно отличаться от действия составляющих его компонентов[160]. Это обстоятельство дополнительно повышает ценность экспериментально измеренных величин, относящихся к конкретным соединениям.

Существенно, что эта величина определена лишь совместно с условиями измерения. IC_{50} зависит от клеточной линии, времени экспозиции, примененной методики оценки жизнеспособности клеток, а для фотоактивируемых соединений — еще и от условий облучения. Как следствие, значения, полученные в разных экспериментальных условиях, сопоставимы лишь с оговорками. Количественная оценка вносимого при этом шума, выполненная для данных о биологической

активности, показала, что почти две трети пар значений, измеренных для одних и тех же соединений в разных условиях, различаются более чем на 0.3 логарифмической единицы, а более четверти — более чем на порядок[12]. Изложенное в разделе 2.1.3 применительно к утрате сведений об условиях эксперимента относится к этим данным в полной мере.

Наконец, экспериментальное определение цитотоксичности для обширных наборов соединений остается трудоемким и дорогостоящим, что и создает запрос на средства предварительной расчетной оценки.

2.5.2. Прогнозирование цитотоксичности

Применительно к органическим соединениям эта задача решается достаточно давно и разнообразными средствами. Построены модели, прогнозирующие активность соединений в отношении раковых клеточных линий по сочетанию молекулярных и геномных признаков[161]. Методами глубокого обучения выявляются структурные фрагменты, ответственные за цитотоксическое действие[162]. Разработаны инструменты раннего отсева цитотоксичных соединений, предназначенные для применения на начальных стадиях разработки[163]. Строятся модели и для отдельных типов клеток, причем особое внимание уделяется интерпретации выявленных зависимостей[164], а также модели, привлекающие наряду со строением сведения о транскриптоме[165]. Опорой для всех этих работ служат крупные ресурсы данных о биологической активности, содержащие миллионы измерений для сотен тысяч соединений[28], а также результаты систематического высокопроизводительного профилирования обширных библиотек соединений на клеточных линиях[166].

Комплексы переходных металлов представлены в этой области несопоставимо беднее, и причина, как отмечается в литературе, состоит прежде всего в трудностях описания координационной связи. Это затруднение проявляется не только на уровне методов, но и на уровне самой инфраструктуры хранения данных. Так, процедура курирования, применяемая при подготовке крупнейшего курируемого ресурса сведений о биологической активности, исключает металлоорганические и металлосодержащие соединения из стандартизации. Для них выставляется соответствующий признак, и приведение структуры к единому виду не производится

вовсе — на том основании, что используемый формат хранения структурных данных не позволяет корректно передать координационную связь[106]. По той же причине для таких соединений отключена и процедура выделения родительской структуры. Авторы приводят в пример транс-дихлородиаминоплатину, депонированную в виде набора несвязанных фрагментов, обычное удаление протонов у которой оставило бы один лишь ион платины. Способы преодоления этого затруднения — дескрипторы, не содержащие прямых сведений о строении, порождение трехмерных структур, применение больших языковых моделей, графовые сети, распознающие координационное окружение, и новые машиночитаемые форматы, учитывающие присутствие металла, — рассмотрены в разделе 2.3 настоящего обзора.

Второе затруднение состоит в отсутствии представительных наборов экспериментальных данных о свойствах металлоорганических и координационных соединений. Восполняется этот недостаток двумя путями, также описанными выше: комбинаторным получением библиотек с последующим сплошным испытанием и систематическим извлечением данных из первичной литературы. Применительно к цитотоксичности первый из них применен, в частности, для комплексов железа(III), библиотека которых была получена комбинаторно и испытана целиком, что позволило выявить соединения высокой активности[124].

Немногочисленные выполненные работы отчетливо обнаруживают следствия этого дефицита. Так, при построении модели антибактериальной активности комплексов рутения имевшихся 288 соединений оказалось недостаточно для количественного прогнозирования, вследствие чего измеренные величины были преобразованы в бинарную классификацию[167]. Переход от прогнозирования значения к прогнозированию класса представляет собой типичный для этой области вынужденный прием, обусловленный объемом доступных данных. Предпринимаются и попытки одновременного прогнозирования спектральных свойств и цитотоксичности для одного ряда соединений - комплексов платины с присоединенным красителем BODIPY[168].

Наиболее полно оба затруднения преодолены в недавней работе, посвященной комплексам платины. Вручную собранный набор из 17732 значений IC₅₀ сочетался в

ней с представлением, передающим связи металл - лиганд и заряды, на основании чего построена модель глубокого обучения, примененная затем для отбора из библиотеки в 214373 соединения. Девятнадцать отобранных комплексов были синтезированы и испытаны, причем прогноз оправдался в 72 % случаев, а созданное методами порождающего конструирования соединение преодолевало устойчивость к цисплатину как в опытах на клеточных культурах, так и на животных. Результаты сделаны доступными в виде общедоступного ресурса[169]. Эта работа показывает, что при наличии данных достаточного объема и пригодного представления комплексы переходных металлов поддаются тем же методам, что и органические соединения.

Сходное по построению исследование выполнено и для комплексов иридия(III). Библиотека из 336 соединений была получена и испытана в едином потоке, причем для каждого соединения определялись образование активных форм кислорода, липофильность, цитотоксичность в темноте и при облучении, поглощение клетками и внутриклеточное распределение. На полученном таким образом внутренне согласованном наборе обучены модели, прогнозирующие как физико-химические, так и биологические свойства по дескрипторам, вычисляемым полуэмпирическим методом сильной связи, после чего проведен виртуальный отбор из библиотеки более чем в 200000 соединений. Обнаружилось, что области химического пространства, отвечающие противоопухолевой и антибактериальной активности, заметно расходятся[123].

Таким образом, применительно к цитотоксичности комплексов переходных металлов положение определяется совместным действием двух обстоятельств: отсутствием общепринятого машиночитаемого представления таких соединений и отсутствием систематически собранных экспериментальных данных сопоставимого с органическими соединениями объема. Устранение второго из них — систематический сбор, стандартизация и верификация экспериментальных данных о цитотоксичности комплексов переходных металлов и построение на их основе моделей классификации — составляет одну из задач настоящей работы.

2.6. Заключительные замечания и постановка задачи

Рассмотренные в настоящем обзоре предметные области - растворимость органических соединений, люминесцентные свойства комплексов иридия(III) и цитотоксичность комплексов переходных металлов - на первый взгляд объединены лишь применяемым к ним расчетным аппаратом. Сопоставление их, однако, обнаруживает общность более глубокую. Во всех трех случаях развитие расчетных методов упирается в состояние экспериментальных данных, и обстоятельства, этому препятствующие, повторяются от области к области.

Эти обстоятельства, разобранные выше применительно к каждой из областей, суть три: отсутствие общепринятого машиночитаемого представления координационных соединений, недостаточный объем систематически собранных экспериментальных данных и утрата сведений об условиях измерения при составлении сводных наборов. Существенно, что второе из них не сводится к первому. Лучшим тому доказательством служит растворимость органических соединений: представление здесь никаких затруднений не вызывает, а данные тем не менее оставались разрознены по тысячам публикаций и относились подавляющей частью к воде. Причины, таким образом, действуют независимо, и устранение одной из них не устраняет другой.

Из изложенного вытекает, что восполнение дефицита структурированных экспериментальных данных представляет собой самостоятельную научную задачу, решение которой предшествует построению моделей и определяет предел их возможностей. Этим и определяются цель и задачи настоящей работы, сформулированные во введении.

Следует оговорить пределы такого подхода. Создание баз данных устраняет лишь второе из названных выше обстоятельств. Затруднение, связанное с представлением координационных соединений, им не снимается, а обходится. Собранные сведения приводятся к виду, пригодному для работы существующими средствами, но общепринятого способа описания координационной связи не возникает. Не устраняется полностью и третье обстоятельство, поскольку условия измерения могут быть сохранены лишь в той мере, в какой они приведены в исходных публикациях.

3. Материалы и методы

3.1. Сбор и стандартизация экспериментальных данных

Все четыре набора данных собирались по единому порядку, состоящему из четырех последовательных ступеней: отбор рецензируемых журналов и поиск в них по ключевым словам, ручной просмотр найденного и отсев непригодных работ, ручное извлечение сведений в табличный вид и, наконец, чистка и верификация собранного массива (рис. 12). Этот порядок сложился при создании первого набора и в дальнейшем воспроизводился без изменений в главном, различаясь лишь в том, что определяется предметом: в перечне просматриваемых журналов, в строках запроса и в составе извлекаемых сведений. Строки запроса и глубина поиска для каждого набора приведены в табл. 2, а прочие ступени изложены ниже.



Рисунок 12. Общий порядок сбора данных, единый для всех четырех наборов

Таблица 2. Особенности сбора данных для каждого из созданных наборов

Набор	Ключевые слова и место поиска	Глубина поиска
BigSolDB 2.0	«solubility» в заглавии	до апреля 2025 года
MixtureSolDB	«solubility + binary» и «solubility + mixture» в заглавии либо в кратком изложении	до октября 2025 года
IrLumDB	«iridium» и «complex» в заглавии	до октября 2024 года
IrCytoToxDB	«iridium» и «Ir(III)» в заглавии	по 2022 год включительно

Набор	Ключевые слова и место поиска	Глубина поиска
MetalCytoToxD B	первый ярус, заглавие: «Ruthenium», «Ru», «Ru(ii)», «Ru(iii)», «RuII», «cycloruthenated», «organoruthenium»; «Iridium», «Ir(iii)», «Organoiridium»; «Rhodium», «Rh(i)», «Rh(iii)», «organorhodium»; «Rhenium», «Re(i)», «organorhenium»; «Osmium», «Os(ii)», «Os(iv)», «organoosmium». Второй ярус, краткое изложение: «cancer», «toxic», «therapy», «biological», «tumor», «drug», «cellular», «apoptosis», «therapeutic», «antiproliferative», «tumour», «pharmaceutical», «clinical», «bioactive», «apoptotic», «IC ₅₀ »	до мая 2025 года

3.1.1. Отбор публикаций и критерии включения

Поиск велся при помощи общедоступной поисковой системы Cobalt[170]. Она охватывает свыше 130 миллионов публикаций и около двух миллиардов ссылок между ними и позволяет вести поиск по заглавиям, по полным текстам либо по авторам, ограничивая выдачу заданным перечнем журналов и годами издания. Следует оговорить, что эта система создана автором настоящей работы совместно с другими разработчиками вне рамок диссертации. Самостоятельным результатом работы она не является и рассматривается здесь исключительно как использованное средство.

Выбор ее определялся возможностью ограничить поиск кругом журналов. Универсальные поисковые системы ищут по всем изданиям сразу, вследствие чего выдача разрастается до величин, не поддающихся ручному просмотру: запрос «iridium» дает в них сотни тысяч работ. Отбор же по журналам, в которых нужны сведения появляются систематически, сводит выдачу к тысячам, то есть к числу, которое можно просмотреть. Поиск велся по заглавиям, а для двух наборов вторым ярусом также по тексту краткого изложения, как указано в табл. 2. Интерфейс системы с примером такого отбора показан на рис. 13.

Ограничение поиска заглавиями означает, что запрос должен быть нарочито широким: узкие сочетания слов отсекают большую часть пригодных работ, о чем свидетельствуют числа, приведенные в разделе 4.3.1. Естественным следствием такого построения является то, что подавляющая часть найденного отсеивается на следующей ступени вручную. Поиск ограничен работами на английском языке.

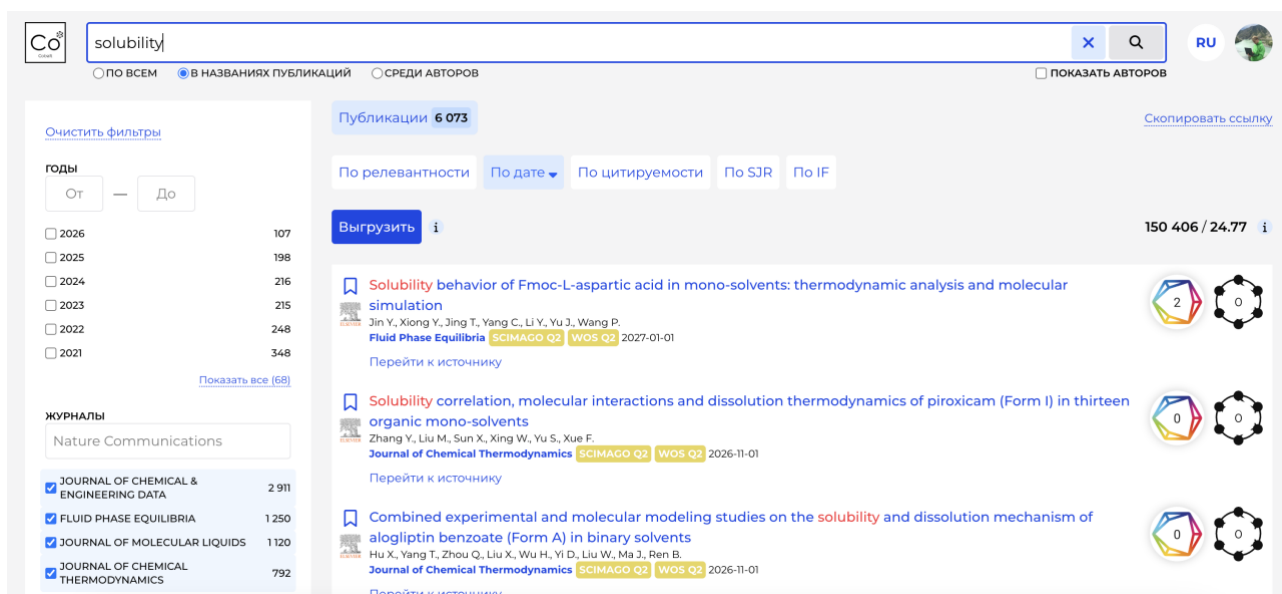


Рисунок 13. Интерфейс поисковой системы Cobalt на примере запроса «solubility» по заглавиям публикаций. Слева - отбор по годам и по журналам с указанием числа найденных работ в каждом из них

Круг просматриваемых журналов задавался предметом. Сведения о растворимости публикуются систематически в небольшом числе изданий по физической химии растворов, и перечень их короток: Journal of Molecular Liquids, Journal of Solution Chemistry, Journal of Chemical and Engineering Data, Journal of Chemical Thermodynamics, Chinese Journal of Chemical Engineering и Fluid Phase Equilibria, а для бинарных смесей также Physics and Chemistry of Liquids. Для комплексов переходных металлов такого узкого круга не существует, вследствие чего просматривались 36 журналов издательств Springer Nature, ACS, RSC, Wiley, MDPI и Elsevier по неорганической, металлоорганической и медицинской химии. Перечень их с числом найденных и отобранных статей приведен в табл. 10, а распределение публикаций-источников по журналам для расширенного набора - в табл. 12.

Отбор для наборов о цитотоксичности комплексов велся в два яруса. Сначала по вхождению в заглавие названий металлов и производных от них обозначений отбирались статьи по каждому металлу, а затем в полученном множестве отбирались работы, относящиеся к биологическим свойствам, по вхождению соответствующих слов в текст краткого изложения. Поиск при этом велся по подстроке, вследствие чего

«toxic» захватывало также «cytotoxic», «toxicity» и сходные образования. Ключевые слова обоих ярусов приведены в табл. 2.

Критерии включения задавались явно и различались по предметным областям. Для наборов о растворимости требовалось, чтобы величина была приведена в источнике непосредственно в мольных долях, а для смесей - чтобы состав был задан в мольных либо массовых долях. Отказ от работ, требующих пересчета из иных величин, устраняет целый класс ошибок, разобранный в разделе 2.1.2, ценой некоторого сокращения охвата. Включались органические соединения, органические соли и сольваты определенного состава. Принимались только измерения при атмосферном давлении и в отсутствие добавок. Полиморфные модификации при составлении BigSolDB 2.0 включались лишь тогда, когда в источнике прямо указано, что изучалась наиболее устойчивая из них, поскольку в противном случае измеренная величина относится к неопределенной твердой фазе, о чем говорится в разделе 2.1.3. В версии 2.2 это ограничение заменено разметкой: обозначение твердой формы, приведенное в первоисточнике, вынесено на уровень отдельного измерения, вследствие чего сопоставляемые в одной работе устойчивая и метастабильная формы сохраняются в наборе как различные записи и становятся предметом самостоятельного разбора.

Для наборов о комплексах переходных металлов включались только одноядерные соединения, что необходимо для последовательного описания комплекса перечислением его лигандов. Для IrLumDB охват был дополнительно ограничен октаэдрическими бис-циклометаллированными комплексами иридия(III). Основания этого ограничения изложены в разделе 4.2.1. Числа отобранных и отсеянных публикаций на каждой ступени приводятся для каждого набора в главе 4.

3.1.2. Извлечение данных из первичных источников

Извлечение велось вручную, чтением полного текста статьи и сопроводительных материалов, с занесением сведений в табличный вид. Автоматическое извлечение не применялось. Основания этому изложены в разделе 2.1.2: интересующие величины приводятся в первичных источниках то в тексте, то в таблицах, то на рисунках, единицы и обозначения не унифицированы, а условия эксперимента нередко рассеяны по разным частям работы, вследствие чего надежность

автоматического разбора оказывается недостаточной, а проверка его результатов по трудоемкости не уступает извлечению вручную.

Наряду с самой целевой величиной извлекались сведения об условиях, при которых она получена: без них запись не поддается однозначному прочтению. Для растворимости это температура и состав среды, то есть растворитель, а для смесей оба растворителя, доля первого из них и вид доли - мольная или массовая. Для люминесцентных свойств это растворитель, время жизни возбужденного состояния и условия измерения квантового выхода. Для цитотоксичности это клеточная линия, время воздействия и значение IC_{50} цисплатина, измеренное в той же работе и служащее средством сравнения между статьями, а для записей о фотодинамическом действии дополнительно длина волны, время и плотность мощность облучения. Для всех наборов без исключения фиксировался цифровой идентификатор публикации-источника, что позволяет вернуться к первоисточнику для любой записи и служит, кроме того, основанием для разбиения выборки по источнику, о котором говорится в разделе 3.3.2.

3.1.3. Стандартизация структур, идентификаторов и единиц

Структуры растворенных веществ и растворителей приводились к нотации SMILES. Основным источником служил ресурс PubChem. Для соединений, в нем отсутствующих, структура строилась средствами ChemDraw 18.0. Все полученные строки канонизировались средствами RDKit, что устраняет расхождения между записями одной и той же структуры. Обозначения растворителей приводились к единому виду для устранения неоднозначности: так, «1-propanol», «propanol» и «propan-1-ol» сводились к «n-propanol», а «2-butanol», «butan-2-ol» и «sec-butyl alcohol» - к «sec-butanol». Сведения о растворенном веществе - регистрационный номер CAS, идентификатор и название - извлекались через программный интерфейс PubChem, а признак одобрения Управлением по санитарному надзору за качеством пищевых продуктов и медикаментов США (FDA) - через программный интерфейс ChEMBL.

Растворимость в первичных источниках приводится в мольных долях, тогда как для построения моделей употребительны логарифмические шкалы, дающие более равномерное распределение. Пересчет выполнялся следующим образом.

Для индивидуальных растворителей мольная доля переводилась в молярную концентрацию по соотношению:

$$S = \frac{x}{(1-x)M_p} 1000\rho$$

где S - растворимость в моль/л, x - мольная доля растворенного вещества, ρ - плотность растворителя при температуре измерения, г/см³, а M_p - молярная масса растворителя, г/моль. Это соотношение дает число молей вещества, приходящееся на литр растворителя, и пренебрегает вкладом самого растворенного вещества в объем раствора. Это допущение правомерно, пока растворимость не слишком велика, и выполняется для подавляющей части набора. Плотность растворителя бралась из собранных экспериментальных значений при той же температуре, а при их отсутствии вычислялась по линейному приближению $\rho = aT + b$. Для 60 растворителей значения плотности извлекались из первичных публикаций, причем для каждого привлекалось не менее двух независимых работ. Для десяти наиболее употребительных, среди которых вода, метанол, толуол и предельные углеводороды, использовались справочные данные NIST Chemistry WebBook. Коэффициент детерминации каждого приближения превышал 0.995. Десятичный логарифм полученной величины и составляет LogS.

Для бинарных смесей растворителей от молярной шкалы пришлось отказаться, поскольку плотности смесей во всем диапазоне составов и температур недоступны. Растворимость пересчитывалась в массовую шкалу, в граммы вещества на 100 граммов растворителя. Вид соотношения зависит от того, в каких долях задан состав смеси в источнике. Если доля первого растворителя приведена как мольная,

$$S = \frac{x}{1-x} M_B / (wM_1 + (1-w)M_2) \cdot 100$$

а если как массовая,

$$S = x / (1-x) M_B (w/M_1 + (1-w)/M_2) \cdot 100$$

где S - растворимость в г/100 г, x - мольная доля растворенного вещества, w - доля первого растворителя в смеси до внесения вещества, M_B - молярная масса

растворенного вещества, а M_1 и M_2 - молярные массы первого и второго растворителей. Для полиэтиленгликолей молярная масса приписывалась по номинальному среднему значению той марки, которая указана в источнике. Наряду с массовой шкалой в наборе сохраняется и исходная молярная доля, а также десятичные логарифмы обеих величин.

Для цитотоксичности величины IC_{50} , приводимые в источниках в микромолях на литр, переводились в отрицательный десятичный логарифм молярной концентрации по соотношению $pIC_{50} = 6 - \lg IC_{50}$, где IC_{50} выражена в мкмоль/л. Этот переход принят в фармакологической литературе и делает распределение величин близким к нормальному, а различия в разных диапазонах концентраций - сопоставимыми.

3.1.4. Верификация и контроль качества

Ошибки, которым подвержен собираемый набор, тройкого рода. Первая - пропуск пригодной публикации; ей противодействует нарочитая широта запроса, о которой сказано в разделе 3.1.1. Вторая - ошибка при извлечении, стандартизации и пересчете величин. Третья принадлежит самому первоисточнику: рецензируемый характер публикации не исключает ни экспериментальных, ни систематических погрешностей, ни прямых опечаток в таблицах. Настоящий раздел посвящен средствам, направленным против второй и третьей.

Основным средством служила перекрестная проверка. Два исполнителя с подготовкой в области физической химии независимо друг от друга извлекали данные из одних и тех же публикаций, после чего третий сличал полученные наборы и сводил их воедино. Всякое расхождение между ними разрешалось обращением к первоисточнику. Этот прием существен потому, что случайная ошибка одного исполнителя обнаруживается несовпадением, тогда как при извлечении в одиночку она осталась бы незамеченной.

Сведенный набор проходил затем сплошную формальную проверку. Исправлялись молярные доли растворенного вещества, превышающие единицу, и доли растворителей, превышающие единицу, - величины физически невозможные и свидетельствующие об ошибке переноса. Пересматривались записи с температурой, несовместимой с состоянием раствора: для индивидуальных растворителей

устранялись значения выше температуры кипения, отвечающие измерениям в субкритических средах и потому несопоставимые с прочими, а для смесей пересматривались записи, отстоящие от температуры кипения любого из компонентов менее чем на десять градусов, для чего температуры кипения растворителей собирались отдельным файлом. Все идентификаторы публикаций-источников проверялись через Crossref API, чем устраняются как опечатки, так и ссылки на несуществующие работы. Отдельно пересматривались случаи, когда для одной пары вещество - растворитель разные работы дают значения, различающиеся на порядок и более: такое расхождение указывает либо на ошибку извлечения, либо на различие условий, не отраженное в записи, и разрешалось оно возвращением к обоим статьям. Наконец, устранялись случаи, когда одному названию отвечали две различные структуры или одной структуре два названия, - обыкновенный источник неоднозначности при сведении данных из разных источников.

Особого решения потребовало устранение дубликатов. Удалялись лишь записи, совпадающие одновременно по веществу, растворителю, температуре и самому значению и восходящие, судя по всему, к одному измерению. При полном совпадении сохранялся более ранний источник. Независимые же повторные измерения разных групп в одинаковых условиях сохранялись намеренно. Это решение принято вопреки распространенной практике усреднения, и оправдание его двоякое: усреднение уничтожает сведения о воспроизводимости, а различить усредненную величину и единичное измерение впоследствии невозможно. Сохраненные повторы позволили оценить внутреннюю согласованность собранных данных, чему посвящен раздел 4.1.4.

Для наборов о комплексах переходных металлов к перечисленному добавлялась сплошная ручная сверка структурных формул, единиц и числовых значений с первоисточником, поскольку строение комплекса задается перечислением лигандов и ошибка в одном из них не обнаруживается никакой формальной проверкой.

3.1.5. Структура наборов и депонирование

Каждый набор представляет собой таблицу в формате с разделителями, в которой строка отвечает одному экспериментальному измерению, а не соединению. Состав столбцов каждого набора приведен в табл. 3, 4, 9 и 11.

Устройство столбцов подчинено одному правилу: записанное в источнике и вычисленное при обработке разделяются. Величина сохраняется в том виде, в каком она приведена в первоисточнике, а полученные из нее пересчетом значения помещаются в отдельные столбцы. Это правило существенно для наборов о цитотоксичности, где часть значений приведена в источниках не числом, а предельной оценкой вида «более 100 мкмоль/л»: разделение позволяет сохранить такие записи, не смешивая их с измеренными числами. Там, где обработка вносила допущения, они помечены отдельным признаком. Так, в MixtureSolDB признаком отмечены записи, отвечающие предельным составам смеси.

Наборы снабжены вспомогательными файлами, содержащими те справочные величины, которые потребовались при обработке и собирались наравне с основными данными: для BigSolDB 2.0 это плотности растворителей при различных температурах, для MixtureSolDB - температуры кипения растворителей. Каждое значение в них сопровождается указанием источника, вследствие чего пересчет может быть воспроизведен или пересмотрен независимо.

Депонирование наборов, присвоение им постоянных идентификаторов, лицензирование и версионирование изложены в разделе 4.4, где рассматривается и соответствие принципам FAIR, сформулированным в разделе 2.1.4.

3.2. Представление соединений

3.2.1. Органические соединения и растворители

Растворенное вещество и растворитель задаются структурными формулами в виде строк SMILES, приведенных к каноническому виду средствами RDKit. Каноническая запись служит и ключом при сведении данных из разных публикаций, и признаком, по которому выявляются дубликаты. Наряду с ней для каждого соединения сохраняются регистрационный номер CAS и идентификатор PubChem CID, а также тривиальное название в том виде, в каком оно приведено в источнике. Эта избыточность намеренна: строка SMILES обеспечивает машинную обработку, номера позволяют сверить отнесение с внешними ресурсами, а название - с первоисточником.

Единицей записи служит система: растворенное вещество, среда и температура. Температура выносится в отдельный столбец наравне с составом среды и величиной растворимости, вследствие чего каждой паре вещество - растворитель отвечает, как правило, серия измерений. Следствия такого устройства данных разбираются в разделе 4.1.3. Среда задается для чистых растворителей строкой SMILES и названием, а для бинарных смесей - строками SMILES обоих растворителей, долей первого из них и указанием на то, мольная эта доля или массовая. Последнее необходимо, поскольку в первичных публикациях употребляются обе шкалы, и без явного указания невозможно определить, какое из приведенных выше соотношений применимо для пересчета. Записи, отвечающие предельным составам смеси, то есть чистым растворителям, помечены отдельным признаком.

3.2.2. Комплексы иридия(III) по структурным формулам лигандов

Строка SMILES к координационным соединениям в общем случае неприменима по причинам, разобранным в разделе 2.3.1. Взаимное расположение лигандов вокруг иона иридия(III) во всех рассматриваемых соединениях, однако, однотипно, что позволяет описывать комплекс через его лиганды, не задавая координационную связь в явном виде. Этот подход заимствован из работы по прогнозированию антибактериальной активности комплексов рутения[167].

Воплощен он следующим образом. Структурная формула каждого лиганда записывалась строкой SMILES, приводилась к каноническому виду средствами RDKit[171] и преобразовывалась в расширенный отпечаток связности ECFP4[172] длиной 2048 бит средствами пакета molfeat[173]. Отпечаток комплекса получался как линейная комбинация отпечатков входящих в него лигандов.

Для задачи о цитотоксичности к отпечаткам добавлялись вычисляемые признаки. Средствами RDKit определялись коэффициент распределения по Криппену[174], площадь топологической полярной поверхности[175] (TPSA) и молекулярная масса, причем не для комплекса как целого и не для каждого лиганда порознь, а для лигандной системы, взятой как одна несвязная молекула вида L1.L2.L3. Основания такого выбора изложены в разделе 4.3.2. К ним присоединялись суммарный заряд комплекса, время воздействия, приведенное к единой шкале преобразованием

StandardScaler из Scikit-learn[176], и клеточная линия в унитарном кодировании. Для многометалльной модели вектор признаков дополнялся тремя величинами, относящимися к металлическому центру: порядковым номером, степенью окисления и числом валентных электронов.

Отбор данных для обучения проводился так, чтобы выборка оставалась однородной по условиям измерения. Для люминесцентных свойств брались только измерения в дихлорметане, поскольку положение максимума испускания зависит от полярности среды. Исключались лиганды, содержащие бор, бром и кремний, ввиду редкости этих элементов в наборе. Для квантового выхода дополнительно требовалось прямое указание источника на удаление кислорода. Для цитотоксичности из обучения исключались неопухолевые клеточные линии и оставлялись только одноядерные комплексы.

3.3. Модели машинного обучения

3.3.1. Алгоритмы и подбор гиперпараметров

Испытывались пять алгоритмов: метод ближайших соседей[177], метод опорных векторов[178] и три реализации градиентного бустинга над решающими деревьями - LightGBM[179], CatBoost[180] и XGBoost[181]. Выбор определялся тем, что признаковое описание здесь разреженное и высокоразмерное, а объем выборки составляет тысячи записей, то есть находится в той области, где ансамбли деревьев устойчиво превосходят как простые метрические методы, так и нейросетевые модели.

Гиперпараметры подбирались средствами Optuna[182]. Подбирались скорость обучения, число и глубина деревьев, доля объектов и доля признаков, привлекаемых при построении каждого дерева, и наименьшее число объектов в листе, для метода ближайших соседей - число соседей, для метода опорных векторов - параметры C и ϵ . Наилучшие найденные значения для каждого алгоритма и каждой задачи приведены в сопроводительных материалах к статьям 2 и 5 из списка опубликованных работ по теме диссертации (см. раздел 7 «Список публикаций, в которых изложены основные результаты диссертации»). Подбор велся на отдельной отложенной части выборки, составлявшей пятую долю для длины волны максимума испускания и десятую для квантового выхода, а не на разбиениях перекрестной проверки. Вложенная

перекрестная проверка при этом не применялась, а отложенная часть, по которой выбирались гиперпараметры, входила затем в набор, на котором проводилась перекрестная проверка. Поэтому сообщаемые оценки качества могут иметь оптимистическое смещение, величина которого отдельно не оценивалась. Проспективной проверки на синтезированных соединениях это не касается, поскольку они не участвовали ни в подборе, ни в обучении. Для моделей цитотоксичности подбор велся тем же порядком на отложенной пятой доле выборки, отделенной по структурам, тогда как итоговое качество оценивалось перекрестной проверкой с разбиением по источнику, описанным в разделе 3.3.2. Преобразования признаков - приведение молекулярной массы и времени воздействия к единой шкале и унитарное кодирование клеточной линии - целевой величины не используют и потому сведений о ней между частями выборки не переносят.

Задачи ставились сообразно устройству величин. Длина волны максимума испускания и квантовый выход прогнозировались как задачи регрессии. Для квантового выхода наряду с этим решалась задача классификации на три класса, поскольку распределение величины сильно смещено к малым значениям и для приложений существенна не столько сама величина, сколько отнесение соединения к слабо, умеренно или сильно излучающим. Цитотоксичность ставилась как задача бинарной классификации с порогом 10 мкмоль/л, что вызвано присутствием в источниках предельных оценок вида «более 100 мкмоль/л». Основания выбора порога изложены в разделе 4.3.2. Значение pIC_{50} при этом сохранялось как целевая величина для вспомогательных моделей регрессии.

Для интерпретации построенных моделей применялся метод SHAP[183], оценивающий вклад каждого признака в отдельный прогноз.

3.3.2. Схемы разбиения выборки и валидация

Качество на обучающей выборке оценивалось десятикратной перекрестной проверкой средствами Scikit-learn[176]. Способ разбиения выборки при этом различался. Для задачи о цитотоксичности сопоставлялись два способа: по структуре, когда соединения распределялись между частями по каноническим строкам SMILES, и по источнику, когда все записи одной публикации целиком попадали либо в

обучающую, либо в проверочную часть. Вторым способ принят как основной, поскольку первый допускает утечку между близкими по строению соединениями из одной статьи. Количественное сопоставление обоих приведено в разделе 4.3.2. Для задач о люминесцентных свойствах разбиение выполнялось по соединениям: выборка предварительно приводилась к одной записи на комплекс отбором измерений в дихлорметане и удалением повторов, совпадающих по всем трем лигандам, вследствие чего повторные измерения одного и того же соединения не могли оказаться одновременно в обучающей и проверочной частях.

Помимо перекрестной проверки применялись три способа внешней проверки. Проверка по времени выхода публикаций: модель обучалась на работах, вышедших до заданного года, и испытывалась на более поздних, что воспроизводит условия действительного применения. Год отсечения при этом варьировался, чтобы судить об устойчивости оценок к объему обучающей выборки. Проверка исключением одного металла: модель обучалась на данных четырех металлов и испытывалась на пятом, чем оценивалась переносимость сведений между подмножествами. Проспективная проверка на специально синтезированных соединениях, не входивших ни в обучающую, ни в проверочную часть и полученных после построения модели.

Достоверность обученных моделей проверялась, кроме того, перемешиванием меток: соответствие между соединениями и классами разрушалось случайной перестановкой, после чего обучение и перекрестная проверка повторялись без изменений. Эта проверка показывает, опирается ли достигнутое качество на действительную связь признаков с меткой; систематических смещений самой выборки она не выявляет.

Качество оценивалось общепринятыми величинами: для регрессии средней абсолютной ошибкой, среднеквадратичной ошибкой и коэффициентом детерминации, для классификации точностью, площадью под характеристической кривой обнаружения и матрицей ошибок. Матрица ошибок приводится по отдельным клеточным линиям, поскольку сводная величина скрывает различия между ними. Все величины сопровождаются стандартным отклонением по десяти разбиениям перекрестной проверки. При проверке по времени выхода публикаций употребляются,

кроме того, две величины, принятые при оценке виртуального скрининга. Доля попаданий определяется как отношение числа действительно цитотоксичных соединений к числу отобранных среди десятой части выборки с наибольшей прогнозируемой вероятностью, причем перед отбором отсеиваются соединения, близкие к обучающим по мере сходства Танимото при пороге 0.9. Коэффициент обогащения есть отношение доли попаданий к доле цитотоксичных соединений во всей проверочной выборке, то есть показывает, во сколько раз отбор по модели превосходит случайный.

3.3.3. Программные средства и вычислительные ресурсы

Весь программный код написан на языке Python. Обработка структур и вычисление дескрипторов выполнены средствами RDKit[171] и molfeat[173], обучение моделей - средствами[179] LightGBM, CatBoost[180], XGBoost[181] и Scikit-learn[176], подбор гиперпараметров - средствами Optuna[182], интерпретация моделей - средствами библиотеки SHAP. Обработка табличных данных велась средствами pandas и NumPy[184], построение рисунков - средствами matplotlib[185]. Расчеты не требовали графических ускорителей и выполнялись на персональном компьютере. Исходный код обработки данных и построения моделей размещен в репозиториях, указанных в табл. 15. Там же в файлах requirements.txt закреплены версии всех использованных библиотек, что обеспечивает воспроизводимость расчетов.

3.4. Синтез и физико-химические измерения

3.4.1. Исходные вещества

Гидрат хлорида иридия(III) $\text{IrCl}_3 \cdot 3\text{H}_2\text{O}$ (Aldrich, 99.9 %), безводный карбонат калия K_2CO_3 (хч, Химмед), ацетилацетон (чда, Химмед), трифторметансульфонат серебра AgOTf (Aldrich), 2,2'-бипиридин (98 %, Acros), 4,4'-диметил-2,2'-бипиридин (99 %, Aldrich), 4,4'-бис(метоксикарбонил)-2,2'-бипиридин, тетра-н-бутиламмоний перхлорат $(\text{n-Bu}_4\text{N})\text{ClO}_4$ (98 %, Acros), аргон газообразный (99.999 %, НИИ КМ), силикагель 60 (Merck).

2-этоксиэтанол (99 %, Aldrich), ацетонитрил (99.98 %, для ВЭЖХ, MACRON), этанол (96 %), хлористый метилен (хч, Химмед), хлороформ (ч, Химмед), диэтиловый эфир (чда, Химмед), этилацетат (хч, Химмед), гексан (хч, Химмед), метанол (99.99 %, Aldrich),

Fluka), деионизованная вода. Растворители перегоняли и высушивали при необходимости по стандартным методикам.

Циклометаллирующие лиганды 1-бензил-2-фенил-1Н-бензо[d]имидазол (bim), 1-бензил-2-(4-трет-бутилфенил)-1Н-бензо[d]имидазол (tbim), 1-бензил-2-(4-хлорфенил)-1Н-бензо[d]имидазол (cbim) и 1-бензил-2-(3,4-диметоксифенил)-1Н-бензо[d]имидазол (mbim) получены по описанной ранее методике. Ранее не описанные лиганды 1-бензил-5,6-дихлор-2-фенил-1Н-бензо[d]имидазол (dcbim), 1-бензил-2-(4-трет-бутилфенил)-5,6-дихлор-1Н-бензо[d]имидазол (tdcbim), 1-бензил-5,6-дихлор-2-(4-хлорфенил)-1Н-бензо[d]имидазол (tcbim), 1-бензил-5,6-диметил-2-фенил-1Н-бензо[d]имидазол (dmbim) и 1-бензил-2-(п-толил)-1Н-нафто[2,3-d]имидазол (mnim) синтезированы аналогичным образом с выходами от 59 до 82 %. Циклометаллированный хлорид $[\text{Ir}(\text{phi})_2\text{Cl}]$, где phi - 1,2-дифенил-1Н-фенантро[9,10-d]имидазол, получен по описанной ранее методике.

3.4.2. Методы исследования

Спектры ЯМР ^1H и $^{13}\text{C}\{^1\text{H}\}$ регистрировали при 25 °С на спектрометре Bruker Avance 400, химические сдвиги приведены в миллионных долях относительно остаточных сигналов растворителя. Масс-спектры высокого разрешения измеряли на масс-спектрометре Bruker microTOF-Q ESI-TOF с ионизацией электрораспылением и времяпролетным анализатором. Массы и изотопное распределение молекулярных ионов сопоставляли с рассчитанными.

Электронные спектры поглощения измеряли на спектрофотометре ОКБ Спектр СФ-2000 в кварцевых кюветах при комнатной температуре. Спектры фотolumинесценции регистрировали на спектрофлуориметре Perkin-Elmer LS-55 в растворах в хлористом метиле, насыщенных аргоном. Удаление кислорода необходимо, поскольку он тушит фосфоресценцию комплексов иридия(III), о чем говорится в разделе 2.4.1.

Циклические вольтамперограммы измеряли на потенциостате Elins P20 в трехэлектродной ячейке с платиновым рабочим электродом, платиновым вспомогательным электродом и серебряным квазиреференсным электродом. Вещества

растворяли в насыщенном аргонном хлористом метиле, содержащем 0.25 М (н-Вu₄N)ClO₄. Развертка потенциала 200 мВ/с, ферроцен использовали в качестве внутреннего стандарта.

Сбор дифракционных данных для монокристаллов проводили при использовании излучения CuK α (λ = 1.54178 Å) либо MoK α (λ = 0.71073 Å) на дифрактометре Bruker D8 VENTURE. Учет поглощения вводили по измерениям интенсивностей эквивалентных отражений (SADABS, Bruker). Структуры расшифровывали прямым методом и уточняли полноматричным анизотропным методом наименьших квадратов по F² для всех неводородных атомов в пакете программ Olex2. Атомы водорода помещали в рассчитанные позиции и уточняли по модели «наездника». Строение установлено для девяти из двенадцати полученных комплексов. Кристаллографические сведения депонированы в Кембриджскую структурную базу данных под номерами 2393803-2393811. Фазовую чистоту порошкообразных образцов подтверждали методом порошковой рентгеновской дифракции на дифрактометре Bruker D8 Advance (CuK α , λ = 1.5418 Å, Ni-фильтр, геометрия на отражение). Рентгенодифракционные исследования и измерения люминесценции выполнены в Центре коллективного пользования и Центре цвета ИОНХ РАН.

3.4.3. Синтез комплексов иридия(III)

Все операции по получению комплексов иридия(III) проводили в атмосфере сухого аргона. Очистку и последующие манипуляции с полученными соединениями выполняли на воздухе. Всего в настоящей работе получено двенадцать ранее не описанных комплексов: девять с ацетилацетонат-ионом и три с производными 2,2'-бипиридина в качестве вспомогательного лиганда.

Ранее не описанные циклометаллирующие лиганды охарактеризованы следующим образом.

1-Бензил-5,6-дихлор-2-фенил-1H-бензо[d]имидазол (dcbim). Белые кристаллы, выход 64 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 5.41 (с, 2H), 7.03-7.10 (м, 2H), 7.29 (с, 1H), 7.32-7.40 (м, 3H), 7.43-7.55 (м, 3H), 7.63-7.70 (м, 2H), 7.93 (с, 1H).

Спектр ЯМР $^{13}\text{C}\{^1\text{H}\}$, δ , м.д. (CDCl_3): 48.21, 111.46, 120.66, 125.40, 126.45, 126.60, 127.76, 128.55, 128.68, 128.77, 128.89, 130.10, 134.82, 134.99, 141.92, 155.62.

1-Бензил-2-(4-трет-бутилфенил)-5,6-дихлор-1H-бензо[d]имидазол (tdcbim).
Белый порошок, выход 59 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 1.34 (с, 9H), 5.45 (с, 2H), 7.06-7.13 (м, 2H), 7.28 (с, 1H), 7.31-7.41 (м, 3H), 7.46-7.53 (м, 2H), 7.60-7.67 (м, 2H), 7.95 (с, 1H).

Спектр ЯМР $^{13}\text{C}\{^1\text{H}\}$, δ , м.д. (CDCl_3): 30.74, 34.54, 48.29, 111.44, 120.42, 125.36, 125.60, 126.60, 127.76, 128.52, 128.91, 134.70, 134.98, 153.71, 155.57.

1-Бензил-5,6-дихлор-2-(4-хлорфенил)-1H-бензо[d]имидазол (tcbim).
Коричневый порошок, выход 71 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 5.39 (с, 2H), 7.02-7.09 (м, 2H), 7.30 (с, 1H), 7.32-7.41 (м, 3H), 7.41-7.47 (м, 2H), 7.56-7.64 (м, 2H), 7.92 (с, 1H).

Спектр ЯМР $^{13}\text{C}\{^1\text{H}\}$, δ , м.д. (CDCl_3): 48.21, 111.43, 120.72, 125.28, 126.67, 126.92, 127.12, 127.90, 128.87, 128.99, 130.04, 134.78, 134.90, 136.46, 141.79, 154.39.

1-Бензил-5,6-диметил-2-фенил-1H-бензо[d]имидазол (dmbim). Белый порошок, выход 82 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 2.34 (с, 3H), 2.40 (с, 3H), 5.41 (с, 2H), 6.99 (с, 1H), 7.08-7.15 (м, 2H), 7.27-7.38 (м, 3H), 7.40-7.50 (м, 3H), 7.65 (с, 1H), 7.66-7.72 (м, 2H).

Спектр ЯМР $^{13}\text{C}\{^1\text{H}\}$, δ , м.д. (CDCl_3): 19.95, 20.20, 47.84, 110.18, 119.62, 125.49, 127.24, 128.27, 128.64, 128.76, 129.24, 129.92, 131.15, 131.84, 134.30, 136.32, 141.41, 152.96.

1-Бензил-2-(п-толил)-1H-нафто[2,3-d]имидазол (mnim). Бежевый порошок, выход 79 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 2.43 (с, 3H), 5.51 (с, 2H), 7.14-7.21 (м, 2H), 7.27-7.44 (м, 7H), 7.59 (с, 1H), 7.64-7.71 (м, 2H), 7.79-7.87 (м, 1H), 8.00-8.08 (м, 1H), 8.36 (с, 1H).

Спектр ЯМР $^{13}\text{C}\{^1\text{H}\}$, δ , м.д. (CDCl_3): 21.12, 48.25, 105.88, 116.26, 123.22, 123.97, 125.54, 126.30, 127.09, 127.38, 128.10, 128.73, 128.85, 129.19, 130.18, 130.23, 135.97, 136.42, 140.37, 142.55, 157.75.

Димерные циклометаллированные хлориды получали по общей методике. Гидрат хлорида иридия(III) (50 мг, 0.142 ммоль) и соответствующий бензимидазольный лиганд (0.355 ммоль) суспендировали в смеси 2-этоксиэтанола (10 мл) и воды (3 мл). Смесь кипятили с обратным холодильником в течение 12 часов в атмосфере аргона. Образовавшийся осадок отфильтровывали, промывали этанолом и диэтиловым эфиром и сушили в вакууме при 50 °С. Выход неочищенного продукта составлял от 50 до 84 %. Малая растворимость этих соединений не позволила получить пригодные для интерпретации спектры ЯМР, вследствие чего димерные хлориды вводили в дальнейшие превращения без дополнительной характеристики.

Ацетилацетонатные комплексы получали по следующей общей методике. В восьмимиллилитровую виалу помещали 30 мг соответствующего димерного хлорида, 18 мг карбоната калия и 8 мкл ацетилацетона и суспендировали в 4 мл ацетонитрила. Суспензию продували аргоном в течение пяти минут, виалу герметично закрывали и выдерживали при 65 °С в течение 24 ч при перемешивании. Образовавшиеся желто-оранжевые твердые продукты отфильтровывали, тщательно промывали деионизованной водой и сушили в вакууме. Выходы составили от 18 до 72 %. Таким образом получены комплексы $[\text{Ir}(\text{bim})_2(\text{acac})]$, $[\text{Ir}(\text{tbim})_2(\text{acac})]$, $[\text{Ir}(\text{cbim})_2(\text{acac})]$, $[\text{Ir}(\text{mbim})_2(\text{acac})]$, $[\text{Ir}(\text{dcbim})_2(\text{acac})]$, $[\text{Ir}(\text{tdcbim})_2(\text{acac})]$, $[\text{Ir}(\text{tcbim})_2(\text{acac})]$, $[\text{Ir}(\text{dmbim})_2(\text{acac})]$ и $[\text{Ir}(\text{mnim})_2(\text{acac})]$.

$[\text{Ir}(\text{bim})_2(\text{acac})]$. Выделенный продукт не требовал дополнительной очистки. Желтый порошок, выход 46 %.

Спектр ЯМР ^1H , δ , м.д. (DMSO-d_6): 7.82 (д, $J = 8.1$ Гц, 2H), 7.60-7.50 (м, 4H), 7.40-7.30 (м, 8H), 7.27-7.21 (м, 2H), 7.17-7.12 (м, 4H), 6.63 (тд, $J = 7.5, 1.3$ Гц, 2H), 6.47 (тд, $J = 7.5, 1.2$ Гц, 2H), 6.22-6.09 (м, 6H), 5.21 (с, 1H), 1.71 (с, 6H).

HRMS (ESI) m/z : $[\text{M}]^+$ расч. для $\text{C}_{45}\text{H}_{37}\text{IrN}_4\text{O}_2$ 858.2543; получено 858.2545.

$[\text{Ir}(\text{tbim})_2(\text{acac})]$. Выделенный продукт не требовал дополнительной очистки. Желтый порошок, выход 55 %.

Спектр ЯМР ^1H , δ , м.д. (DMSO-d_6): 7.79 (д, $J = 7.9$ Гц, 2H), 7.58-7.54 (м, 2H), 7.44 (д, $J = 8.4$ Гц, 2H), 7.40-7.27 (м, 10H), 7.21-7.17 (м, 4H), 6.67 (дд, $J = 8.3, 2.1$ Гц, 2H), 6.31 (д, $J = 2.0$ Гц, 2H), 6.17 (д, $J = 17.7$ Гц, 2H), 6.05 (д, $J = 17.7$ Гц, 2H), 5.18 (с, 1H), 1.71 (с, 6H), 0.80 (с, 18H).

HRMS (ESI) m/z : $[\text{M}]^+$ расч. для $\text{C}_{53}\text{H}_{53}\text{IrN}_4\text{O}_2$ 970.3797; получено 970.3813. Сигнал иона $[\text{M}]^+$ перекрывается с сигналом иона $[\text{M}+\text{H}]^+$.

$[\text{Ir}(\text{cbim})_2(\text{acac})]$. Очищали колоночной хроматографией на оксиде алюминия, элюент - хлористый метилен с последующим переходом к смеси хлористый метилен - метанол 20:1 по объему. Желтый порошок, выход 18 %.

Спектр ЯМР ^1H , δ , м.д. (CD_2Cl_2): 7.65 (д, $J = 8.3$ Гц, 2H), 7.50 (д, $J = 8.0$ Гц, 2H), 7.43-7.31 (м, 12H), 7.20 (д, $J = 7.6$ Гц, 4H), 6.75 (д, $J = 8.3$ Гц, 2H), 6.34 (с, 2H), 6.01-5.93 (м, 4H), 1.86 (с, 6H). Сигнал ацетилацетонатного фрагмента около 5.35 м.д. перекрывается с остаточным сигналом растворителя.

HRMS (ESI) m/z : $[\text{M}]^+$ расч. для $\text{C}_{45}\text{H}_{35}\text{Cl}_2\text{IrN}_4\text{O}_2$ 926.1749; получено 926.1742.

$[\text{Ir}(\text{mbim})_2(\text{acac})]$. Реакционная смесь содержала небольшое количество осадка. Кристаллический продукт выделялся при медленном упаривании ацетонитрила, его отфильтровывали, тщательно промывали деионизованной водой и сушили в вакууме. Желтый порошок, выход 45 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 7.74-7.70 (м, 2H), 7.40-7.37 (м, 2H), 7.32-7.28 (м, 10H), 7.26-7.22 (м, 4H), 6.97 (с, 2H), 5.98-5.90 (м, 2H), 5.89 (с, 2H), 5.86-5.78 (м, 2H), 5.23 (с, 1H), 3.47 (с, 6H), 3.22 (с, 6H), 1.84 (с, 6H).

HRMS (ESI) m/z : $[M]^+$ расч. для $C_{49}H_{45}IrN_4O_6$ 978.2969; получено 978.2965.

[Ir(dcbim)₂(acac)]. Выделенный продукт не требовал дополнительной очистки. Желтый порошок, выход 59 %.

Спектр ЯМР 1H , δ , м.д. ($CDCl_3$): 7.82 (с, 2H), 7.48-7.46 (м, 4H), 7.39-7.34 (м, 6H), 7.24 (с, 2H), 7.22 (с, 2H), 6.72 (т, $J = 7.7$ Гц, 2H), 6.64 (т, $J = 7.7$ Гц, 2H), 6.46 (д, $J = 7.9$ Гц, 2H), 5.93 (д, $J = 17.5$ Гц, 2H), 5.84 (д, $J = 17.5$ Гц, 2H), 5.25 (с, 1H), 1.85 (с, 6H).

HRMS (ESI) m/z : $[M]^+$ расч. для $C_{45}H_{33}Cl_4IrN_4O_2$ 996.0951; получено 996.0943.

[Ir(tdcbim)₂(acac)]. Очищали колоночной хроматографией на оксиде алюминия, элюент - смесь хлористый метилен - гексан 1:3 по объему. Желтый порошок, выход 37 %.

Спектр ЯМР 1H , δ , м.д. ($CDCl_3$): 7.87 (с, 2H), 7.45 (с, 2H), 7.36-7.32 (м, 8H), 7.25-7.22 (м, 4H), 6.75 (дд, $J = 8.3, 2.0$ Гц, 2H), 6.47 (д, $J = 2.0$ Гц, 2H), 5.97-5.90 (м, 2H), 5.79-5.73 (м, 2H), 5.24 (с, 1H), 1.86 (с, 6H), 0.94 (с, 18H).

HRMS (ESI) m/z : $[M]^+$ расч. для $C_{53}H_{49}Cl_4IrN_4O_2$ 1108.2206; получено 1108.2227.

[Ir(tcbim)₂(acac)]. Выделенный продукт не требовал дополнительной очистки. Желтый порошок, выход 57 %.

Спектр ЯМР 1H , δ , м.д. ($DMSO-d_6$): 8.44 (с, 2H), 7.66 (д, $J = 8.6$ Гц, 2H), 7.58 (с, 2H), 7.33 (т, $J = 7.4$ Гц, 6H), 7.26 (д, $J = 7.4$ Гц, 2H), 7.08 (с, 2H), 7.06 (с, 2H), 6.80 (дд, $J = 8.4, 2.2$ Гц, 2H), 6.21 (с, 4H), 6.06 (д, $J = 2.2$ Гц, 2H), 5.38 (с, 1H), 1.77 (с, 6H).

HRMS (ESI) m/z : $[M]^+$ расч. для $C_{45}H_{30}Cl_6IrN_4O_2$ 1063.0085; получено 1063.0147. Сигнал иона $[M]^+$ перекрывается с сигналом иона $[M+H]^+$.

[Ir(dmbim)₂(acac)]. Очищали перекристаллизацией из смеси хлористый метилен - гексан 1:1 по объему. Желтый порошок, выход 69 %.

Спектр ЯМР 1H , δ , м.д. ($CDCl_3$): 7.53 (с, 2H), 7.40-7.28 (м, 10H), 7.24 (с, 2H), 7.14 (с, 2H), 6.66-6.61 (м, 2H), 6.57-6.51 (м, 2H), 6.49-6.45 (м, 2H), 5.92 (д, $J = 17.4$ Гц, 2H), 5.84 (д, $J = 17.4$ Гц, 2H), 5.21 (с, 1H), 2.37 (с, 6H), 2.34 (с, 6H), 1.83 (с, 6H).

HRMS (ESI) m/z : $[M]^+$ расч. для $C_{49}H_{45}IrN_4O_2$ 914.3172; получено 914.3184. Сигнал иона $[M]^+$ перекрывается с сигналом иона $[M+H]^+$.

$[Ir(mnim)_2(acac)]$. Очищали перекристаллизацией из смеси хлористый метилен - ацетонитрил 1:1 по объему. Оранжевый порошок, выход 72 %.

Спектр ЯМР 1H , δ , м.д. (CD_2Cl_2): 8.17 (с, 2H), 7.97 (д, $J = 8.0$ Гц, 2H), 7.93 (д, $J = 7.7$ Гц, 2H), 7.86 (с, 2H), 7.54-7.33 (м, 16H), 6.61 (д, $J = 8.0$ Гц, 2H), 6.42 (с, 2H), 6.13 (д, $J = 17.5$ Гц, 2H), 6.05 (д, $J = 17.6$ Гц, 2H), 1.99 (с, 6H), 1.96 (с, 6H). Сигнал ацетилацетонатного фрагмента около 5.35 м.д. перекрывается с остаточным сигналом растворителя.

HRMS (ESI) m/z : $[M]^+$ расч. для $C_{55}H_{45}IrN_4O_2$ 986.3171; получено 986.3170.

Малая растворимость полученных ацетилацетонатных комплексов не позволила получить пригодные для интерпретации спектры ЯМР ^{13}C .

Комплексы с дииминовыми вспомогательными лигандами получали по следующей общей методике. Циклометаллированный хлорид иридия(III) (50 мг, 0.052 ммоль) и соответствующее производное 2,2'-бипиридина (0.104 ммоль) смешивали в сухом хлороформе (5 мл) и этаноле (5 мл), после чего прибавляли 130 мкл 0.5 М раствора трифторметансульфоната серебра в метаноле и кипятили с обратным холодильником в течение 8-10 часов в атмосфере аргона в темноте. Полученный раствор отфильтровывали, упаривали досуха, растворяли в хлористом метиле, вновь отфильтровывали и перекристаллизовывали из смеси хлористый метилен - гексан в объемном отношении 1:1. Таким образом получены комплексы $[Ir(phi)_2(bpy)]$, $[Ir(phi)_2(dmbpy)]$ и $[Ir(phi)_2(dcmby)]$, где bpy - 2,2'-бипиридин, $dmbpy$ - 4,4'-диметил-2,2'-бипиридин, $dcmby$ - 4,4'-бис(метоксикарбонил)-2,2'-бипиридин, с выходами 59, 77 и 58 %, соответственно. Комплекс $[Ir(phi)_2(bpy)]$ дополнительно очищали колоночной хроматографией на силикагеле с последовательным изменением состава элюента от смеси этилацетат - гексан 20:1 до смеси этилацетат - метанол 20:1 по объему.

$[Ir(phi)_2(bpy)]$. Желто-оранжевый порошок, выход 59 %.

Спектр ЯМР ^1H , δ , м.д. (ацетон- d_6): 8.80-8.72 (м, 4H), 8.63 (д, $J = 8.4$ Гц, 2H), 8.26 (дд, $J = 8.1, 1.0$ Гц, 2H), 8.15 (дд, $J = 7.8, 1.4$ Гц, 2H), 8.10-8.02 (м, 9H), 8.01-7.98 (м, 3H), 7.94-7.86 (м, 2H), 7.60-7.52 (м, 2H), 7.40 (дд, $J = 7.5, 0.9$ Гц, 2H), 7.37-7.28 (м, 4H), 7.16-7.10 (м, 2H), 6.92-6.75 (м, 6H), 6.63 (дд, $J = 7.8, 1.2$ Гц, 2H).

HRMS (ESI) m/z : $[\text{M}]^+$ расч. для $\text{C}_{64}\text{H}_{42}\text{IrN}_6$ 1087.3100; получено 1087.3105.

$[\text{Ir}(\text{phi})_2(\text{dmbpy})]$. Желтый порошок, выход 77 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 8.53 (д, $J = 8.4$ Гц, 2H), 8.41 (д, $J = 8.3$ Гц, 2H), 8.29 (д, $J = 5.7$ Гц, 2H), 8.01 (дд, $J = 8.2, 0.9$ Гц, 2H), 7.94-7.83 (м, 2H), 7.82-7.77 (м, 2H), 7.65-7.58 (м, 2H), 7.50-7.43 (м, 4H), 7.36 (д, $J = 5.7$ Гц, 2H), 7.25-7.15 (м, 6H), 7.03-6.97 (м, 2H), 6.79-6.69 (м, 4H), 6.68-6.64 (м, 2H), 6.52-6.44 (м, 2H), 2.29 (с, 6H).

HRMS (ESI) m/z : $[\text{M}]^+$ расч. для $\text{C}_{66}\text{H}_{46}\text{IrN}_6$ 1115.3414; получено 1115.3418.

$[\text{Ir}(\text{phi})_2(\text{dcmbpy})]$. Красный порошок, выход 58 %.

Спектр ЯМР ^1H , δ , м.д. (CDCl_3): 8.76 (д, $J = 5.6$ Гц, 2H), 8.56 (д, $J = 8.3$ Гц, 2H), 8.44 (д, $J = 8.5$ Гц, 2H), 8.36-8.31 (м, 4H), 8.08 (т, $J = 7.5$ Гц, 2H), 8.02-7.96 (м, 4H), 7.94-7.88 (м, 2H), 7.87-7.82 (м, 2H), 7.81-7.76 (м, 2H), 7.51 (т, $J = 7.5$ Гц, 2H), 7.32-7.28 (м, 4H), 7.25-7.22 (м, 2H), 7.06 (д, $J = 8.4$ Гц, 2H), 6.90-6.82 (м, 4H), 6.75 (т, $J = 7.6$ Гц, 2H), 6.62-6.56 (м, 2H), 3.96 (с, 6H).

HRMS (ESI) m/z : $[\text{M}]^+$ расч. для $\text{C}_{68}\text{H}_{46}\text{IrN}_6\text{O}_4$ 1203.3211; получено 1203.3209.

Состав и строение всех полученных соединений подтверждены спектроскопией ЯМР и масс-спектрометрией высокого разрешения. Сами спектры приведены в сопроводительных материалах к статье 2 из списка опубликованных работ по теме диссертации.

3.4.4. Измерение спектров люминесценции

Спектры фотолюминесценции регистрировали в растворах в хлористом метиле, насыщенных аргоном. Выбор единого растворителя существен, поскольку испускание бис-циклометаллированных комплексов иридия(III) происходит из состояний с переносом заряда, энергия которых зависит от полярности среды, вследствие чего значения, полученные в разных растворителях, между собой

несопоставимы. Удаление кислорода, необходимость которого отмечена в разделе 3.4.2, приводит, кроме того, условия измерения в соответствие с теми, для которых в базе данных отображены записи.

Определялось только положение максимума испускания. Квантовые выходы для полученных комплексов не измерялись. Всего снято 33 спектра люминесценции. Помимо двенадцати полученных в настоящей работе комплексов, в это число вошли соединения того же структурного типа, синтезированные ранее в той же лаборатории, причем для части из них спектры были зарегистрированы заново в хлористом метиле. Существенно, что все 33 спектра получены в одной лаборатории в одинаковых условиях, вследствие чего проверочная выборка внутренне согласована и свободна от межлабораторного разброса, неизбежного при сопоставлении литературных данных. Максимумы испускания охватывают диапазон от 509 до 629 нм. Проверка построенных моделей на этих данных, выполненная поэтому только для длины волны испускания, изложена в разделе 4.2.3.

3.4.5. Квантово-химические расчеты

Расчеты выполнены в рамках теории функционала плотности с использованием гибридного функционала B3LYP. Для легких элементов применен стандартный базисный набор def2-SVP, для атома иридия - Штутгарт-Дрезденский псевдопотенциал. Расчеты проведены в пакете программ ORCA 5.0[186]. Для оптимизированных структур выполнен расчет частот колебаний, мнимых частот не обнаружено, что подтверждает отвечающие им локальные минимумы. Влияние среды учтено поляризуемой континуальной моделью C-PCM с параметрами хлористого метилена, то есть того же растворителя, в котором проводились измерения. Длину волны испускания вычисляли методом Δ SCF как разность энергий триплетного и синглетного состояний, рассчитанных в точке оптимизированного триплетного минимума.

3.5. Интерактивные онлайн-инструменты

Инструменты представляют собой веб-приложения, написанные на языке Python средствами библиотеки Streamlit и размещенные в облачной среде Streamlit Community Cloud. Данные загружаются из тех же файлов, что депонированы в Zenodo, вследствие

чего расхождение между содержимым приложения и опубликованным набором исключено. Поиск по структуре и все операции над структурами - канонизация, сопоставление по подструктуре и вычисление меры сходства - выполняются средствами RDKit. Структурную формулу пользователь задает либо строкой SMILES, либо рисованием в графическом редакторе, встроенном в приложение через компонент streamlit-ketcher, что избавляет от необходимости владеть строчной записью структур. Рисунки строятся средствами plotly и допускают изменение масштаба и считывание значений. Приложение для IrLumDB, кроме того, содержит обученные модели в виде сохраненных файлов, вследствие чего прогнозирование выполняется без обращения к внешним службам. Исходный код приложений размещен в тех же репозиториях, что и код обработки данных.

4. Обсуждение результатов

4.1. Базы данных растворимости органических соединений

4.1.1. BigSolDB 2.0

Как показано в разделе 2.2.3, наборы данных о растворимости в органических растворителях немногочисленны и по объему уступают водным на порядок, тогда как именно неводные среды определяют выбор условий синтеза, выделения и кристаллизации органических веществ. Восполнение этого пробела составило первую задачу настоящей работы, результатом которой стал набор BigSolDB 2.0. Общий порядок его создания представлен на рис. 14.



Рисунок 14. Общая схема создания базы данных BigSolDB 2.0

Сбор проводился по процедуре, изложенной в разделе 3.1. Ручному просмотру подверглись 824 найденные работы. Полученные сведения были объединены с набором BigSolDB 1.0, содержащим 54273 значения для 839 соединений в 138 растворителях из 771 статьи. Этот набор представляет собой более раннюю работу автора, выполненную с иным составом соавторов. Он был размещен в виде препринта и депонирован в общедоступном репозитории, однако журнального рецензирования не проходил и в число публикаций, по которым выполнена настоящая работа, не входит[187]. При объединении он не переносился в готовом виде: все его записи прошли ту же процедуру проверки и стандартизации, что и вновь собранные, включая перекрестную проверку, снятие дубликатов, сверку идентификаторов источников и приведение структур к каноническому виду. Самостоятельным результатом настоящей работы является поэтому BigSolDB 2.0, а не его предшественник. В итоговом виде BigSolDB 2.0 содержит 103944 экспериментальных значения растворимости для 1448 органических соединений в 213 индивидуальных растворителях, извлеченных из 1595 рецензируемых публикаций, в диапазоне температур от 243 до 425 K.

Собранный массив проходил стандартизацию и контроль качества по порядку, изложенному в разделах 3.1.3 и 3.1.4. Существенно для дальнейшего изложения одно

принятое там правило: устранялись только точные дубликаты, тогда как независимые повторные измерения разных групп сохранялись, поскольку именно они позволяют оценивать воспроизводимость.

Пересчет растворимости из мольных долей в молярную шкалу требует знания плотности раствора. Обычно ее принимают равной плотности чистого растворителя, не оговаривая вносимой этим погрешности. Здесь плотности чистых растворителей были собраны из рецензируемых публикаций для всех растворителей, представленных в наборе более чем сотней записей, и опубликованы отдельным файлом с указанием температуры и источника, что позволило учитывать их температурную зависимость вместо подстановки единственного значения. Приближение, пренебрегающее вкладом растворенного вещества в объем раствора, при этом сохраняется и ограничивает точность пересчета для концентрированных растворов. Порядок пересчета изложен в разделе 3.1.3. Этот набор содержит 2210 значений плотности для 70 растворителей в диапазоне от 243 до 373 K, извлеченных из 146 публикаций и из справочника NIST Chemistry WebBook. Собранные плотности представляют тем самым самостоятельный результат, пригодный к использованию и вне задачи о растворимости. Состав полей набора приведен в табл. 3.

Таблица 3. Описание полей базы данных BigSolDB 2.0

Столбец	Описание	Тип
SMILES_Solute	Структура растворенного вещества в нотации SMILES	строка
Temperature_K	Температура измерения, K	число
Solvent	Название растворителя	строка
SMILES_Solvent	Структура растворителя в нотации SMILES	строка
Solubility(mole_fraction)	Растворимость в мольных долях, как приведено в источнике	число
Solubility(mol/L)	Растворимость, пересчитанная в молярную шкалу, моль/л	число
LogS(mol/L)	Десятичный логарифм растворимости в молярной шкале	число
Compound_Name	Название растворенного вещества	строка

Столбец	Описание	Тип
CAS	Регистрационный номер CAS растворенного вещества	строка
PubChem_CID	Идентификатор растворенного вещества в PubChem	строка
FDA_Approved	Признак одобрения FDA	строка
Source	Цифровой идентификатор публикации-источника	строка

Статистика набора представлена на рис. 15. Распределение по абсолютным значениям растворимости оказывается крайне неравномерным: 52677 записей, то есть более половины набора, отвечают мольной доле ниже 0.01. Для обучения моделей такое представление непригодно, поскольку почти весь диапазон изменения величины сосредоточен вблизи нуля. В логарифмической шкале распределение близко к одномодальному с отчетливым максимумом около минус единицы и хвостами, простирающимися до минус девяти в области крайне малой растворимости и до двух в области высокой. Именно поэтому в качестве целевой величины при построении моделей используется логарифм растворимости, а не сама растворимость. Среди растворителей, как по числу записей, так и по числу различных растворенных веществ, отчетливо преобладают низшие спирты, вода, этилацетат, ацетон и ацетонитрил.

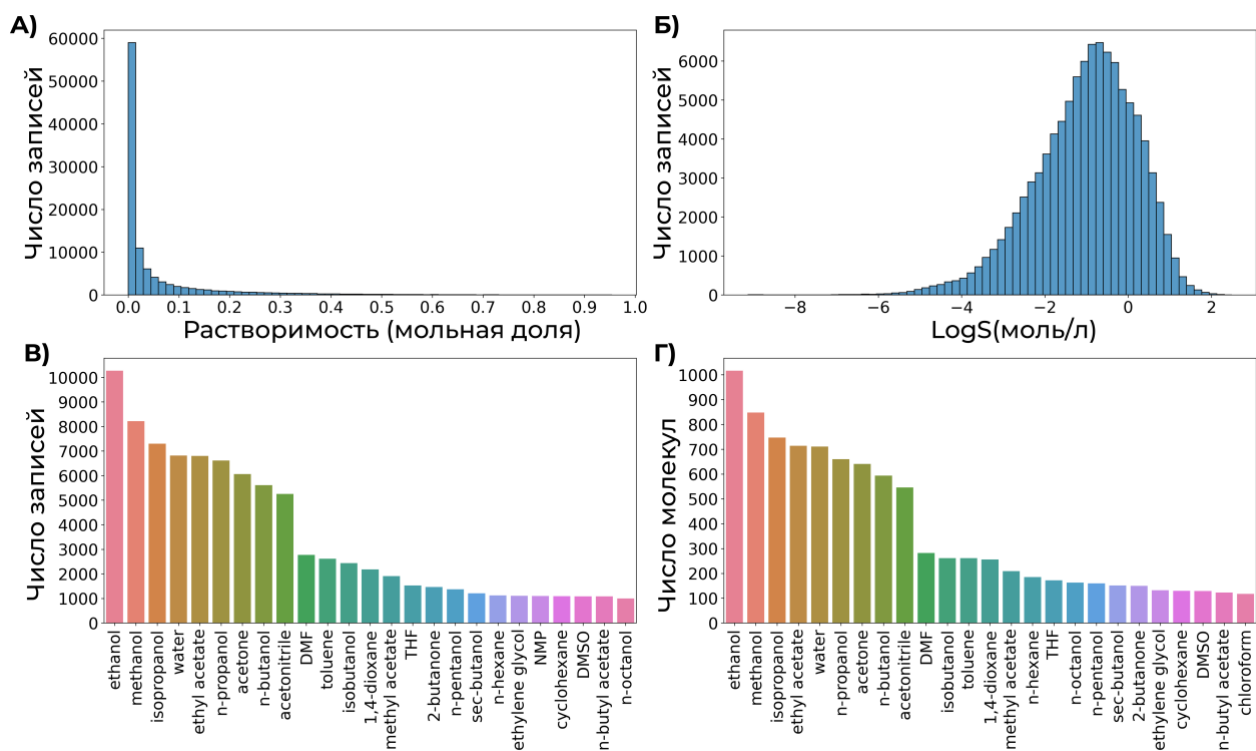


Рисунок 15. Статистика базы данных BigSolDB 2.0. А - распределение записей по растворимости в мольных долях. Б - распределение записей по десятичному логарифму растворимости в молярной шкале. В - двадцать пять наиболее представленных растворителей по числу записей. Г - те же растворители по числу различных растворенных веществ

Сопоставление с наборами, перечисленными в табл. 1, показывает, что по числу записей BigSolDB 2.0 уступает лишь компиляциям, полученным слиянием ранее опубликованных ресурсов, и превосходит все известные нам наборы, собранные непосредственно из первичной литературы. Существенно при этом, что каждое значение сопровождается указанием температуры, растворителя и публикации-источника, тогда как в наборах, полученных агрегацией, эти сведения, как показано в разделе 2.1.1, к моменту второго или третьего заимствования оказываются в значительной мере утраченными.

4.1.2. MixtureSolDB

Как отмечено в разделе 2.2.4, переход от индивидуальных растворителей к их смесям означает качественное усложнение задачи. Зависимость растворимости от состава смеси существенно нелинейна и не выводится из значений для чистых

компонентов, а имевшиеся наборы данных исчислялись десятками тысяч записей при крайне узком круге веществ. Вторая задача работы состояла в создании набора, пригодного для обучения моделей в этих условиях. Результатом стал набор MixtureSolDB, порядок создания которого представлен на рис. 16.

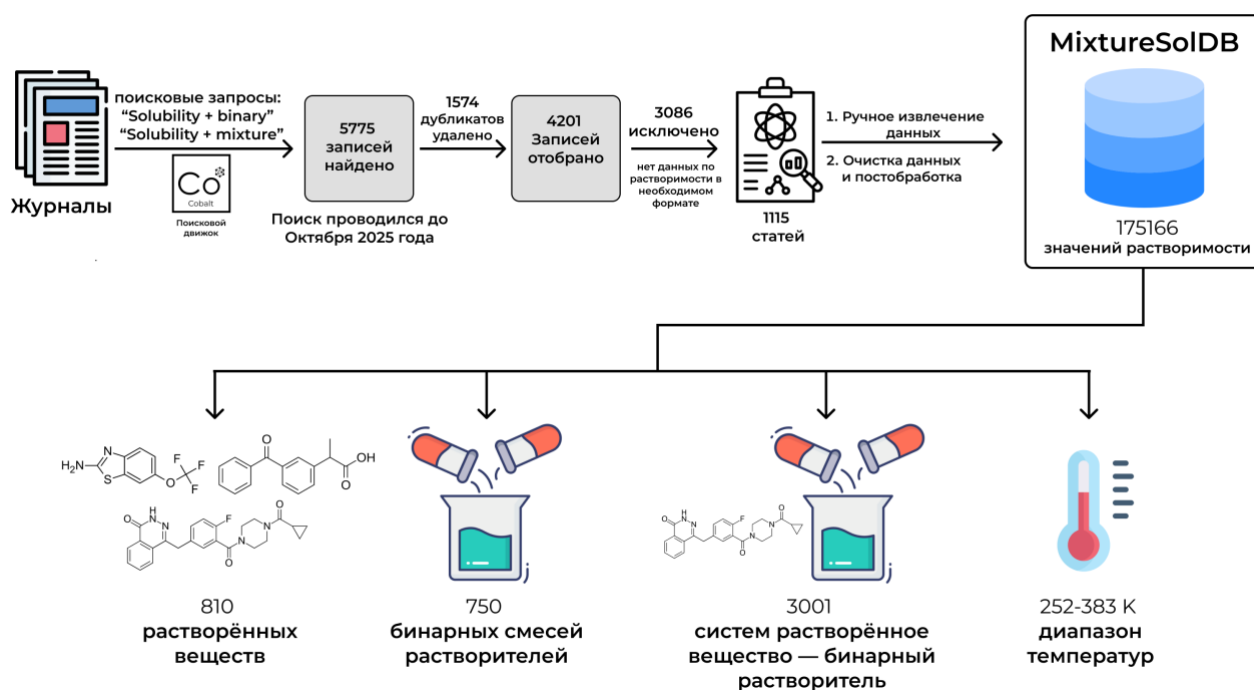


Рисунок 16. Схема отбора публикаций и общее устройство базы данных MixtureSolDB

Ход отбора зафиксирован в виде диаграммы, построенной по образцу, принятому в систематических обзорах: найдено 5775 записей, из них 1574 признаны дубликатами, 4201 уникальная запись просмотрена, 3086 отсеяно ввиду отсутствия данных в пригодном виде, и 1115 статей отобрано для ручного извлечения. Такое изложение процедуры отбора отвечает требованию прослеживаемости, рассмотренному в разделе 2.1.4, и позволяет судить о полноте охвата, а не принимать ее на веру.

Устройство набора подчинено задаче однозначного прочтения каждой записи. Каждой строке присвоен устойчивый идентификатор вида MSD0175166, сохраняющийся при обновлениях набора. Пары растворителей приведены к алфавитному порядку, так что первый компонент всегда предшествует второму, а доли при необходимости пересчитаны соответственно. Этот прием устраняет неоднозначность, при которой одна и та же смесь могла бы быть записана двумя

способами и попасть в набор дважды. Растворимость приведена одновременно в четырех дополняющих друг друга формах: в мольных долях, в граммах на 100 граммов растворителя и в виде десятичных логарифмов обеих величин. Отдельным признаком помечены записи, отвечающие чистым растворителям, то есть предельным составам смеси. Таких записей 28870, или 16.5 % набора, и наличие их позволяет сопоставлять смеси с индивидуальными средами в пределах одного набора, не прибегая к внешним источникам. Состав полей приведен в табл. 4.

Таблица 4. Описание полей базы данных MixtureSolDB

Столбец	Описание	Тип
RecordID	Устойчивый идентификатор записи	строка
SMILES_Solute	Структура растворенного вещества в нотации SMILES	строка
Temperature_K	Температура измерения, К	число
Solubility(mole_fraction)	Растворимость в мольных долях	число
LogS(mole_fraction)	Десятичный логарифм растворимости в мольных долях	число
Solubility(g/100g)	Растворимость в граммах вещества на 100 граммов растворителя	число
LogS(g/100g)	Десятичный логарифм растворимости в массовой шкале	число
Solvent1	Название первого компонента смеси	строка
Solvent2	Название второго компонента смеси	строка
SMILES_Solvent1	Структура первого компонента в нотации SMILES	строка
SMILES_Solvent2	Структура второго компонента в нотации SMILES	строка
Fraction_Solvent1	Доля первого компонента в смеси до внесения вещества	число
Fraction_Solvent2	Доля второго компонента в смеси до внесения вещества	число
Fraction_Type	Вид доли: мольная или массовая	строка
IsPureSolventEndpoint	Признак записи для чистого растворителя	логический
Compound_Name	Название растворенного вещества	строка
CAS	Регистрационный номер CAS растворенного вещества	строка

Столбец	Описание	Тип
PubChem_CID	Идентификатор растворенного вещества в PubChem	число
FDA_Approved	Признак одобрения FDA	логический
Source	Цифровой идентификатор публикации-источника	строка

Отдельно оговорены и пределы полноты набора. Для 130 записей не приведен регистрационный номер CAS, а для 522 - идентификатор PubChem, поскольку соответствующих обозначений для этих структур в названных ресурсах попросту не существует.

В окончательном виде MixtureSolDB содержит 175166 экспериментальных значений растворимости для 810 органических соединений в 750 бинарных смесях, составленных из 139 растворителей, что дает 3001 систему вещество - бинарный растворитель, извлеченных из 1115 публикаций в диапазоне температур от 252 до 383 К. Статистика набора представлена на рис. 17. Как и для индивидуальных растворителей, распределение по абсолютным значениям растворимости резко смещено к малым величинам, тогда как в логарифмической шкале оно близко к одномодальному. Состав набора отчетливо смещен в сторону водно-органических систем: вода входит в качестве одного из компонентов в 57 % записей, а преобладают ее смеси с низшими спиртами. Это смещение отражает распределение интересов исследователей и представляет собой частный случай публикационного смещения, рассмотренного в разделе 2.1.2.

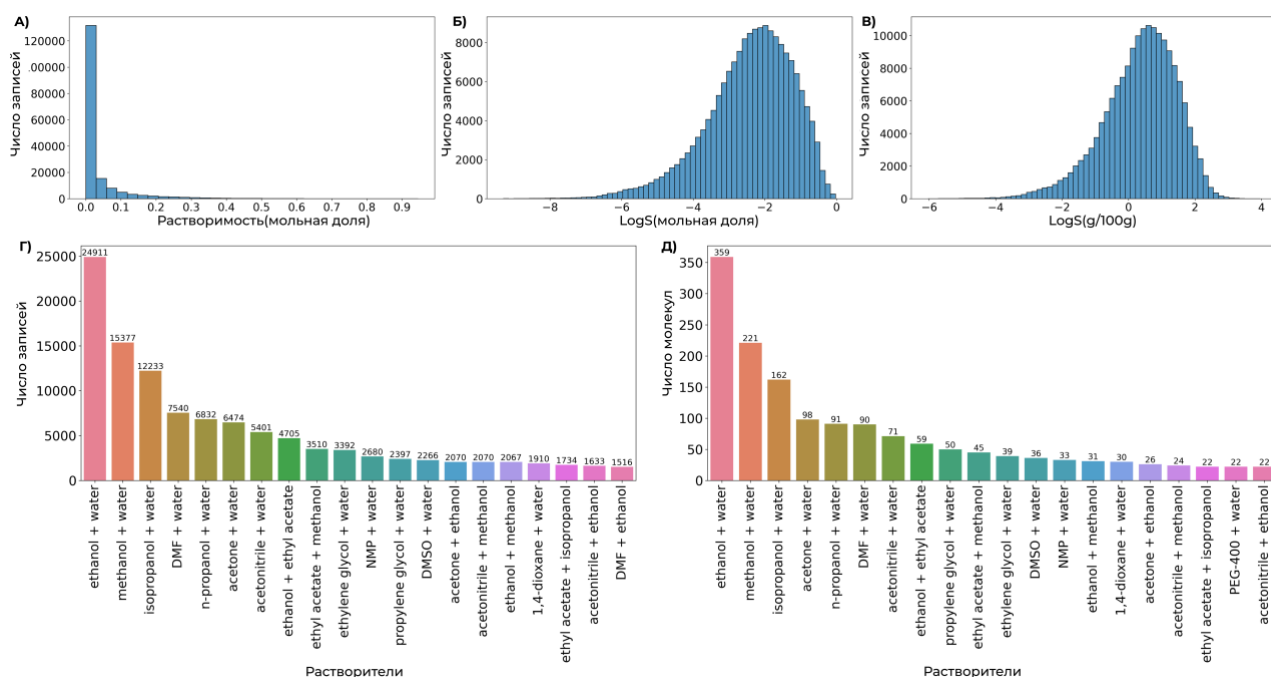


Рисунок 17. Статистика базы данных MixtureSolDB. А - распределение записей по растворимости в мольных долях. Б - распределение записей по десятичному логарифму растворимости в мольных долях. В - распределение записей по десятичному логарифму растворимости в массовой шкале. Г - двадцать наиболее представленных пар растворителей по числу записей. Д - те же пары по числу различных растворенных веществ

Прямое сопоставление с двумя ранее опубликованными общедоступными наборами приведено в табл. 5. Превышение достигнуто по всем шести показателям, однако величина его неодинакова. По числу значений растворимости, источников, растворенных веществ, бинарных смесей и систем вещество - смесь оно составляет от пяти до пяти с половиной раз, тогда как по числу индивидуальных растворителей, из которых смеси составлены, - около полутора. Это различие осмысленно: круг употребительных растворителей ограничен и близок к исчерпанию, тогда как число их сочетаний и число изученных в них веществ далеки от него. Существенно и то, что рост числа записей не сводится к уплотнению уже охваченной области: число различных систем вещество - смесь возросло с 545 до 3001, то есть расширен именно охват химического пространства.

Таблица 5. Сопоставление MixtureSolDB с ранее опубликованными наборами данных о растворимости в бинарных смесях растворителей: BaoDB[83] и MixSolDB[77]

Показатель	BaoDB, 2024[83]	MixSolDB, 2025[77]	MixtureSolDB
Значений растворимости	28098	34734	175166
Растворенных веществ	125	171	810
Растворителей	44	95	139
Бинарных смесей	110	153	750
Систем вещество - смесь	373	545	3001
Источников	126	209	1115

Наличие растворимости в нескольких формах ставит вопрос о выборе целевой величины при построении моделей. Для сопоставительных испытаний предпочтителен логарифм растворимости в мольных долях: эта величина безразмерна и потому сравнима между растворителями различной молярной массы. Для практических же задач, где требуется навеска, удобнее массовая шкала. Обе величины приведены в наборе, вследствие чего выбор остается за пользователем и не навязывается устройством данных.

4.1.3. Температура как полноправная переменная

Оба набора отличает от большинства ранее опубликованных то, что температура сохранена для каждой записи. Значение этого обстоятельства выходит за рамки полноты описания. Как отмечено в разделе 2.2.2, наборы растворимости в воде, на которых обучалось большинство моделей, относятся к одной температуре либо вовсе не содержат сведений о ней, вследствие чего построенные на них модели прогнозируют растворимость лишь при комнатных условиях. Распределение температур в собранных данных представлено на рис. 18.

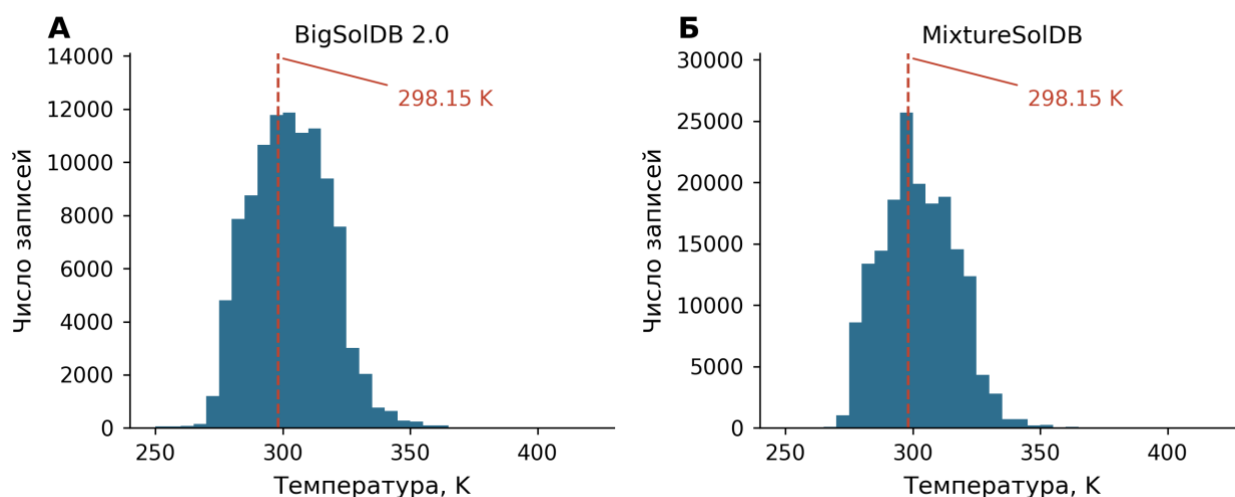


Рисунок 18. Распределение записей по температуре измерения. А - BigSolDB 2.0, Б - MixtureSolDB. Штриховой линией отмечена температура 298.15 К

Медиана в обоих наборах составляет 303.15 К. Стандартной температуре 298.15 К отвечают лишь 8.2 % записей BigSolDB и 12.1 % записей MixtureSolDB, а за пределами промежутка от 293.15 до 313.15 К лежит 48.0 и 43.8 % записей соответственно. Иными словами, около половины собранных данных относится к условиям, при которых модели, обученные на наборах с фиксированной температурой, неприменимы вовсе.

Существеннее, однако, что измерения образуют серии. Растворимость определяется, как правило, при нескольких температурах для одной и той же системы. Если считать системой сочетание растворенного вещества с растворителем, а для смесей - с парой растворителей заданного состава, то в BigSolDB таких систем 11263, а в MixtureSolDB 27055, причем медианное число различных температур на систему составляет девять и восемь. К системам, охваченным не менее чем пятью температурами, относится 99.2 % записей первого набора и 94.8 % второго. Именно это устройство данных позволяет обучать модели самой температурной зависимости, а не интерполировать между разрозненными значениями, и делает температуру полноправной входной переменной наравне со строением вещества и растворителя.

Следует, вместе с тем, оговорить и пределы. Охваченный диапазон определяется условиями, в которых растворимость измеряют на практике: выше 323.15 К лежит 8.0

% записей BigSolDB и 5.6 % записей MixtureSolDB, ниже 283.15 К - 6.9 и 6.1 % соответственно. Приведенные далее оценки качества моделей относятся к этому диапазону, и перенос их на существенно более высокие или низкие температуры был бы неоправдан.

4.1.4. Внутренняя согласованность собранных данных

Как показано в разделе 2.2.2, достигаемая точность прогнозирования растворимости ограничена воспроизводимостью самих измерений. Оценки этого предела, приводимые в литературе, получены, однако, на сторонних наборах, и переносить их на собранные данные без проверки нельзя. Устройство BigSolDB позволяет выполнить такую проверку непосредственно.

Эта возможность возникает благодаря принятому правилу снятия дубликатов. Устранялись лишь записи, совпадающие полностью и, судя по всему, восходящие к одному измерению, тогда как независимые измерения разных групп сохранялись. Вследствие этого в наборе имеются случаи, когда растворимость одного и того же вещества в одном и том же растворителе при одной и той же температуре измерена в нескольких работах порознь. Таких записей 6939, и относятся они к 176 веществам. Попарное сопоставление этих значений, дающее 3935 пар, приводит к среднеквадратичному отклонению 0.391 логарифмической единицы при средней абсолютной разности 0.146 и коэффициенте детерминации 0.90 (рис. 19, А).

Та же оценка была выполнена для MixtureSolDB, причем в расчет принимались только записи для бинарных смесей, а записи для чистых растворителей исключались. Совпадение условий требуется здесь одновременно по веществу, паре растворителей, составу смеси и температуре, вследствие чего пересечения между работами оказываются существенно более редкими. Независимые повторные измерения нашлись для восьми веществ из шестнадцати источников, что дало 460 пар значений. Среднеквадратичное отклонение составило 0.43 логарифмической единицы при коэффициенте детерминации 0.82 (рис. 19, Б). Оценку эту следует принимать лишь как проверку порядка величины: две трети пар приходится на два соединения, флорфеникол и 3,4,5-триметоксибензальдегид, а исключение одного лишь рисперидона, для которого две работы расходятся более чем на логарифмическую

единицу, понижает отклонение до 0.36. Существенно, однако, что разброс в смесях оказывается того же порядка, что и в индивидуальных растворителях, а не заметно большим.

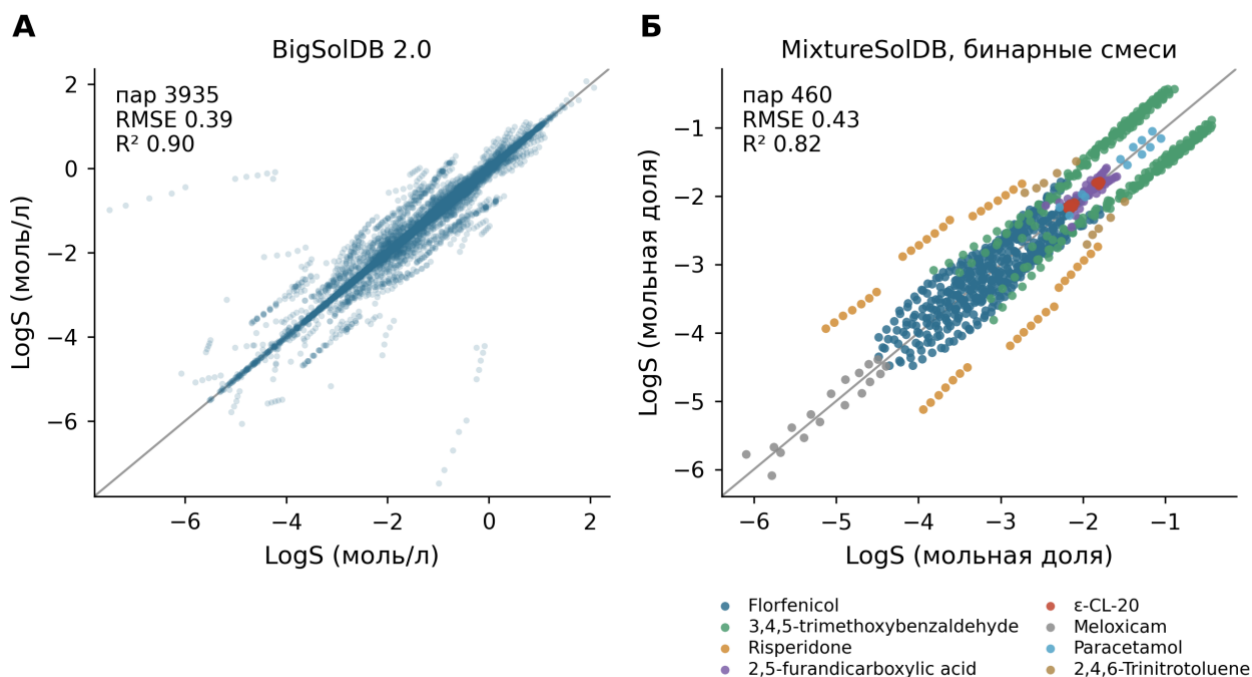


Рисунок 19. Сопоставление независимых измерений растворимости, выполненных разными группами в одинаковых условиях. А - BigSolDB 2.0. Б - MixtureSolDB, только бинарные смеси. Цветом обозначены растворенные вещества. По обеим осям отложена одна и та же величина, измеренная в двух разных работах. Каждая пара нанесена в обеих перестановках, серой линией показано совпадение значений

Полученные величины заслуживают двоякого прочтения. С одной стороны, они несколько ниже оценок межлабораторного разброса, полученных ранее для литературных данных о растворимости[65], и ниже нижней границы алеаторного предела в 0.5-1 логарифмическую единицу, приводимой для доступных наборов[66]. Это свидетельствует о внутренней согласованности собранных данных и, косвенно, о действенности примененной процедуры перекрестной проверки, описанной в разделе 3.1.4. С другой стороны, 0.39 и 0.43 логарифмической единицы задают предел, ниже которого точность модели, обученной на соответствующих данных, опуститься не может, сколь бы сложной ни была ее архитектура. Величина, о которой сообщалось бы

как о существенно меньшей, указывала бы на утечку между обучающей и проверочной выборками.

Существенно и то, что эта оценка получена на самих собранных данных, а не заимствована из литературы. Наборы, полученные слиянием ранее опубликованных ресурсов, такой возможности лишены: в них независимые повторные измерения либо усреднены, либо неотличимы от дубликатов, возникших при последовательном заимствовании, о чем говорится в разделе 2.1.2. Сохранение повторных измерений, представляющееся на первый взгляд избыточностью, оказывается тем самым свойством набора, которое позволяет судить о пределе его собственной точности.

Этот прием применяется и другими. При составлении наборов коллекции SolProp дубликаты также сохранялись намеренно, именно чтобы оценить влияние различных источников на измеряемую величину, и авторы сообщают, что у 12 % повторяющихся записей расхождение превышает 0.2 логарифмической единицы[69]. Для BigSolDB та же доля составляет 18.7 %. Это сопоставление показательно вдвойне. Оно подтверждает, что заметное расхождение между независимыми измерениями свойственно литературным данным о растворимости как таковым, а не является особенностью собранных здесь наборов. Оно же вскрывает и устройство разброса: медианное расхождение в BigSolDB составляет всего 0.026 логарифмической единицы, тогда как 8.6 % пар расходятся более чем на 0.5, а 1.9 % - более чем на логарифмическую единицу. Подавляющее большинство повторных измерений, таким образом, согласуется превосходно, а приведенные выше средние величины определяются сравнительно небольшим хвостом сильно расходящихся случаев. Именно на него и следует направлять дальнейшую проверку данных.

4.1.5. Сведения о методике измерения и твердой фазе

Рассмотренные в разделе 2.2.1 обстоятельства - зависимость измеряемой величины от метода определения, продолжительности установления равновесия и природы твердой фазы - в опубликованных наборах данных о растворимости не отражены. Не отражены они и в версии 2.0: каждая запись сопровождается указанием температуры, растворителя и источника, но не условий, при которых значение получено. Вследствие этого пользователь набора лишен возможности как отобразить

записи по строгости методики, так и оценить, какая доля разброса приходится на различие условий, а не на свойства вещества.

Для устранения этого пробела первичные публикации были просмотрены повторно, и извлеченные из них сведения об условиях эксперимента вынесены в два вспомогательных файла. Первый описывает методику измерения и содержит по одной записи на публикацию: способ определения концентрации, тип равновесия, продолжительность выдержки и отстаивания, способ и скорость перемешивания, число повторов, способ термостатирования и давление. Второй описывает использованное вещество и содержит по одной записи на каждое сочетание публикации, вещества и твердой формы: чистоту и способ ее определения, поставщика, выполненный контроль твердой фазы и его исход, температуру и энтальпию плавления с обязательным указанием происхождения каждой из этих величин. Обозначение твердой формы, приведенное в первоисточнике, вынесено при этом на уровень отдельного измерения, поскольку работы, посвященные полиморфизму, сопоставляют несколько форм одного вещества внутри одной статьи.

Охват полученного слоя неоднороден и сам по себе характеризует сложившуюся практику публикации. Способ определения концентрации удалось установить для всех 1842 публикаций, способ перемешивания - для 94 %, продолжительность выдержки - для 79 %, тогда как скорость перемешивания приводится лишь в 17 % работ. Рентгенофазовый контроль твердой фазы до и после установления равновесия выполнен в 1003 случаях из 2295. Температура и энтальпия плавления приведены одновременно в 1041 случае, причем в 819 из них обе величины измерены в той же работе, что и растворимость, что позволяет вычислять коэффициенты активности без привлечения посторонних источников. Обозначение твердой формы указано для 14813 измерений, охватывающих 201 вещество, из которых 29 представлены более чем одной формой.

Распределение измерений по способу определения концентрации приведено на рис. 20. Преобладают гравиметрический вариант изотермического насыщения и высокоэффективная жидкостная хроматография, на которые вместе приходится более двух третей набора. Лазерный синтетический метод, дающий температуру растворения

при заданном составе, а не концентрацию при заданной температуре, составляет еще шестую часть. Существенно, что доли, вычисленные по числу измерений и по числу публикаций, заметно расходятся: работы, использующие ультрафиолетовую спектрофотометрию, дают в среднем вдвое меньше точек на статью, вследствие чего этот метод, составляя шестую часть публикаций, дает лишь четырнадцатую часть измерений.

Одновременно с введением дополнительных полей в основной файл добавлен десятичный логарифм мольной доли. Эта величина первична, тогда как молярная шкала, как указано в разделе 2.2.1, получается пересчетом с привлечением плотности и потому наследует его погрешность. Сопоставление версий набора приведено в табл. 6.

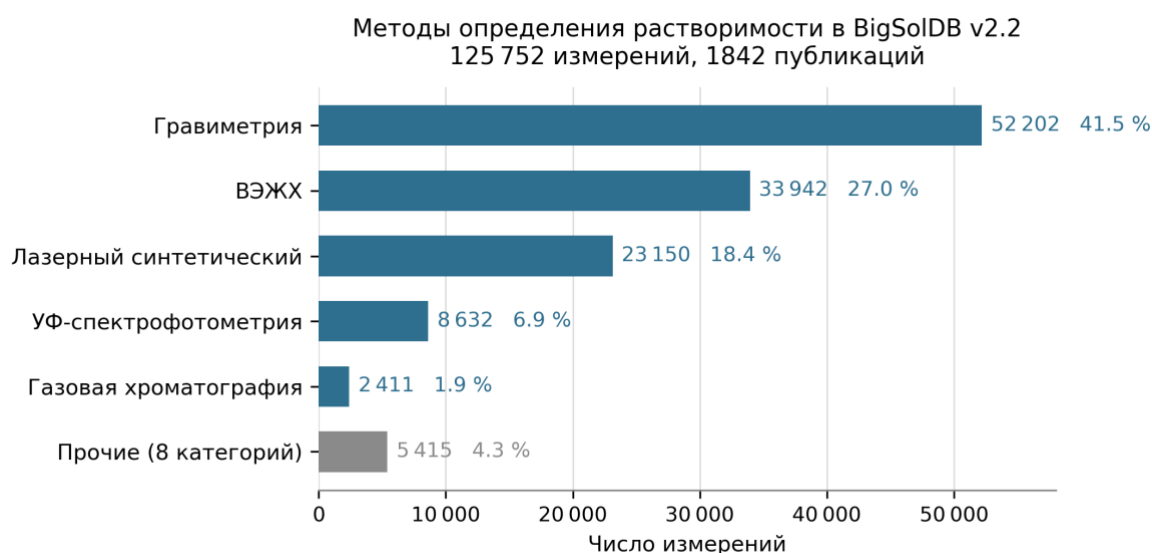


Рисунок 20. Методы определения растворимости в BigSolDB. Приведено число измерений и доля от общего числа. Категории, каждая из которых охватывает менее двух процентов набора, объединены.

Таблица 6. Сопоставление версий набора данных BigSolDB

Показатель	2.0	2.1	2.2
Измерений	103944	112465	125752
Публикаций	1595	1687	1842
Растворенных веществ	1448	1525	1627
Растворителей	213	218	221

Сведения о твердой форме	нет	нет	14813 записей
Сведения о методике измерения	нет	нет	1842 публикации
Сведения о веществе и контроле твердой фазы	нет	нет	2295 записей
Записей с измеренной температурой плавления	нет	нет	75220
Записей с измеренной энтальпией плавления	нет	нет	60831
Записей, для которых обе величины измерены в той же работе	нет	нет	60479
Записей, пригодных для расчета коэффициента активности	нет	нет	59680

Наличие температуры и энтальпии плавления, измеренных в той же работе, что и растворимость, позволяет вычислить коэффициент активности насыщенного раствора без привлечения посторонних источников. Эта величина получается отнесением идеальной растворимости, отвечающей уравнению Шредера - Ле Шателье, к измеренной и характеризует отклонение раствора от идеальности, то есть вклад собственно растворителя, отделенный от вклада кристаллической решетки. Расчет выполнен для 59680 измерений, охватывающих 722 вещества, 146 растворителей и 673 публикации. Исключены записи, для которых обнаружено превращение твердой фазы, лежащие выше температуры плавления, а также относящиеся к веществам, разлагающимся ниже плавления. Полученные величины распределены по классам растворителей закономерным образом (рис. 21). Растворители, не отнесенные ни к одному из выделенных классов, на рисунке не показаны. Отрицательное отклонение от идеальности дают лишь апротонные акцепторы водородной связи, показанные на рисунке по отдельности: N,N-диметилацетамид с медианой 0.34, N-метилпирролидон с 0.37, N,N-диметилформамид с 0.60 и диметилсульфоксид с 0.65. У диметилацетамида коэффициент активности меньше единицы в 79 % записей. Далее в порядке возрастания следуют кетоны и простые эфиры, лежащие вблизи идеальности, сложные эфиры и галогенуглеводороды, нитрилы и спирты, ароматические углеводороды, и, наконец, алканы и вода с медианами 34.8 и 87.6. Этот ряд воспроизводит известную зависимость от полярности и способности к образованию водородной связи, причем

получен он без единого подгоночного параметра, из одних лишь измеренных величин, и служит тем самым независимой проверкой качества собранных данных о плавлении.

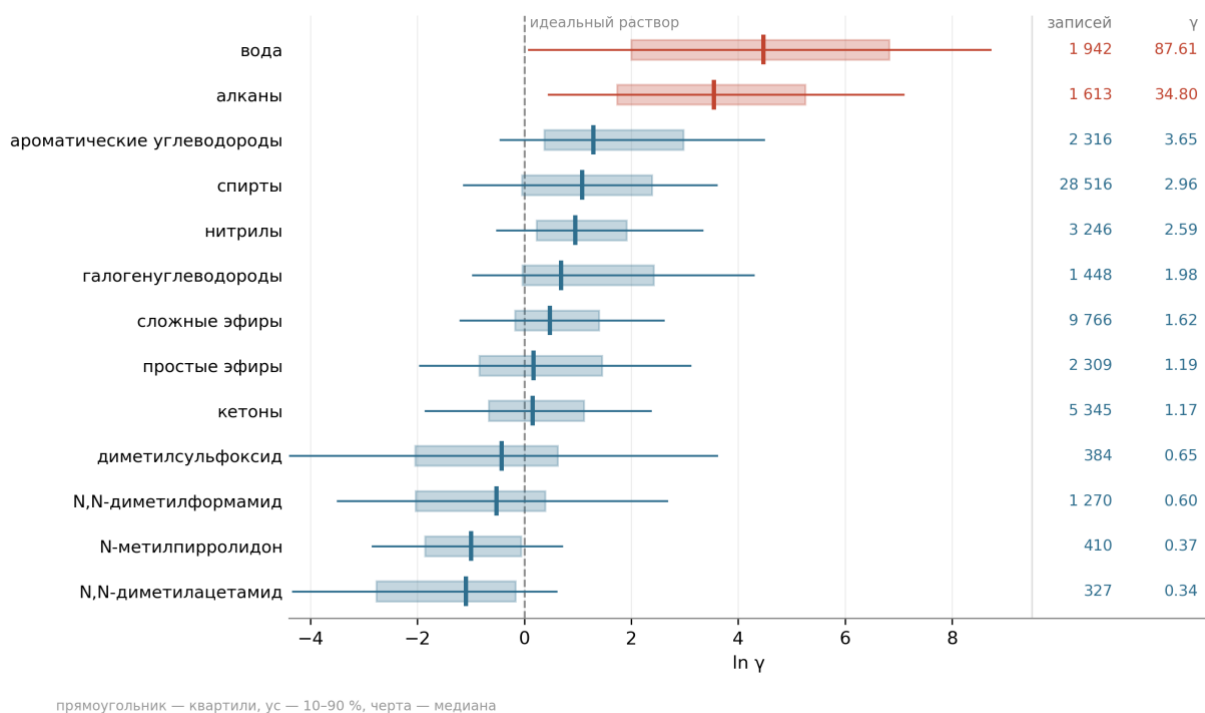


Рисунок 21. Коэффициенты активности насыщенных растворов по классам растворителей. Приведены записи, для которых температура и энтальпия плавления измерены в той же работе, что и растворимость. Штриховая линия отвечает идеальному раствору. Справа указаны число записей и медианное значение коэффициента активности.

Разметка твердой формы, введенная в версии 2.2, позволяет отделить вклад полиморфизма от межлабораторного разброса. Сопоставлялись пары измерений, совпадающие по веществу, растворителю и температуре. Единицей счета служила независимая серия, отвечающая одному сочетанию вещества, растворителя, пары форм и пары источников. Для 24 серий, в которых обе формы измерены в пределах одной работы, а следовательно лаборатория, методика и партия растворителя совпадают, медиана расхождения составляет 0.125 логарифмической единицы, тогда как для 154 серий, отвечающих одной форме у разных групп, - 0.025 (рис. 22). Вклад твердой формы превышает, таким образом, межлабораторный разброс приблизительно в пять раз. Это различие не сводится к нескольким выпадающим значениям: межквартильные интервалы двух групп почти не перекрываются, составляя 0.075-0.409

и 0.016-0.057 соответственно. В отношении растворимостей это соответствует 1.33 по медиане и 2.6 в девяностом процентиле, что согласуется с обобщением литературных данных, согласно которому растворимости полиморфов различаются обычно менее чем вдвое[39], и распространяет этот вывод на органические растворители, тогда как исходная оценка получена для водных сред. Различие сильно зависит от вещества: для пары метастабильной и устойчивой форм глутатиона оно достигает 0.70 логарифмической единицы, то есть пятикратного отношения, тогда как для индометацина не превышает 0.05. Оценка опирается на семь соединений и потому характеризует порядок величины, а не распределение по химическому пространству.

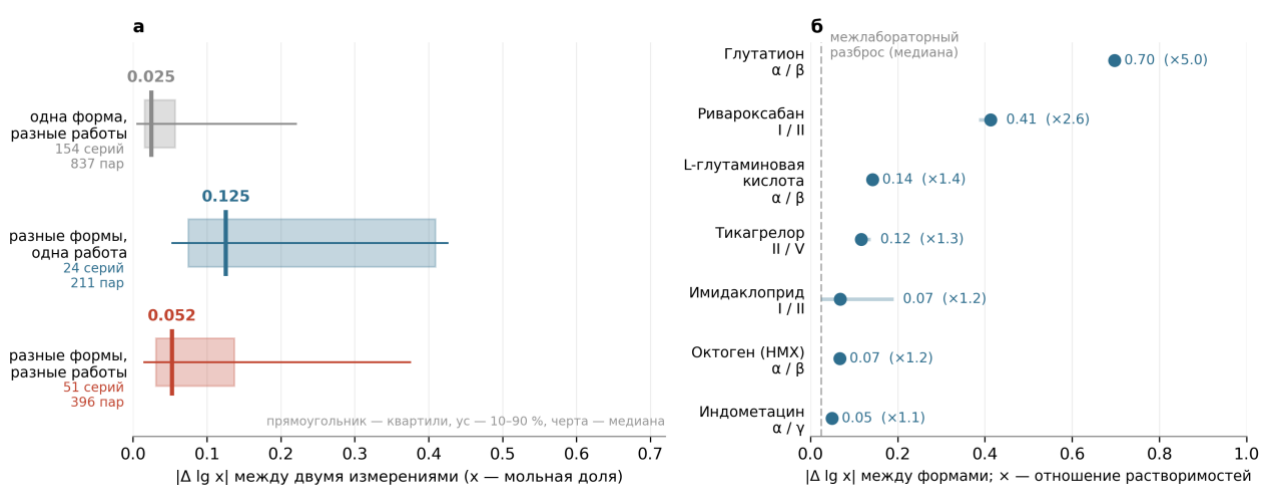


Рисунок 22. Вклад твердой формы в расхождение литературных данных о растворимости. А - распределение модуля разности десятичных логарифмов мольной доли для трех групп пар измерений, совпадающих по веществу, растворителю и температуре. Б - величина различия между формами по отдельным веществам. В скобках указано отношение растворимостей. Прямоугольник отвечает квартилям, ус - интервалу от десятого до девяностого процентиля и черта - медиане.

4.2. Люминесцентные свойства комплексов иридия(III)

4.2.1. База данных IrLumDB

Как показано в разделе 2.4.2, расчетные подходы к прогнозированию фотофизических свойств комплексов иридия(III) опираются либо на квантово-химические вычисления, требующие трехмерных структур, либо на единственный высокопроизводительный набор, ограниченный одним структурным типом.

Восполнение этого пробела потребовало систематического сбора экспериментальных сведений из первичной литературы, результатом которого стала база данных IrLumDB, а на ее основе - модели прогнозирования, проверенные затем на специально синтезированных соединениях (рис. 23).

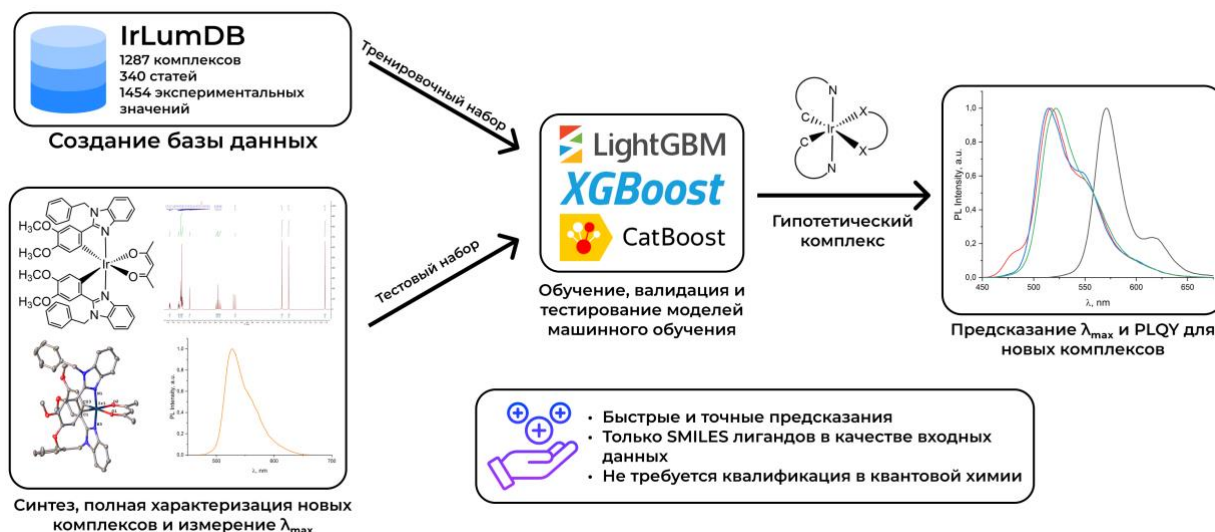


Рисунок 23. Общая схема работы: создание базы данных IrLumDB, синтез и характеристика новых комплексов, обучение моделей и прогнозирование длины волны максимума испускания и квантового выхода

Сбор проводился по описанной в разделе 3.1 процедуре. Извлекались длина волны максимума испускания, квантовый выход и сведения об условиях измерения, перечисленные в разделе 3.1.2.

Охват набора был сознательно ограничен октаэдрическими бис-циклометаллированными комплексами иридия(III), в которых два C,N-циклометаллирующие лиганда расположены транс друг к другу по атомам азота, а третий, вспомогательный бидентатный лиганд находится в транс-положении к связям Ir-C. Основанием послужило то, что именно это семейство наиболее многочисленно среди люминесцентных соединений иридия и допускает наиболее широкую настройку фотофизических свойств. Их геометрические цис-изомеры встречаются крайне редко, а число описанных трис-циклометаллированных комплексов, также широко применяемых в качестве излучателей, по данным Кембриджской структурной базы, уступает числу бис-циклометаллированных примерно в пять раз. Такое ограничение

существенно и с точки зрения представления: в пределах этого семейства взаимное расположение лигандов вокруг иона иридия одинаково, вследствие чего комплекс однозначно задается простым перечислением его лигандов, а место циклометаллирования указывается формальным депротонированием соответствующего атома углерода. Тем самым затруднение, разобранный в разделе 2.3.1 и состоящее в неприменимости линейных нотаций к координационным соединениям, обходится без введения новых форматов записи.

Существенной особенностью набора является то, что для каждой записи фиксировался растворитель, в котором проводилось измерение. Необходимость этого вытекает из зависимости энергии испускания от полярности среды, рассмотренной в разделе 2.4.1. Отдельно фиксировались и условия определения квантового выхода, поскольку присутствие кислорода тушит фосфоресценцию: для большинства записей в источниках указана инертная атмосфера, а все отступления от этих условий вынесены в отдельный столбец. Изложенное в разделе 2.1.3 применительно к утрате сведений об условиях эксперимента относится к рассматриваемым данным в полной мере, и приведенное разделение позволяет отбирать для обучения моделей однородные подмножества.

В окончательном виде база данных содержит 1454 отдельных измерения люминесцентных свойств для 1287 комплексов иридия(III), извлеченных из 340 рецензируемых публикаций, и депонирована в общедоступном репозитории под постоянным идентификатором (табл. 15).

Разбор собранных данных обнаруживает ряд закономерностей, характеризующих как само семейство соединений, так и сложившуюся практику его исследования.

Прежде всего, в области 450-650 нм набор покрывает пространство свойств полностью: для каждой длины волны в этом промежутке представлены комплексы с любыми значениями квантового выхода, от близких к нулю до близких к единице (рис. 24). За пределами же этой области, при длинах волн свыше 650 нм, обнаруживается заметное смещение к малым квантовым выходам. Это смещение не является

особенностью выборки и отвечает закону энергетического зазора, разобранным в разделе 2.4.1: с уменьшением энергии излучательного перехода скорость безызлучательной дезактивации возрастает. Тем самым набор воспроизводит независимо установленную физическую закономерность, что служит косвенным свидетельством его достоверности.

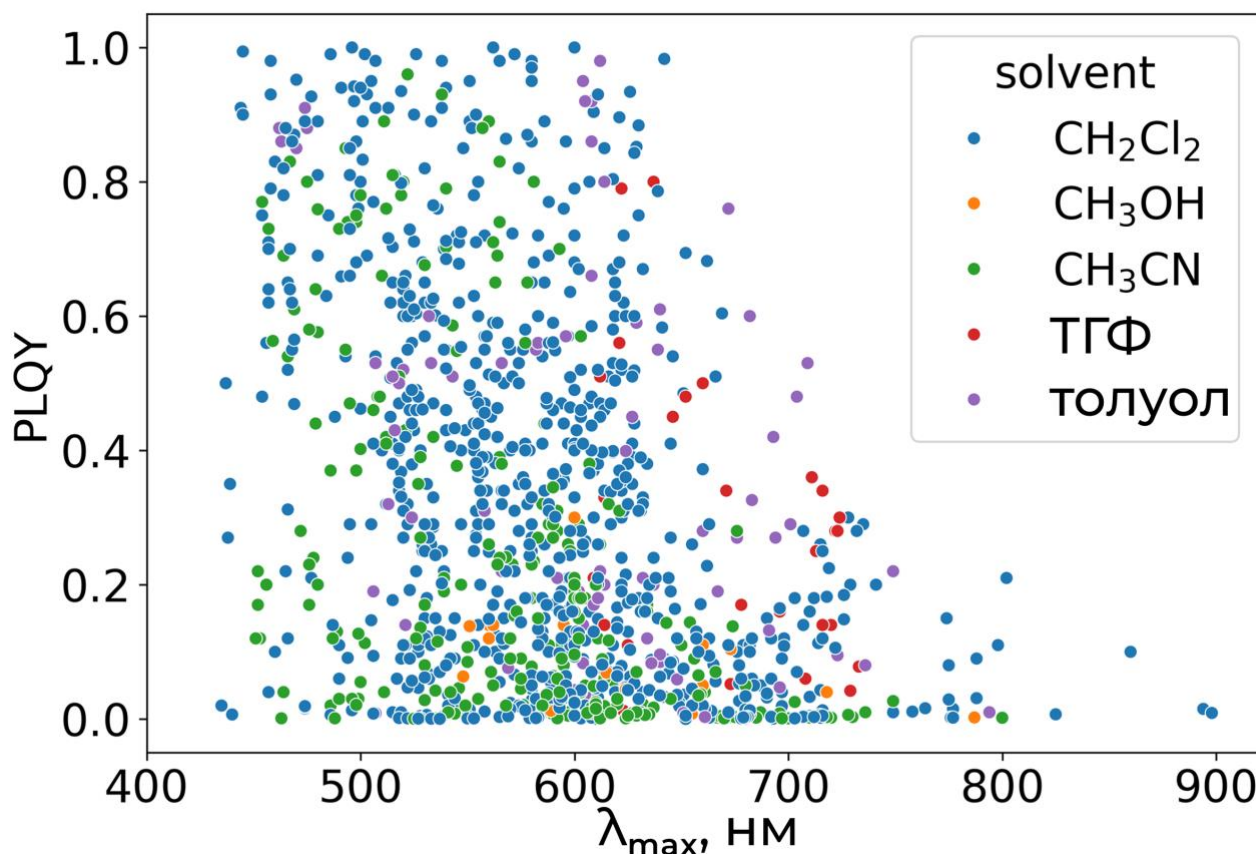


Рисунок 24. Распределение квантовых выходов и максимумов испускания бис-циклометаллированных комплексов иридия(III) в базе данных IrLumDB. Показаны пять наиболее употребительных растворителей

Распределение по длинам волн испускания оказывается существенно неравномерным (рис. 25). Основная часть соединений излучает в зелено-желто-оранжевой области: на промежуток 500-650 нм приходится 1023 записи из 1454, тогда как к обоим краям видимого диапазона число описанных соединений резко убывает. Распределение квантовых выходов смещено к малым значениям, причем большинство комплексов оказывается слабоизлучающими. Последнее обстоятельство показывает,

что подавление безызлучательной дезактивации остается нерешенной задачей конструирования излучателей на основе комплексов иридия(III).

Состав лигандного окружения отражает принятую в этой области стратегию поиска (рис. 25, Г и Д). Среди циклометаллирующих лигандов преобладают производные 2-фенилпиридина, причем с большим отрывом лидирует незамещенный 2-фенилпиридин, за которым следуют его фторированное производное и лиганды с расширенной системой сопряжения. Среди вспомогательных лигандов с еще большим отрывом преобладает ацетилацетонат, далее следуют пиколинат и 2,2'-бипиридин. Такое распределение отвечает обычному приему, при котором один тип лигандов оставляют неизменным и сравнительно простым, а второй подвергают систематическому изменению.

Что до последствий этих изменений, разбор набора показывает, что варьирование электронодонорной и электроноакцепторной способности заместителей, равно как и изменение размера сопряженной системы лигандов, позволяет смещать энергию триплетного состояния лишь в сравнительно узких пределах, задаваемых самим типом лиганда. Для существенного изменения энергии триплета требуются значительные структурные или электронные преобразования, причем и в этом случае высокий квантовый выход не гарантирован. Это наблюдение важно для оценки применимости расчетных средств отбора: оно означает, что перебор в пределах одного лигандного семейства дает ограниченный выигрыш, а поиск следует вести по возможно более широкому структурному пространству.

Наконец, две особенности характеризуют условия, в которых свойства измеряют. Катионные и нейтральные комплексы представлены в наборе сопоставимо, 638 и 801 запись соответственно, что указывает на устойчивый интерес к обоим типам соединений. Растворителем измерения в 59 % записей служит дихлорметан, поскольку он удобен в обращении и обеспечивает достаточную растворимость большинства бис-циклометаллированных комплексов иридия(III). Это преобладание и определило выбор подмножества для обучения моделей, описанный ниже.

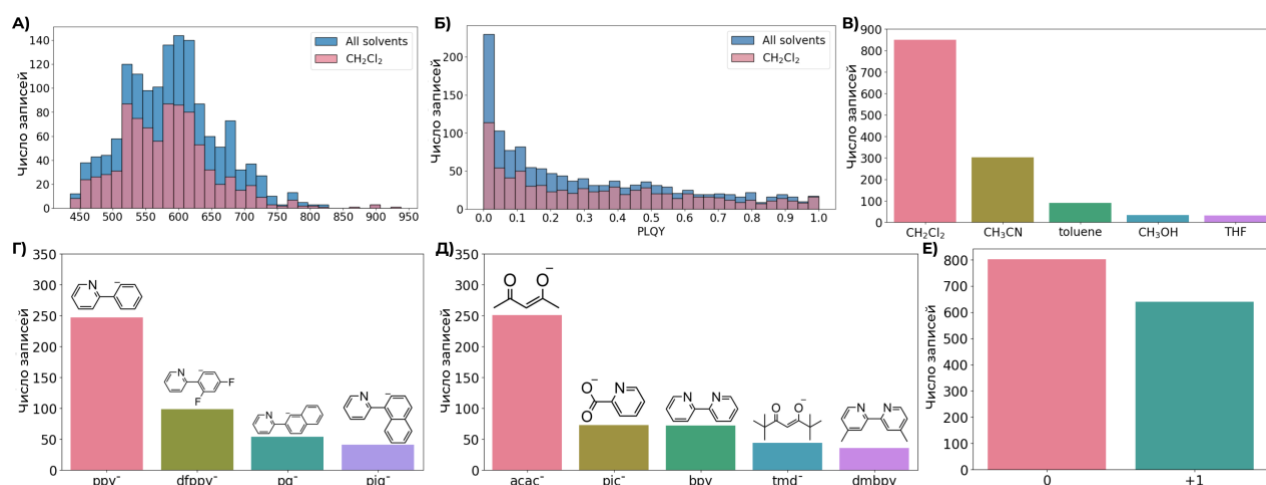


Рисунок 25. Статистика базы данных IrLumDB. А - распределение максимумов испускания. Б - распределение квантовых выходов. В - распределение по растворителям. Г - наиболее употребительные циклометаллирующие лиганды. Д - наиболее употребительные вспомогательные лиганды. Е - распределение комплексов по суммарному заряду

4.2.2. Модели прогнозирования длины волны максимума испускания и квантового выхода

Способ представления комплекса, изложенный в разделе 3.2.2, заимствован из работы по прогнозированию антибактериальной активности комплексов рутения[167] и распространен на существенно более разнообразное пространство бис-циклометаллированных комплексов иридия(III) (рис. 26). Тем самым описание строится из тех же элементов, из которых химик мысленно собирает комплекс, а сведения о координационной связи в явном виде не требуются.

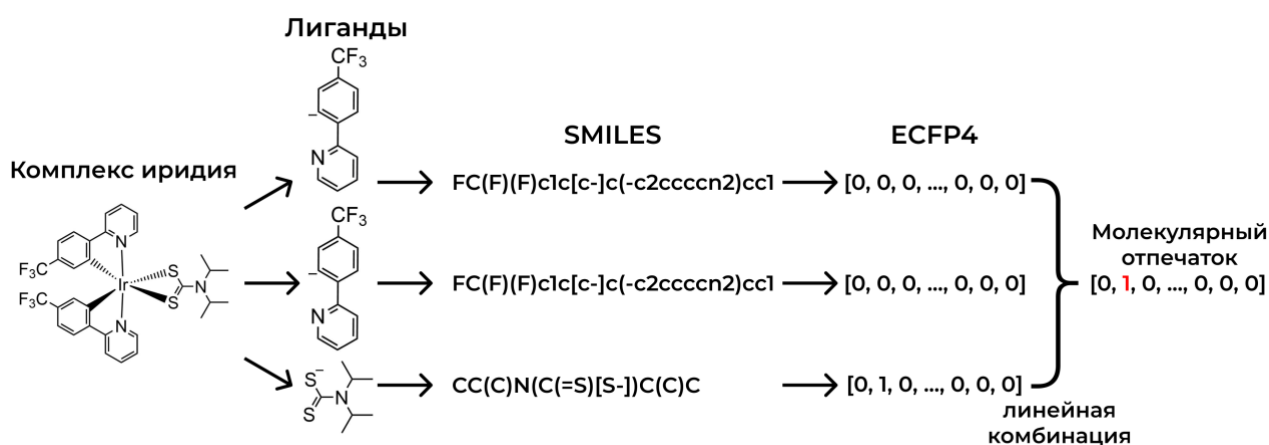


Рисунок 26. Построение молекулярного отпечатка комплекса иридия(III) как линейной комбинации отпечатков ECFP4 его лигандов

Отбор однородных подмножеств для обучения проводился по правилам, изложенным в разделе 3.2.2. Для длины волны максимума испускания он оставил 818 записей, а для квантового выхода, где отбор был строже, - 748 записей.

Испытывались пять алгоритмов, перечисленных в разделе 3.3.1. Качество оценивалось десятикратной перекрестной проверкой.

Результаты прогнозирования длины волны максимума испускания приведены в табл. 7. Три реализации градиентного бустинга дают близкие между собой величины и заметно превосходят метод ближайших соседей и метод опорных векторов: средняя абсолютная ошибка при перекрестной проверке составляет 18.3-18.8 нм против 25.3 и 21.5 нм соответственно, а коэффициент детерминации достигает 0.84-0.85. На экспериментальной проверочной выборке лучший результат показал CatBoost со средней абсолютной ошибкой 16.9 нм. Величина порядка 18 нм отвечает примерно трем процентам от самой длины волны и достаточна для того, чтобы отнести соединение к нужной области спектра, что и составляет практическую задачу при отборе излучателей.

Таблица 7. Качество прогнозирования длины волны максимума испускания при десятикратной перекрестной проверке и на экспериментальной проверочной выборке

Алгоритм	Перекрестная проверка			Экспериментальная проверочная выборка		
	MAE, нм	RMSE, нм	R ²	MAE, нм	RMSE, нм	R ²
kNN	25.3 ± 2.1	39.2 ± 4.8	0.69 ± 0.06	29.3	36.6	0.18
SVR	21.5 ± 1.5	31.5 ± 2.3	0.80 ± 0.04	21.0	25.8	0.59
LightGBM	18.3 ± 2.0	27.3 ± 3.3	0.84 ± 0.06	18.5	21.6	0.72
CatBoost	18.8 ± 1.5	27.4 ± 2.4	0.84 ± 0.05	16.9	20.3	0.75
XGBoost	18.4 ± 1.4	26.8 ± 1.8	0.85 ± 0.04	19.6	22.5	0.69

Прямое сопоставление достигнутой точности с работами, рассмотренными в разделе 2.4.2, затруднено и требует оговорок. Модели, построенные на комбинаторно

полученном массиве, прогнозируют среднюю энергию испускания со средней абсолютной ошибкой от 0.021 до 0.072 эВ[135], однако относятся они к иной величине, выраженной в иных единицах, и получены на 1380 соединениях единственного структурного типа, охарактеризованных единообразно в одной лаборатории. То же относится и к составным моделям, обученным на 1046 соединениях с двумя одинаковыми циклометаллированными лигандами из того же массива[141]. Настоящая работа решает задачу существенно менее благоприятную: 1287 соединений разнообразного строения, извлеченных из 340 публикаций разных лабораторий, измеренных в разных условиях и разными приборами. Существеннее, однако, различие во входных данных. Наиболее точные из указанных моделей используют признаки, получаемые полуэмпирическим квантово-химическим расчетом, а значит, требуют трехмерной структуры и отдельного вычисления для каждого нового соединения. Авторы одной из этих работ прямо указывают, что исключение расчета электронной структуры из построения признаков необходимо для преодоления вычислительного затруднения[141]. Предложенный здесь подход обходится структурными формулами лигандов, вследствие чего оценка свойств гипотетического соединения не требует ни построения трехмерной структуры, ни квантово-химического расчета, ни соответствующей квалификации.

Прогнозирование квантового выхода оказалось задачей более трудной. Средняя абсолютная ошибка лучших моделей составила 0.13 при значениях самой величины от нуля до единицы (табл. 8). Причина видна из разбора набора: распределение квантовых выходов сильно смещено к малым значениям, а сама величина чувствительна к условиям измерения сильнее, чем положение максимума испускания. Достигнутый уровень, однако, следует признать удовлетворительным, если сопоставить его с тем, что дает та же задача для органических соединений. Модель, обученная на наборе данных о флуоресцентных красителях в растворах, прогнозирует квантовый выход со средней абсолютной ошибкой 0.13 и энергию испускания с ошибкой 0.080 эВ, что сопоставимо с точностью нестационарной теории функционала плотности[188]. Иными словами, для комплексов иридия(III) достигнута та же точность прогнозирования квантового выхода, что и для органических красителей, - при том что органических соединений в наборах данных на порядки больше, а представление их в

машиночитаемом виде не составляет затруднения. Вместе с тем на соединениях из литературы, не входивших в обучающую выборку, модель воспроизводит квантовый выход с точностью, достаточной для отбора: для пяти известных интенсивноизлучающих комплексов прогнозируемые величины отличаются от измеренных не более чем на 0.10 (рис. 27).

Таблица 8. Качество прогнозирования квантового выхода при десятикратной перекрестной проверке

Алгоритм	MAE	RMSE	R ²
kNN	0.14 ± 0.01	0.21 ± 0.01	0.44 ± 0.10
SVR	0.13 ± 0.01	0.18 ± 0.02	0.58 ± 0.07
LightGBM	0.13 ± 0.01	0.18 ± 0.01	0.57 ± 0.05
CatBoost	0.13 ± 0.01	0.18 ± 0.02	0.58 ± 0.04
XGBoost	0.13 ± 0.01	0.19 ± 0.01	0.56 ± 0.04

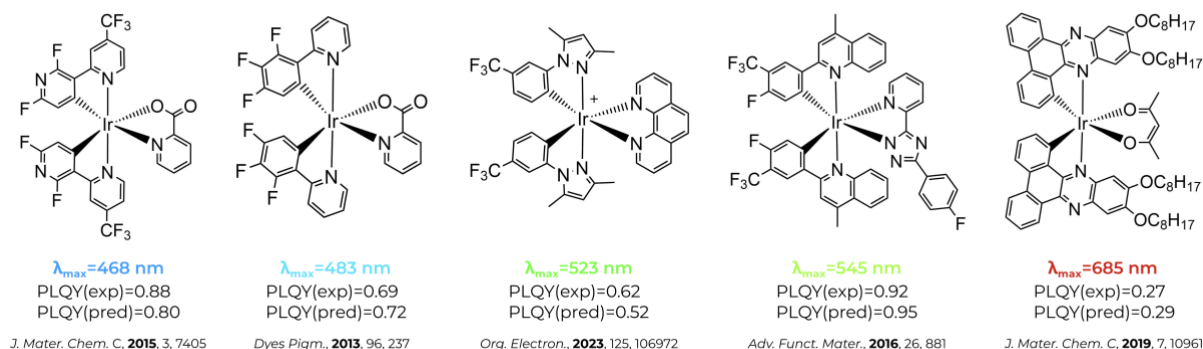


Рисунок 27. Проверка модели прогнозирования квантового выхода на сильноизлучающих комплексах иридия(III) из литературы, не входивших в обучающую выборку

Ввиду этого задача была поставлена также в виде классификации. Основанием послужило то, что для разных приложений требуются противоположные свойства: там, где нужно испускание света, желателен наибольший квантовый выход, а там, где требуется образование синглетного кислорода или выделение теплоты, напротив, полезны безызлучательные пути дезактивации. Соединения разделялись на три класса - с низким (от 0 до 0.1), умеренным (от 0.1 до 0.5) и высоким (от 0.5 до 1) квантовым выходом, каковое разделение отвечает и распределению самих данных: в

дихлорметане на эти классы приходится 201, 323 и 200 записей соответственно. Точность полученного классификатора составила 71.6, 74.9 и 69.0 % для низкого, умеренного и высокого классов при общей точности 72.4 % (табл. 9, рис. 28). Существенно, что ошибки приходятся преимущественно на смежные классы, то есть модель почти не путает слабоизлучающие соединения с интенсивноизлучающими, что для задачи предварительного отбора и является определяющим.

Таблица 9. Точность классификации комплексов по величине квантового выхода при десятикратной перекрестной проверке

Алгоритм	Точность, %
kNN	68.1 ± 4.1
SVC	70.3 ± 3.2
LightGBM	71.4 ± 2.3
CatBoost	72.4 ± 4.1
XGBoost	70.2 ± 3.3

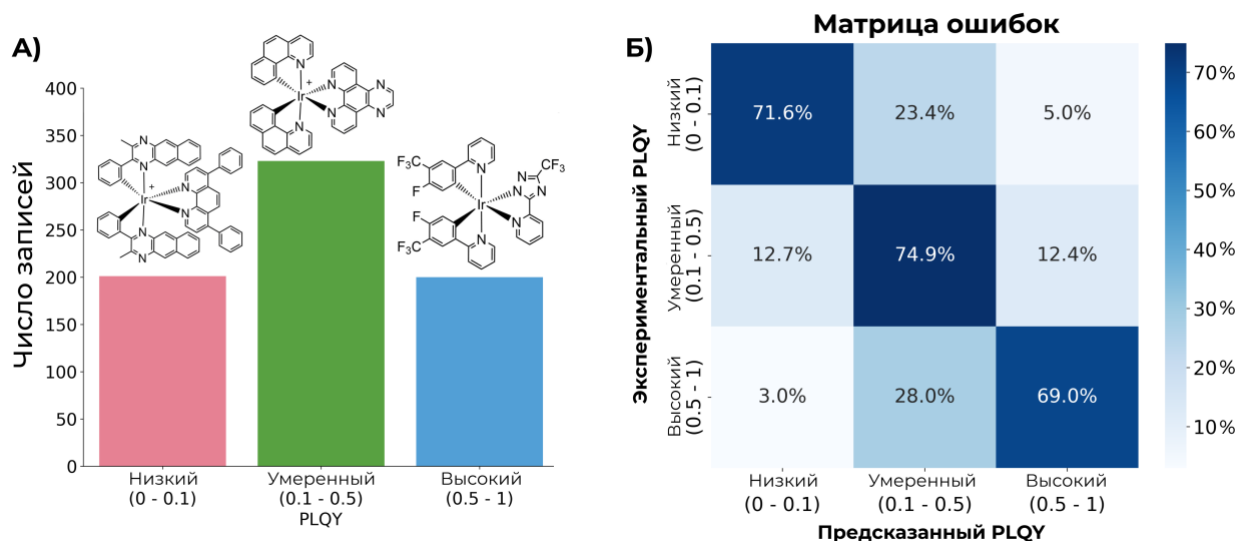


Рисунок 28. Классификация комплексов по величине квантового выхода. А - распределение записей по трем классам для измерений в дихлорметане и примеры соединений каждого класса. Б - матрица ошибок классификатора при десятикратной перекрестной проверке

Разбор построенных моделей методом SHAP показал, что модель опирается на химически осмысленные фрагменты: расширение сопряженной системы

циклометаллирующего лиганда смещает прогнозируемый максимум в длинноволновую область, тогда как введение атомов фтора и электроноакцепторных заместителей действует противоположным образом. Проверка на соединениях из литературы, не входивших в обучающую выборку, подтвердила применимость подхода и позволила проследить, какие именно структурные изменения модель связывает со смещением испускания (рис. 29).

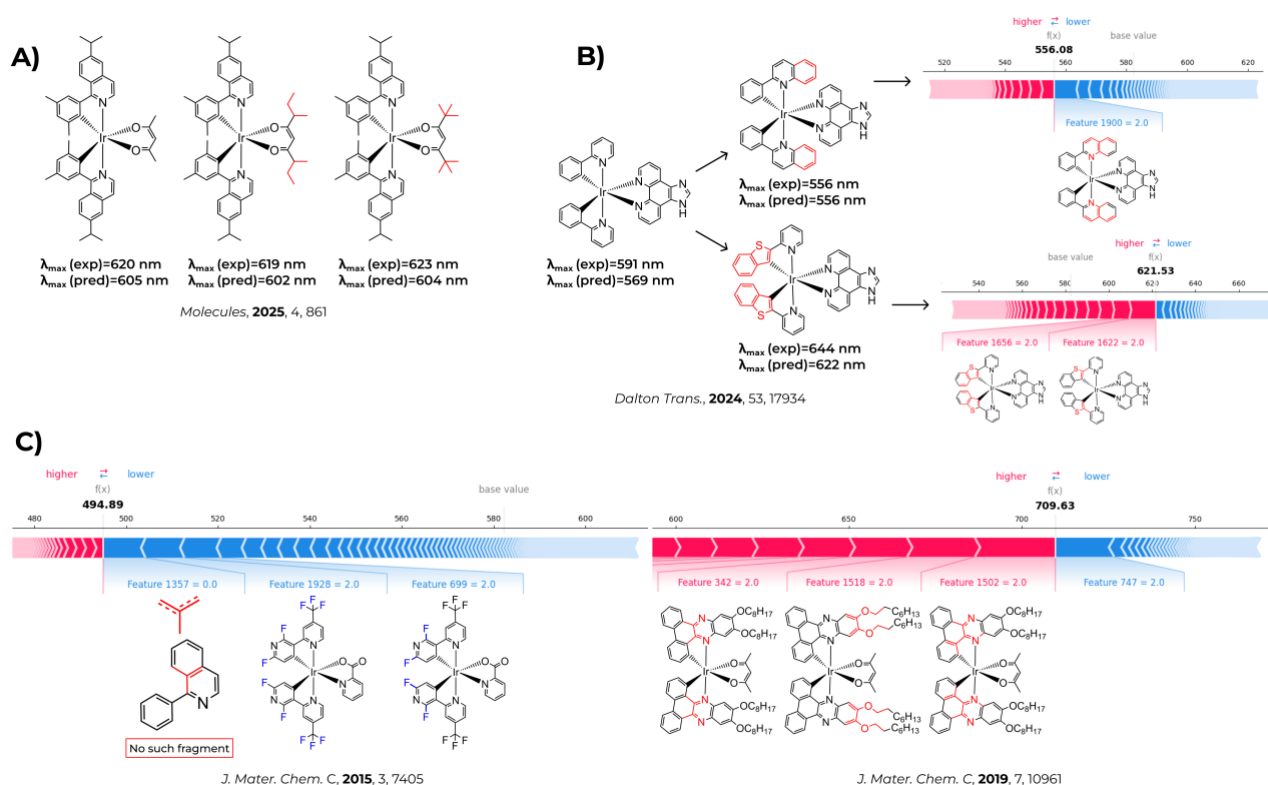


Рисунок 29. Проверка модели на соединениях из литературы, не входивших в обучающую выборку, и разбор вкладов отдельных структурных фрагментов методом SHAP. А, Б - прогнозируемые и измеренные длины волн максимума испускания. В - вклады признаков, понижающих и повышающих прогнозируемое значение

4.2.3. Экспериментальная проверка на синтезированных комплексах

Проверка модели на данных, полученных независимо от обучающей выборки, является наиболее строгим испытанием ее пригодности, и потому была проведена на специально синтезированных соединениях.

Двенадцать новых комплексов иридия(III) получены по методикам, изложенным в разделе 3.4.3, с выходами от 18 до 77 % (рис. 30). Циклометаллирующими лигандами

служили производные N-бензилбензимидазола, пять из которых также не были описаны ранее, а вспомогательными - ацетилацетонат и производные 2,2'-бипиридина. Состав и строение полученных соединений подтверждены спектроскопией ядерного магнитного резонанса, масс-спектрометрией высокого разрешения и рентгеноструктурным анализом, а однофазность образцов - порошковой рентгеновской дифракцией.

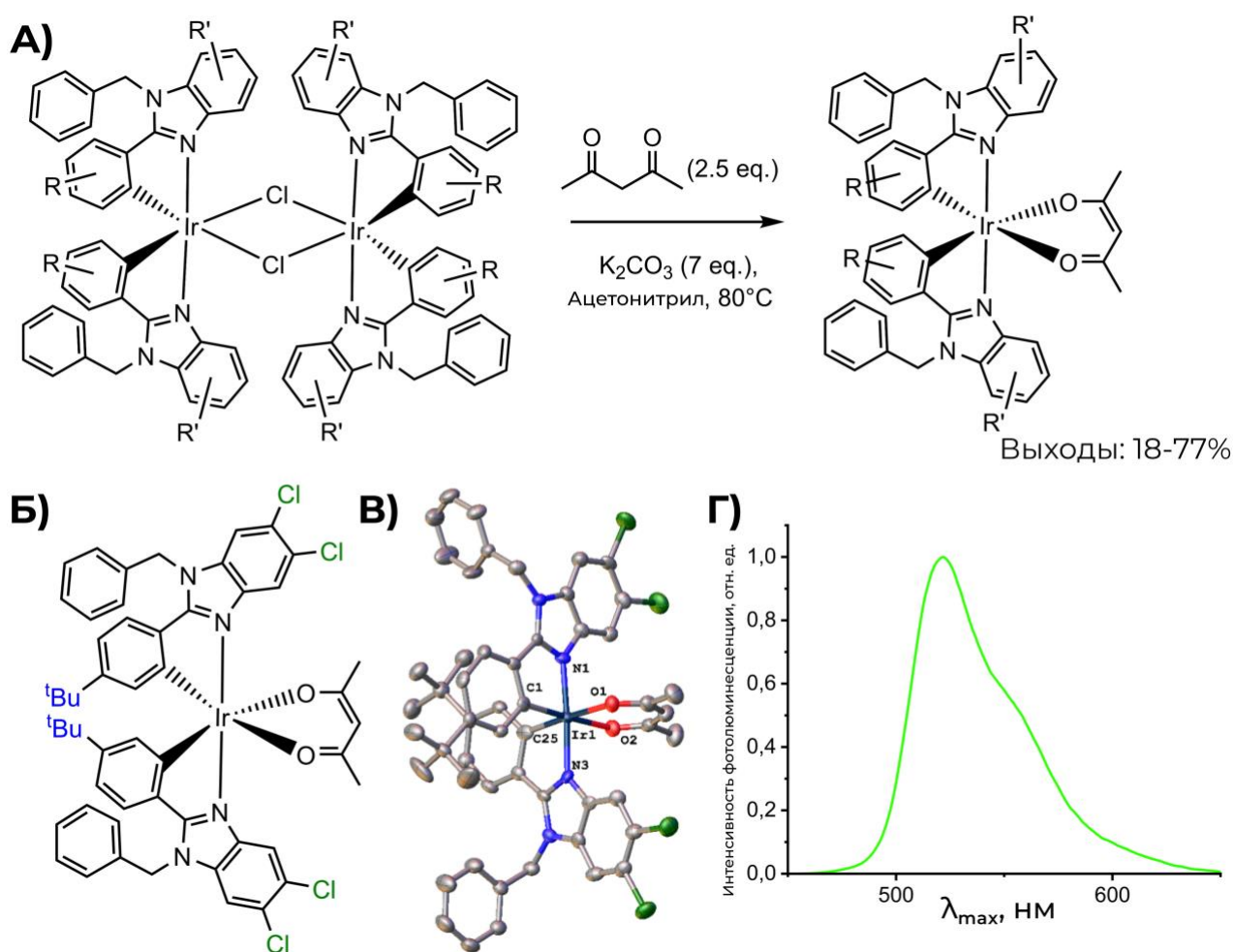


Рисунок 30. Синтез и характеристика полученных комплексов иридия(III). А - общая схема получения. Б - структурная формула одного из соединений. В - его строение по данным рентгеноструктурного анализа. Г - спектр люминесценции в дихлорметане

Проверочная выборка из 33 спектров люминесценции, собранная так, как изложено в разделе 3.4.4, охватывает область испускания от зеленой до красной части спектра, что позволило проверить модель по всей практически значимой области (рис. 31). Средняя абсолютная ошибка прогнозирования длины волны максимума

испускания составила 16.9 нм, то есть оказалась не хуже полученной при перекрестной проверке на литературных данных. Это обстоятельство существенно: обыкновенно качество моделей при переходе к независимым данным заметно падает, что и было показано в разделе 2.2.2 применительно к моделям растворимости. Сохранение точности означает, что построенное представление и обучающая выборка охватывают то химическое пространство, в котором ведется поиск новых излучателей.

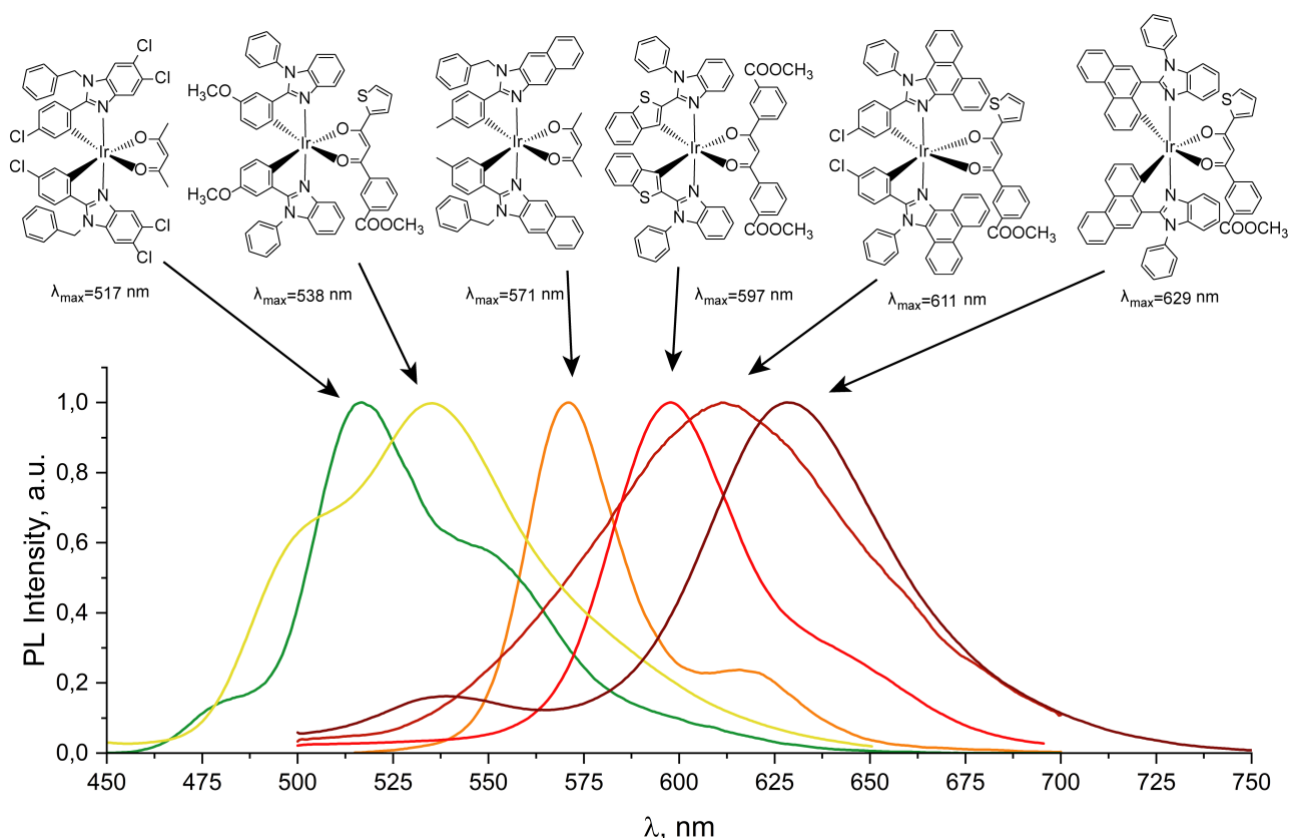


Рисунок 31. Спектры люминесценции полученных комплексов иридия(III) в дихлорметане и отвечающие им структурные формулы

Для тех же 33 соединений выполнен, кроме того, квантово-химический расчет длины волны испускания методом ΔSCF , изложенный в разделе 3.4.5. Сопоставление обоих подходов с опытом приведено на рис. 32. Расчет обнаруживает систематическое завышение длины волны испускания при сохранении общей зависимости, тогда как прогнозы моделей машинного обучения группируются вблизи прямой точного согласия. Систематические отклонения такого рода при вычислительном исследовании комплексов переходных металлов отмечаются и другими авторами.

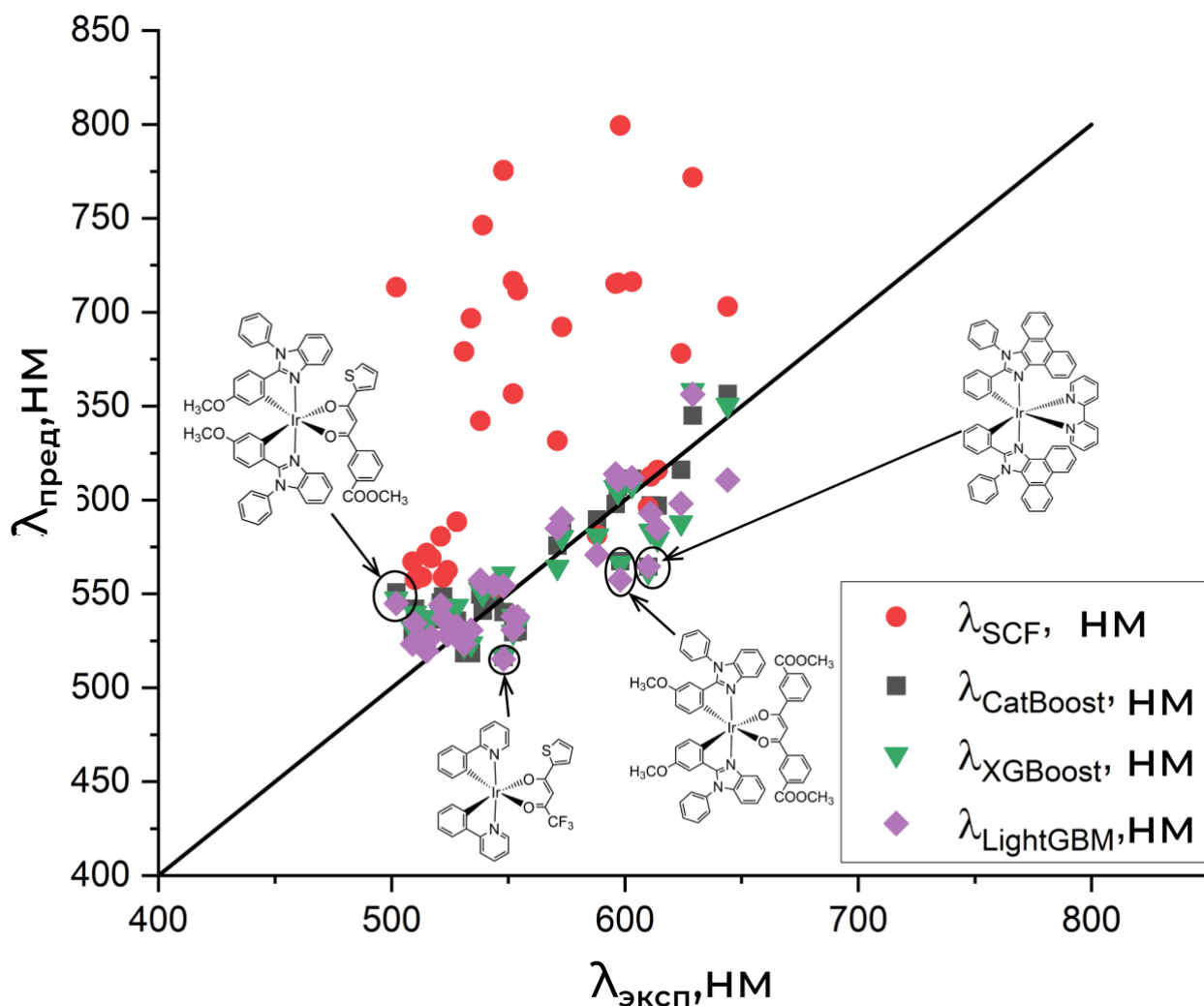


Рисунок 32. Сопоставление прогнозов моделей машинного обучения и квантово-химического расчета с измеренными длинами волн максимума испускания для 33 комплексов проверочной выборки. Прямая отвечает точному согласию. Отмечены соединения с наибольшей ошибкой прогнозирования

Причина расхождения поддается объяснению из устройства самих спектров. Некоторые из них содержат несколько полос, отвечающих различным триплетным подуровням, тогда как расчет относится к низшему триплетному состоянию, которое не обязательно дает наиболее интенсивную полосу. Модель же обучена непосредственно на положениях максимумов испускания и потому воспроизводит именно ту величину, которая измеряется.

Отсюда следует практический вывод, определяющий место построенных моделей среди существующих средств. Они не заменяют квантово-химический расчет там, где требуется описание электронного строения, но превосходят его в задаче быстрой и точной оценки положения максимума испускания, не требуя ни трехмерных структур, ни квалификации в квантовой химии, ни сколько-нибудь значительных вычислительных затрат. Именно это и делает их пригодными для предварительного отбора среди тысяч гипотетических соединений, о котором говорилось в разделе 2.4.2.

4.3. Цитотоксичность комплексов переходных металлов

4.3.1. Базы данных IrCytoToxDB и MetalCytoToxDB

Как показано в разделе 2.5.2, крупнейшие ресурсы сведений о биологической активности исключают металлосодержащие соединения из стандартизации, вследствие чего данные о цитотоксичности комплексов переходных металлов оказываются рассеяны по первичной литературе. Восполнение этого пробела проводилось в два приема: сначала для комплексов одного металла, а затем, по отработанной таким образом схеме, для пяти.

Первым был собран набор IrCytoToxDB, охватывающий комплексы иридия(III). Поиск, описанный в разделе 3.1.1, дал 6247 статей, из которых ручным просмотром выделены 222, содержащие сведения о цитотоксичности, за 2008-2022 годы (рис. 33). Существенно, что сужение запроса не годилось: по сочетанию «*iridium anticancer*» находится 136 работ, по «*iridium cytotoxicity*» - 22, тогда как ручной отбор из широкой выборки дал на порядок больше. Это обстоятельство подтверждает изложенное в разделе 2.1.2 о производительности разных путей сбора данных.

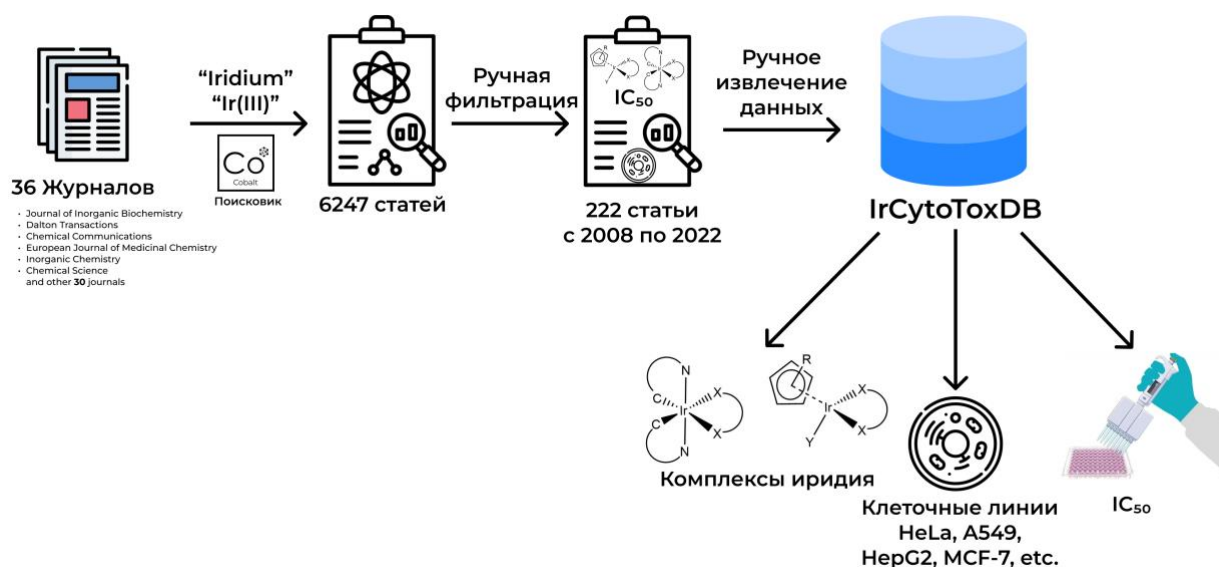


Рисунок 33. Схема сбора базы данных IrCytoToxDB

Способ представления комплекса был выбран тот же, что и для люминесцентных свойств, - разложением на лиганды, обозначаемые L1-L4. Применительно к цитотоксичности он потребовал обобщения, поскольку набор охватывает два существенно различных структурных типа. Для бис-циклометаллированных комплексов лигандам L1 и L2 отвечают циклометаллирующие лиганды, а L3 и L4 - вспомогательные. Для полусэндвичевых комплексов L1 отвечает лиганду на основе циклопентадиенила, L2 - бидентатному, L3 - монодентатному. Формально ковалентная связь с ионом иридия(III) обозначалась отрицательным зарядом на связывающем атоме.

Таблица 10. Журналы, просмотренные при сборе базы данных IrCytoToxDB, и число статей до и после отбора

Журнал	Найдено статей	Отобрано статей
Journal of Inorganic Biochemistry	46	26
Dalton Transactions	556	25
Chemical Communications	390	18
European Journal of Medicinal Chemistry	21	14
Inorganic Chemistry	913	14

Журнал	Найдено статей	Отобрано статей
Chemical Science	101	12
Journal of Biological Inorganic Chemistry	19	10
Angewandte Chemie International Edition	225	9
Dyes and Pigments	127	8
Journal of Organometallic Chemistry	689	8
Inorganica Chimica Acta	326	6
Journal of Medicinal Chemistry	10	6
Inorganic Chemistry Frontiers	32	6
Organometallics	790	6
Inorganic Chemistry Communications	71	5
Chemistry - A European Journal	150	5
New Journal of Chemistry	108	5
Zeitschrift für anorganische und allgemeine Chemie	76	4
Journal of the American Chemical Society	795	4
RSC Advances	122	4
Scientific Reports	35	3
Polyhedron	161	2
European Journal of Inorganic Chemistry	239	2
ACS Applied Materials and Interfaces	80	2
ChemBioChem	4	2
Biomaterials	17	2
ChemMedChem	5	2
Pharmaceutics	3	2

Журнал	Найдено статей	Отобрано статей
Chemico-Biological Interactions	4	2
Journal of Materials Chemistry B	22	2
Chemistry - An Asian Journal	55	1
ChemPlusChem	7	1
Applied Organometallic Chemistry	11	1
Molecules	33	1
ACS Medicinal Chemistry Letters	2	1
Bioorganic Chemistry	2	1
Всего	6247	222

Состав полей набора приведен в табл. 11. В окончательном виде IrCytoToxDB содержит 2694 измеренных значения концентрации полумаксимального ингибирования для 803 комплексов иридия(III) против 127 клеточных линий. Распределение записей по годам публикации обнаруживает неуклонный рост внимания к биологическим приложениям соединений иридия (рис. 34). Перечень просмотренных журналов с числом статей до и после отбора приведен в табл. 10. Среди клеточных линий выделяются две приблизительно равные по представленности, HeLa и A549, тогда как на десять наиболее употребительных приходится 70 % всех данных (рис. 35). Распределение самих величин оказывается двумодальным: соединения сосредоточены у крайне высокой цитотоксичности, ниже 10 мкмоль/л, и крайне низкой, выше 100 мкмоль/л (рис. 36), - обстоятельство, которое впоследствии определило постановку задачи прогнозирования как классификации.

Таблица 11. Описание полей базы данных IrCytoToxDB

Столбец	Описание	Тип
L1	строка SMILES лиганда L1, связанного с ионом иридия	строка

Столбец	Описание	Тип
L2	строка SMILES лиганда L2	строка
L3	строка SMILES лиганда L3	строка
L4	строка SMILES лиганда L4	строка
Counterion	строка SMILES противоиона, если комплекс заряжен	строка
Abbreviation_in_the_article	обозначение комплекса, принятое в статье-источнике	строка
IC50_Dark(M*10 ⁻⁶)	значение IC ₅₀ в темноте в том виде, в каком оно приведено в источнике	число или строка
IC50_Dark_standard_error(M*10 ⁻⁶)	стандартная ошибка IC ₅₀ в темноте	число
IC50_Light(M*10 ⁻⁶)	значение IC ₅₀ при облучении в том виде, в каком оно приведено в источнике	число или строка
IC50_Light_standard_error(M*10 ⁻⁶)	стандартная ошибка IC ₅₀ при облучении	число
Excitation_Wavelength(nm)	длина волны возбуждения, отвечающая значению IC ₅₀ при облучении	число или строка
Irradiation_Time(minutes)	время облучения, отвечающее значению IC ₅₀ при облучении	число
Irradiation_Power(W*m ⁻²)	плотность мощности источника света	число
Cell_line	обозначение клеточной линии	строка
Time(h)	время воздействия комплекса на клеточную линию	число
DOI	цифровой идентификатор публикации-источника	строка

Столбец	Описание	Тип
Year	год публикации источника	число
Comments	дополнительные замечания	строка

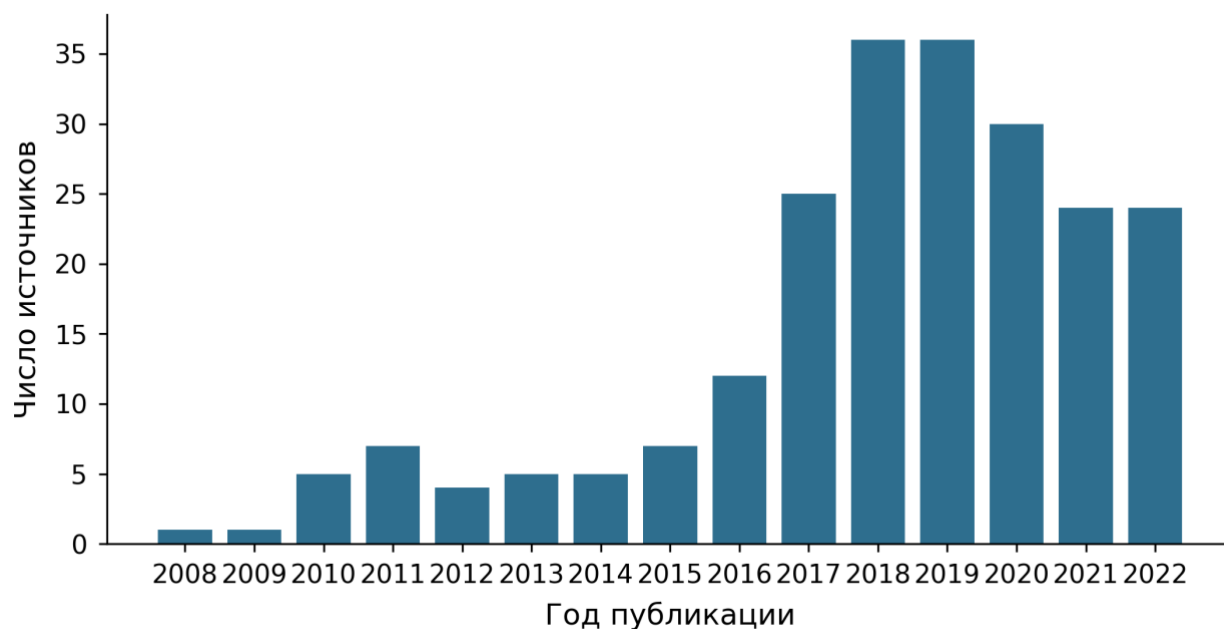


Рисунок 34. Распределение источников базы данных IrCytoToxDB по году публикации

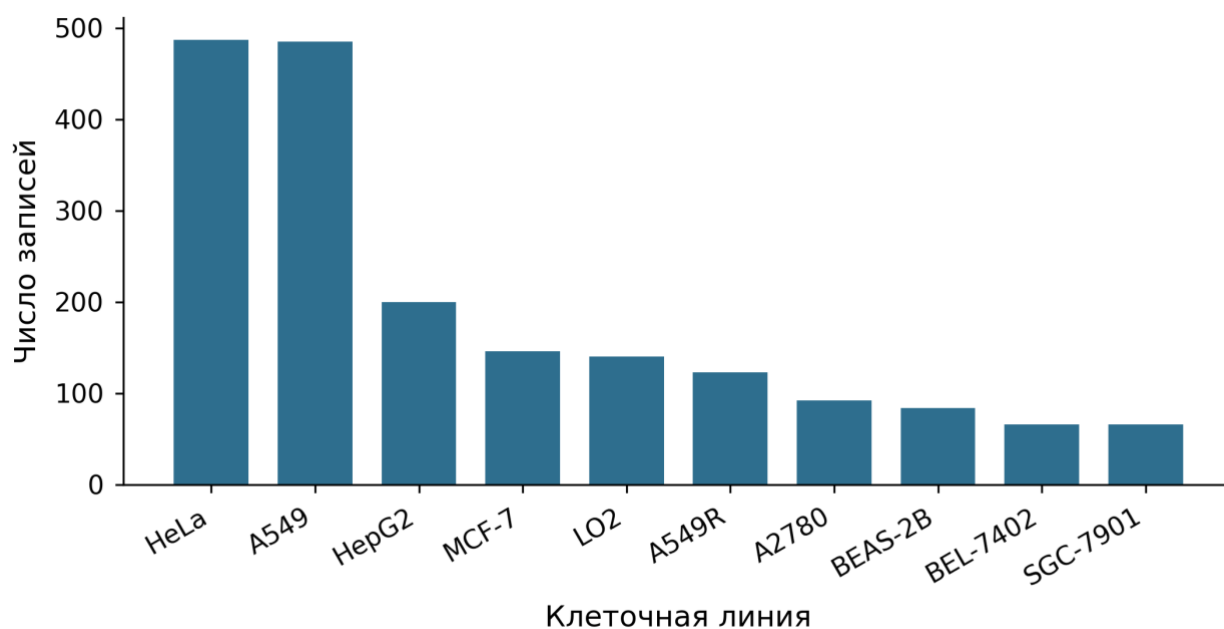


Рисунок 35. Наиболее употребительные клеточные линии в базе данных IrCytoToxDB

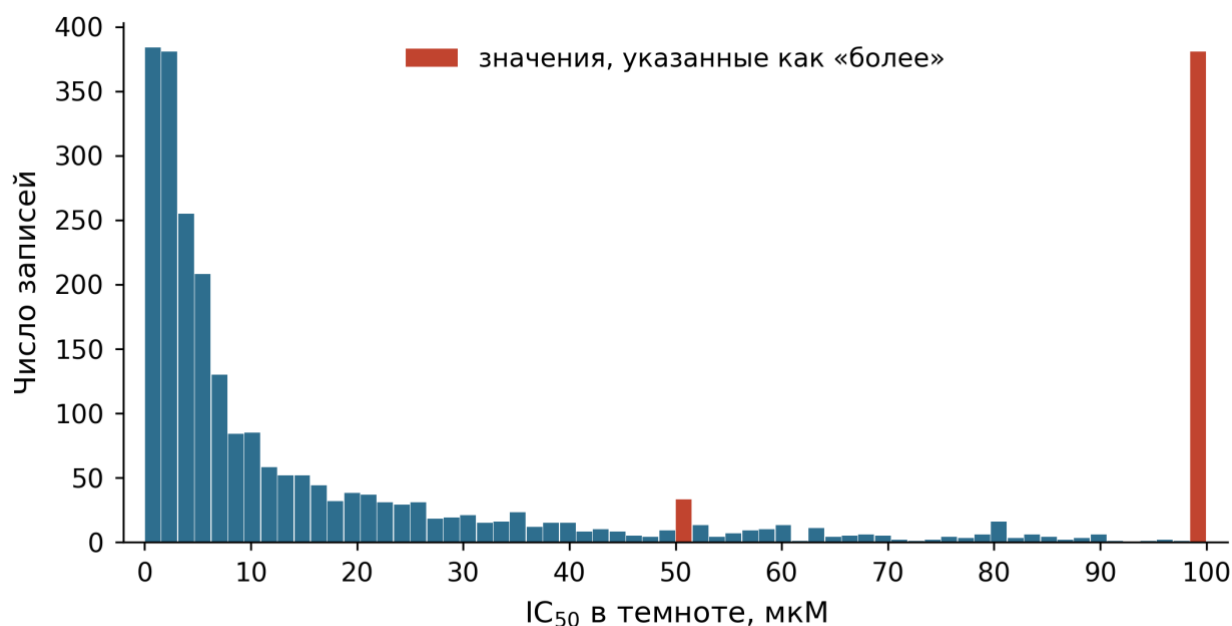


Рисунок 36. Распределение значений IC_{50} в темноте в базе данных IrCytoToxDB. Красным выделены столбцы, отвечающие значениям, приведенным в источниках как «более» порогового

Отработанная таким образом схема была затем распространена на пять металлов. Набор MetalCytoToxDB собирался по тому же порядку, но в два яруса отбора, описанных в разделе 3.1.1 (рис. 37). Первый ярус дал более 50 тысяч статей, второй сузил их до поддающегося ручному просмотру числа. Отобранные работы просматривались вручную, и в итоге выделена 1921 публикация, из которых извлечено 26500 значений концентрации полумаксимального ингибирования для 7050 комплексов против 754 клеточных линий.



Рисунок 37. Схема сбора базы данных MetalCytoToxDB

Распределение записей по металлам крайне неравномерно и отражает сложившиеся предпочтения исследователей. На рутений приходится около 19 тысяч записей, на иридий около 4.5 тысячи, тогда как родий, осмий и рений представлены примерно тысячей каждый (рис. 38). Эта неравномерность существенна: она означает, что для трех металлов из пяти построить самостоятельную модель невозможно, и именно она обусловила обращение к переносу сведений между подмножествами, описанному ниже. Распределение публикаций-источников по журналам приведено в табл. 12.

Таблица 12. Двадцать журналов, давших наибольшее число публикаций при сборе базы данных MetalCytoToxDB

Журнал	Число публикаций
Dalton Transactions	207
Journal of Inorganic Biochemistry	193
Inorganic Chemistry	152
Journal of Medicinal Chemistry	78
European Journal of Medicinal Chemistry	73
Journal of Organometallic Chemistry	64
Inorganica Chimica Acta	62
Organometallics	53
European Journal of Inorganic Chemistry	48
Chemical Communications	47
Journal of Biological Inorganic Chemistry	46
Chemistry - A European Journal	42
Polyhedron	42
Applied Organometallic Chemistry	39
New Journal of Chemistry	39
Angewandte Chemie	31
Molecules	30
Journal of Coordination Chemistry	27
Inorganic Chemistry Frontiers	27

Журнал	Число публикаций
Chemical Science	25
Прочие журналы	596

Прочие распределения набора устроены следующим образом. Среди клеточных линий преобладают A549, HeLa и MCF-7, причем представлены как опухолевые, так и неопухолевые линии, а также две линии, устойчивые к цисплатину. Время воздействия сосредоточено у 48 часов, за которыми следуют 72 и 24 часа. Распределение самих значений повторяет замеченное ранее для иридия: основная масса приходится на область малых концентраций, а на 50 и 100 мкмоль/л приходятся выраженные пики, отвечающие записям, приведенным в первоисточниках как предельные оценки вида «более 50» и «более 100 мкмоль/л». Цитотоксичность в отношении неопухолевых линий систематически ниже, чем в отношении опухолевых, что отвечает ожиданиям, но создает существенную несбалансированность классов.

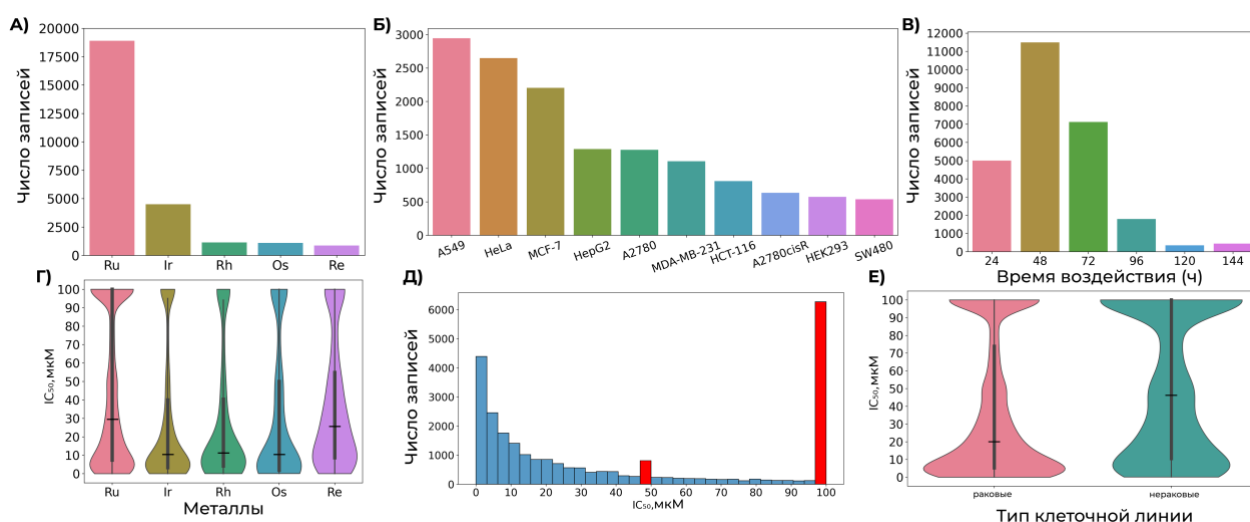


Рисунок 38. Статистика базы данных MetalCytoToxDB. А - распределение записей по металлам. Б - десять наиболее употребительных клеточных линий. В - время воздействия. Г - распределение значений IC_{50} по металлам. Д - распределение значений IC_{50} , красным отмечены записи, приведенные в первоисточниках как предельные оценки. Е - сопоставление опухолевых и неопухолевых клеточных линий

Структурный состав набора также неоднороден. Среди комплексов рутения преобладают полусэндвичевые соединения с ареновым лигандом, которых около 1750, за ними следуют полипиридилные комплексы с 2,2'-бипиридином и 1,10-

фенантролином, около 500 и 470 соответственно, и замыкают ряд соединения с трифенилфосфином, которых около 150. Среди комплексов иридия сопоставимо представлены бис-циклометаллированные и полусэндвичевые соединения, около 500 и 390 (рис. 39). Для трех остальных металлов преобладающий структурный тип оказывается единственным: у осмия и родия это полусэндвичевые соединения с ареновым и циклопентадиенильным лигандом, которых около 180 и 120, соответственно, у рения - трикарбонильные комплексы с бидентатным лигандом и галогенидом, которых около 70 (рис. 40). Эта однородность имеет двойное следствие. С одной стороны, она облегчает перенос сведений между металлами, поскольку строение комплексов сходно. С другой - означает, что за пределами преобладающего типа обучающих данных для этих металлов нет вовсе. Это соотношение важно для оценки применимости моделей: оно показывает, в каких областях структурного пространства обучающих данных достаточно, а в каких прогноз следует принимать с осторожностью. Однороден набор и по степени окисления металла: комплексы рутения представлены почти исключительно соединениями в степени окисления +2, которых около 17.5 тысячи записей против 1.3 тысячи для степени +3, а комплексы иридия - исключительно соединениями иридия(III).

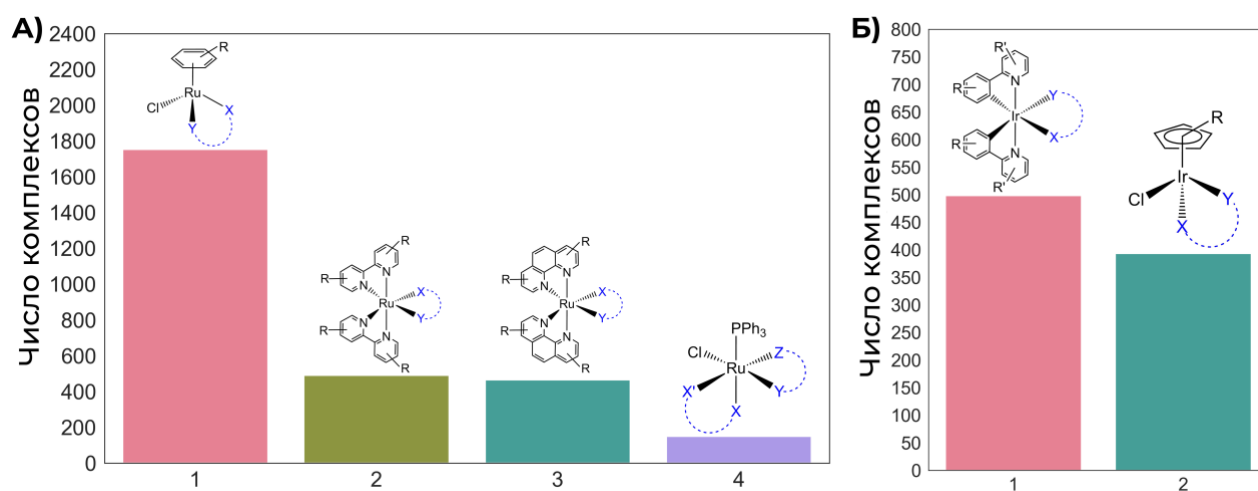


Рисунок 39. Структурные типы комплексов в базе данных MetalCytoToxDB. А - комплексы рутения. Б - комплексы иридия

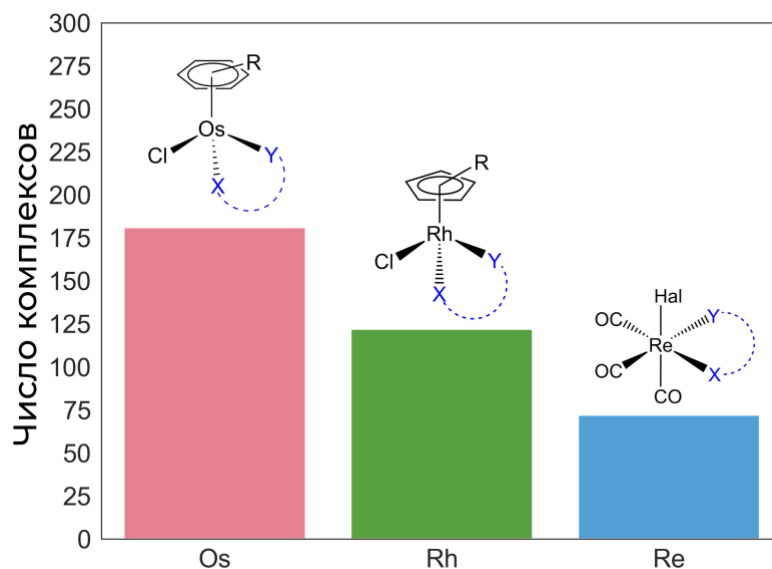


Рисунок 40. Преобладающие структурные типы комплексов осмия, родия и рения в базе данных MetalCytoToxDB

Помимо самой цитотоксичности набор позволяет оценить и избирательность действия. Для комплексов рутения, у которых в одной и той же публикации имеются измерения и на опухолевых, и на неопухолевых линиях, вычислен индекс селективности, определенный в разделе 2.5.1 как отношение концентраций полумаксимального ингибирования. Единицей счета служило сочетание соединения и публикации: для каждого из них бралось срединное значение концентрации по всем неопухолевым линиям работы и такое же по всем опухолевым, после чего вычислялось их отношение. Время воздействия при этом не согласовывалось, вследствие чего сопоставляемые величины могут относиться к разной его продолжительности. Всего таких сочетаний соединения и публикации оказалось 1779. Отношение, однако, вычислимо не для всех: в 349 случаях предельной оценкой вида «более 100 мкмоль/л» задана величина для обеих линий, и тогда отношение не определено вовсе, поскольку двум нижним границам отвечает любое значение — как 0.1, так и 10. Еще в 179 случаях предельной оценкой задана одна лишь неопухолевая линия, а в 94 — только опухолевая; такие записи дают лишь одностороннюю границу индекса. Подстановка предельных оценок как чисел обращала бы отношение в единицу по одному только способу записи и создавала бы у нуля искусственный пик, поэтому распределение построено по 1157 сочетаниям, для которых обе величины измерены (рис. 41). Оно

имеет максимум при значении индекса около двух и растянутый правый край, простирающийся до полутора логарифмических единиц: избирательность, превышающая единицу, обнаруживается у трех четвертей таких сочетаний, но различающей опухолевые и неопухолевые клетки в десятки раз обладает лишь малая их часть. Существенно и другое: приведенные величины вычислены для различных сочетаний опухолевой и неопухолевой линий, вследствие чего строгое сопоставление их между собой невозможно.

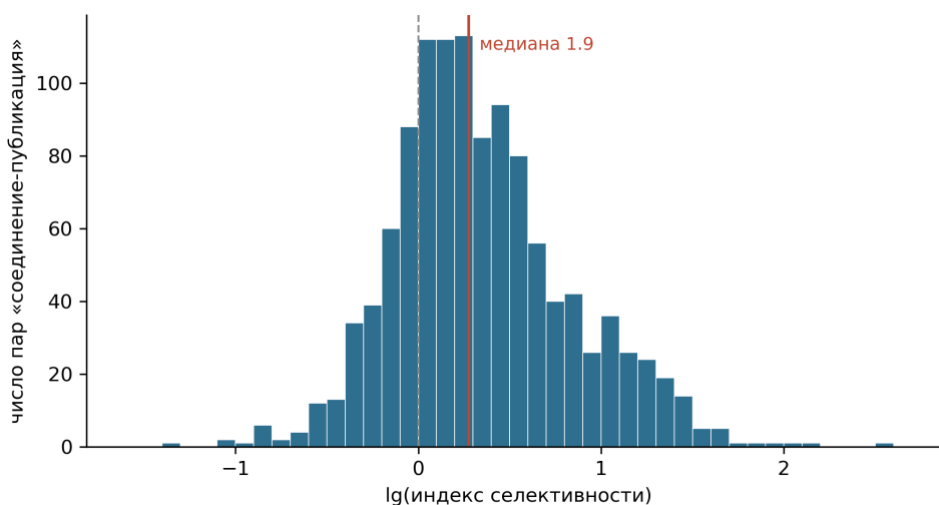


Рисунок 41. Распределение десятичного логарифма индекса селективности для комплексов рутения в базе данных MetalCytoToxDB. Построено по 1157 сочетаниям соединения и публикации, для которых обе концентрации полумаксимального ингибирования измерены; сочетания, где хотя бы одна из них задана предельной оценкой, не включены. Штриховой чертой показано значение индекса, равное единице, сплошной - срединное значение

Собранные данные позволяют, кроме того, оценить воспроизводимость самих измерений цитотоксичности - задачу, для решения которой обыкновенно недостает материала. Эта возможность возникает благодаря тому, что в подавляющем большинстве работ по металлосодержащим противоопухолевым средствам цисплатин измеряют одновременно с исследуемыми соединениями в качестве средства сравнения, и значения его извлекались наравне с прочими. В наборе содержится 14309 таких

значений из 978 публикаций для 567 клеточных линий, что превращает цисплатин во внутренний эталон, измеренный сотнями независимых групп.

Результат оказывается обескураживающим (рис. 42). Для линии A549, представленной 1547 значениями из 382 публикаций, срединная величина pIC_{50} составляет 4.84, что отвечает концентрации полумаксимального ингибирования 14.4 мкмоль/л, тогда как крайние значения охватывают промежуток от 3.90 до 6.85, то есть без малого три порядка по концентрации. Для HeLa при 1489 значениях из 351 публикации крайние величины pIC_{50} составляют 4.15 и 6.30, для MCF-7 при 1260 значениях из 325 публикаций - 4.00 и 6.28, что отвечает размаху в 143 и 187 раз соответственно. Даже межквартильный размах, отсекающий крайние четверти, охватывает для этих линий промежуток примерно от 7 до 21 мкмоль/л, то есть трехкратный.

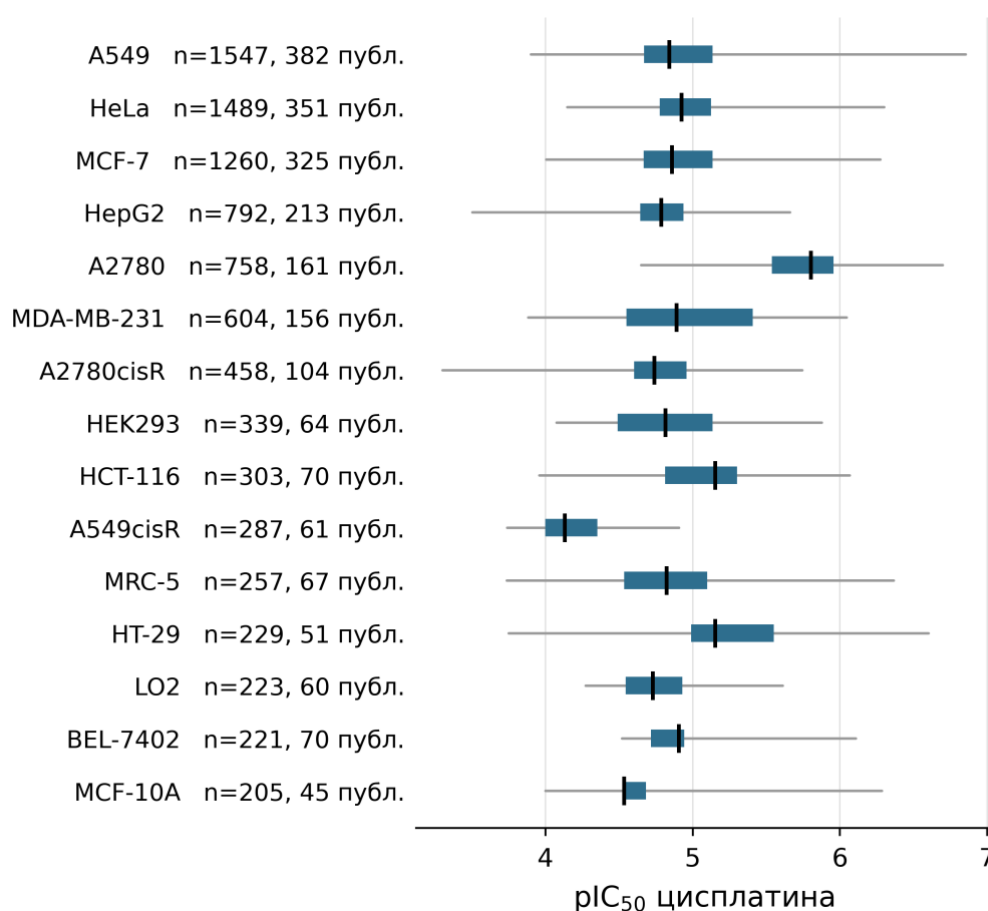


Рисунок 42. Разброс величин pIC_{50} цисплатина, измеренных разными группами на одних и тех же клеточных линиях. Серой чертой показан полный размах, столбцом -

межквартильный, вертикальной чертой - срединное значение. Величина pIC_{50} определена как отрицательный десятичный логарифм концентрации полумаксимального ингибирования в микромолях на литр, увеличенный на шесть

Это наблюдение не является особенностью настоящего набора. Расхождение величин цитотоксичности цисплатина, измеренных разными группами на одних и тех же линиях, отмечалось и ранее: разбор 41 работы, охвативший 56 опытов, выполненных с 2018 по 2022 год, обнаружил чрезвычайно широкие доверительные промежутки, причем неоднородность результатов превышала 99.7 % и объяснялась, следовательно, различиями в постановке опыта, а не случайностью[189]. Собранный в настоящей работе материал превышает объем указанного разбора более чем на два порядка и подтверждает его вывод на несопоставимо большей выборке.

Следствия отсюда двоякие. С одной стороны, они объясняют изложенное выше: почему качество моделей не связано с числом записей для отдельной клеточной линии и почему при таком разбросе разумнее решать задачу классификации, а не регрессии. С другой - они очерчивают предел достижимого. Пока цитотоксичность одного и того же вещества на одной и той же линии различается в разных работах на два порядка, ожидать от расчетных средств точности, превышающей воспроизводимость самого эксперимента, не приходится. Это положение в точности повторяет разобранный в разделе 2.2.2 применительно к растворимости, где предел точности моделей также определился воспроизводимостью измерений, а не устройством самих моделей.

Отдельного разбора заслуживает та часть набора, которая относится к фотодинамической терапии. Наряду с обычной цитотоксичностью для 2158 записей из 268 публикаций, охватывающих 757 соединений, приведены значения концентрации полумаксимального ингибирования, измеренные и в темноте, и при облучении, а также условия облучения: длина волны возбуждения, время и плотность мощности. Это подмножество составляет около восьми процентов набора, причем доля его неуклонно растет - с четырех процентов записей в 2014 году до семнадцати в 2024.

Отношение двух измеренных величин дает индекс фототоксичности, показывающий, во сколько раз облучение усиливает действие соединения.

Распределение его по металлам приведено на рис. 43. Наибольшие значения дают комплексы иридия со срединным усилением в 20 раз, за ними следуют осмий и рутений. У половины всех записей облучение усиливает действие более чем вдесятеро, у пятнадцати процентов - более чем в сто раз, причем наибольшее наблюдавшееся усиление превышает три порядка. Вместе с тем у 11 % записей индекс не превосходит единицы, то есть облучение не дает никакого выигрыша, и такие соединения фотосенсибилизаторами по существу не являются.

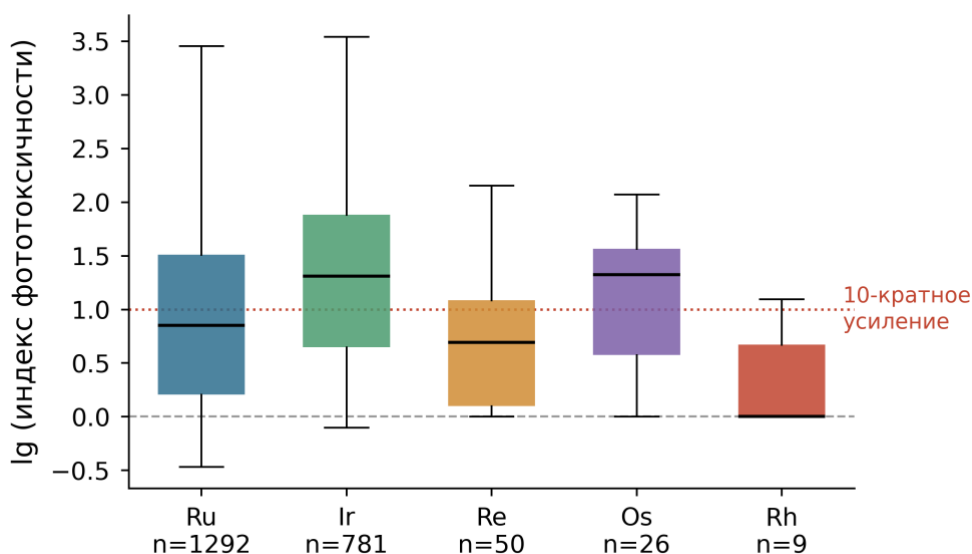


Рисунок 43. Индекс фототоксичности комплексов переходных металлов, вычисленный как отношение концентрации полумаксимального ингибирования в темноте к таковой при облучении. Пунктиром отмечены отсутствие выигрыша от облучения и десятикратное усиление

Существенное, однако, оказывается распределение по длине волны возбуждения (рис. 44). Срединная длина волны составляет 460 нм, почти три четверти записей приходится на область короче 500 нм, и лишь 11 % попадают в терапевтическое окно 600-850 нм, где поглощение тканями наименьшее. Между тем именно проникновение света вглубь ткани и составляет главное затруднение фотодинамической терапии: синий свет проникает на доли миллиметра. Область, таким образом, изучает преимущественно соединения, возбуждаемые светом, который до опухоли не дойдет. Это наблюдение объясняет, почему в последние годы появились работы, целенаправленно посвященные поиску соединений с поглощением в терапевтическом

окне, о которых говорилось в разделе 2.4.2. Существенно при этом, что соединения, возбуждаемые в окне, по действенности прочим не уступают: срединный индекс фототоксичности для них составляет 19 против 8 для области 450-500 нм, то есть это ограничение историческое.

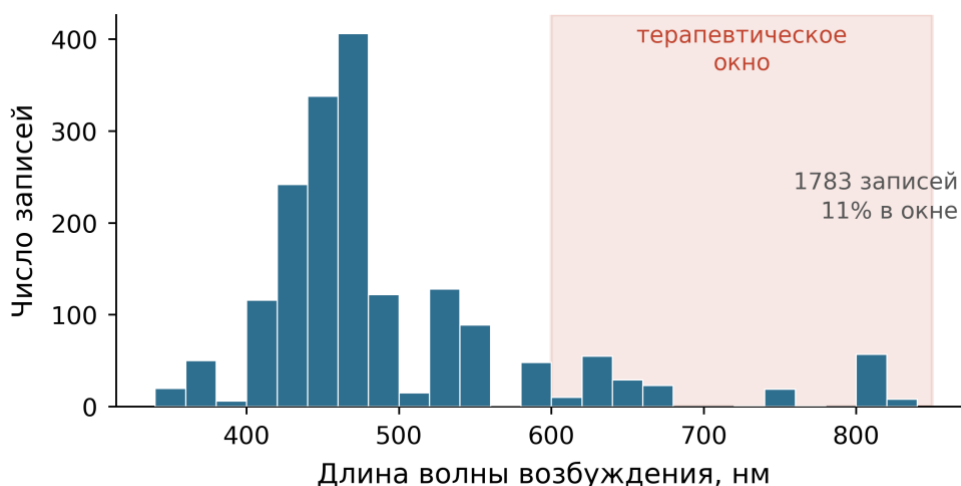


Рисунок 44. Распределение записей по длине волны возбуждения. Затенена область терапевтического окна 600-850 нм

Наконец, следует отметить обстоятельство, важное для оценки качества данных. Условия облучения сообщаются в первоисточниках существенно полнее, чем условия эксперимента в работах по растворимости: длина волны приведена в 96 % записей, время облучения в 95 %, плотность мощности в 87 %. Это сопоставление показательно. Требования к описанию опыта складываются вместе с самой областью, и в молодой области, где значимость условий очевидна каждому исследователю, они оказываются строже, чем в области с вековой историей измерений.

Устройство расширенного набора отличается от устройства предшественника в трех отношениях, и эти различия вытекают из расширения охвата. Прежде всего, лиганды перестали записываться по отдельным столбцам L1-L4 и объединены в один столбец в виде несвязной молекулы. Причина в том, что при переходе от одного металла к пяти число лигандов у комплекса перестало быть предсказуемым: оно меняется от двух до шести, и заранее отведенного числа столбцов не хватило бы для одних соединений, оставаясь пустым для других. Тот же довод определил впоследствии и способ вычисления признаков, о котором говорится ниже. Далее,

наряду с величинами в том виде, в каком они приведены в источниках, введены отдельные столбцы с приведенными к единому виду числовыми значениями, что позволяет отличать записанное от вычисленного и сохранять предельные оценки вида «более 100 мкмоль/л», не смешивая их с измеренными числами. Наконец, добавлены значение IC₅₀ цисплатина, измеренное в той же работе и служащее средством сравнения между статьями, и признаки металлического центра - символ, степень окисления, порядковый номер и число валентных электронов, - без которых перенос сведений между металлами был бы невозможен.

Состав полей расширенного набора приведен в табл. 13. Оба набора депонированы в общедоступном репозитории под постоянными идентификаторами (табл. 15). MetalCytoToxDB превосходит предшественника по объему примерно в десять раз и является, насколько нам известно, крупнейшим открытым собранием сведений о цитотоксичности комплексов переходных металлов.

Таблица 13. Описание полей базы данных MetalCytoToxDB

Столбец	Описание	Тип
SMILES_Ligands	строки SMILES всех лигандов комплекса, записанные как одна несвязная молекула	строка
Counterion	строка SMILES противоиона, если комплекс заряжен	строка
Abbreviation_in_the_article	обозначение комплекса, принятое в статье-источнике	строка
IC50_Dark(M*10 ⁻⁶)	значение IC ₅₀ в темноте в том виде, в каком оно приведено в источнике	строка
IC50_Dark_standard_error(M*10 ⁻⁶)	стандартная ошибка IC ₅₀ в темноте	число
IC50_Dark_value	числовое значение IC ₅₀ в темноте, приведенное к единому виду	число

Столбец	Описание	Тип
IC50_Light($M \cdot 10^{-6}$)	значение IC ₅₀ при облучении в том виде, в каком оно приведено в источнике	строка
IC50_Light_standard_error($M \cdot 10^{-6}$)	стандартная ошибка IC ₅₀ при облучении	число
IC50_Light_value	числовое значение IC ₅₀ при облучении, приведенное к единому виду	число
Excitation_Wavelength(nm)	длина волны возбуждения	число
Irradiation_Time(minutes)	время облучения	число
Irradiation_Power($W \cdot m^{-2}$)	плотность мощности источника света	число
Cell_line	обозначение клеточной линии	строка
Time(h)	время воздействия комплекса на клеточную линию	число
DOI	цифровой идентификатор публикации-источника	строка
Year	год публикации источника	число
IC50_Cisplatin($M \cdot 10^{-6}$)	значение IC ₅₀ цисплатина, измеренное в той же работе как средство сравнения	строка
IC50_Cisplatin_standard_error($M \cdot 10^{-6}$)	стандартная ошибка IC ₅₀ цисплатина	число
IC50_Cisplatin_value	числовое значение IC ₅₀ цисплатина, приведенное к единому виду	число
Charge_complex	суммарный заряд комплексного иона	число
Metal	символ металла	строка
Oxidation_state	степень окисления металла	число

Столбец	Описание	Тип
Atomic_number	порядковый номер металла	число
Valence_e	число валентных электронов металла	число

4.3.2. Модели классификации по уровню цитотоксичности

Постановка задачи определялась устройством самих данных. Признаковое описание комплекса строилось из отпечатков его лигандов, к которым добавлялись признаки металлического центра (рис. 45). От регрессии пришлось отказаться по двум причинам. Во-первых, значительная часть значений приведена в первоисточниках не числом, а предельной оценкой вида «более 100 мкмоль/л», и обучение на таких записях воспроизводило бы условное обозначение отсутствия величины. Во-вторых, для практического отбора соединений точная величина менее важна, чем отнесение к одному из двух классов. Соединения разделялись поэтому на цитотоксичные, с концентрацией полумаксимального ингибирования ниже 10 мкмоль/л, и нецитотоксичные, с концентрацией не ниже этой величины. Порог 10 мкмоль/л не является всеобщим, но представляет собой разумное соглашение между биологической значимостью и распределением данных между классами[166,190], и он же принят во многих хемоинформатических исследованиях[191–193].

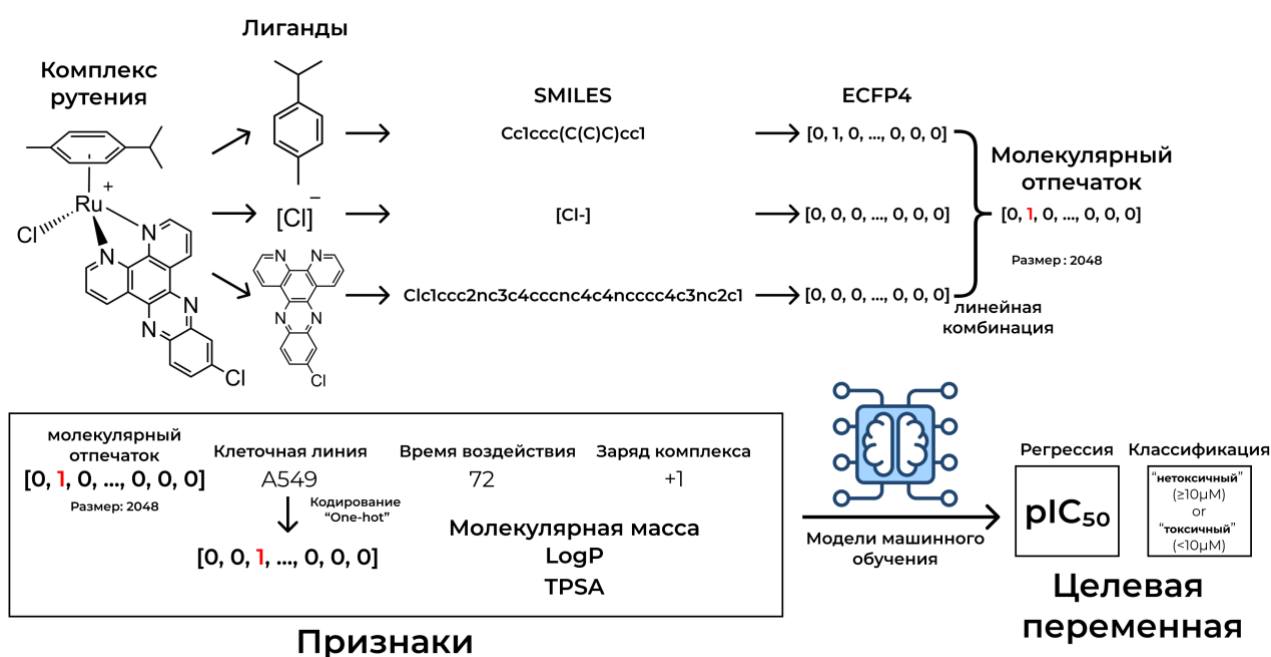


Рисунок 45. Построение признакового описания комплекса: отпечатки лигандов и признаки металлического центра

Неопухолевые клеточные линии были исключены из обучения. Основанием послужило то, что цитотоксичность в отношении них систематически ниже, отчего классы оказываются несбалансированными настолько, что модель вырождается в отнесение всех соединений к преобладающему классу. Целевой величиной служил показатель pIC_{50} , определенный в разделе 3.1.3.

Обучающая выборка после удаления дубликатов, отвечающих одному и тому же сочетанию соединения, клеточной линии и времени воздействия в разных источниках, составила 11294 записи для 4399 комплексов рутения против 23 опухолевых клеточных линий (рис. 46).

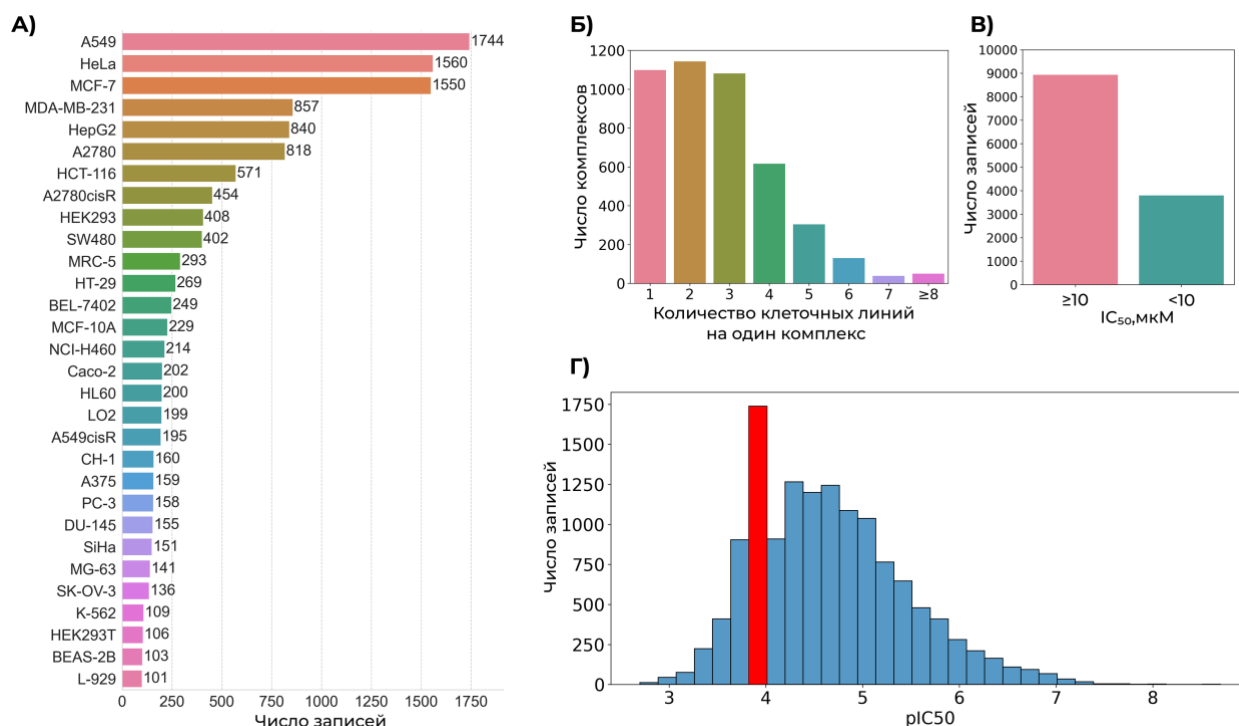


Рисунок 46. Обучающая выборка для комплексов рутения. А - распределение записей по клеточным линиям. Б - число клеточных линий, приходящихся на одно соединение. В - распределение записей между классами. Г - распределение значений pIC_{50} , красным отмечены записи, отвечающие предельным оценкам

Существенным при оценке качества оказался способ разбиения выборки. Сопоставление двух схем, описанных в разделе 3.3.2, приведено в табл. 14. Первый

способ дает заметно более высокие оценки - точность 0.836 и ROC-AUC 0.894 против 0.773 и 0.810 при разбиении по источнику. Это различие не случайно и объясняется утечкой данных: близкие по строению соединения из одной статьи попадают при разбиении по структуре и в обучающую, и в проверочную часть, отчего модель отчасти воспроизводит запомненное. Изложенное в разделе 2.1.3 применительно к утечке через дублирование получает здесь прямое количественное подтверждение, и достоверной следует признавать вторую, меньшую оценку.

Таблица 14. Качество классификации комплексов рутения по уровню цитотоксичности при двух схемах разбиения выборки

Алгоритм	Разбиение по структуре (SMILES)		Разбиение по источнику (DOI)	
	Точность	ROC-AUC	Точность	ROC-AUC
LightGBM	0.836 ± 0.010	0.894 ± 0.008	0.773 ± 0.017	0.810 ± 0.024
CatBoost	0.828 ± 0.007	0.881 ± 0.007	0.770 ± 0.018	0.801 ± 0.030
XGBoost	0.844 ± 0.015	0.888 ± 0.010	0.772 ± 0.019	0.806 ± 0.023

Качество модели заметно разнится между клеточными линиями (рис. 47). Наилучшие результаты получены для линий DU-145 и A375, где ROC-AUC превышает 0.86, тогда как для A2780, Сасо-2 и РС-3 он опускается ниже 0.74. Связи между числом записей для линии и достигаемым качеством при этом не обнаруживается: линии с полутора тысячами записей и с полутора сотнями дают сопоставимые оценки. Само по себе это не указывает на причину различий, поскольку подвыборки разнятся и по химическому составу, и по соотношению классов, и по трудности самой задачи для отдельной линии. Заметная доля разнородности вносится, однако, самими измерениями. Размах значений IC50 внутри каждой клеточной линии охватывает почти весь доступный промежуток концентраций (рис. 48 и 49): для большинства линий межквартильный размах простирается от единиц до сотни микромолей на литр. Эти распределения объединяют измерения для разных соединений и потому отражают одновременно и действительные различия их активности, и разброс условий опыта; отделить одно от другого по ним нельзя. Такое разделение возможно там, где одно и то же вещество измерено многократно: для цисплатина, о котором сказано выше (рис. 42), разброс повторных измерений на одной клеточной линии достигает почти трех

порядков по концентрации, что показывает, насколько сильно условия опыта способны влиять на измеряемую величину. Их относительный вклад в общий размах по сравнению с природой самого комплекса и клеточной линией требует отдельной оценки и в настоящей работе не определялся. Медианы при этом различаются между линиями сравнительно мало, от 8 до 48 мкмоль/л.

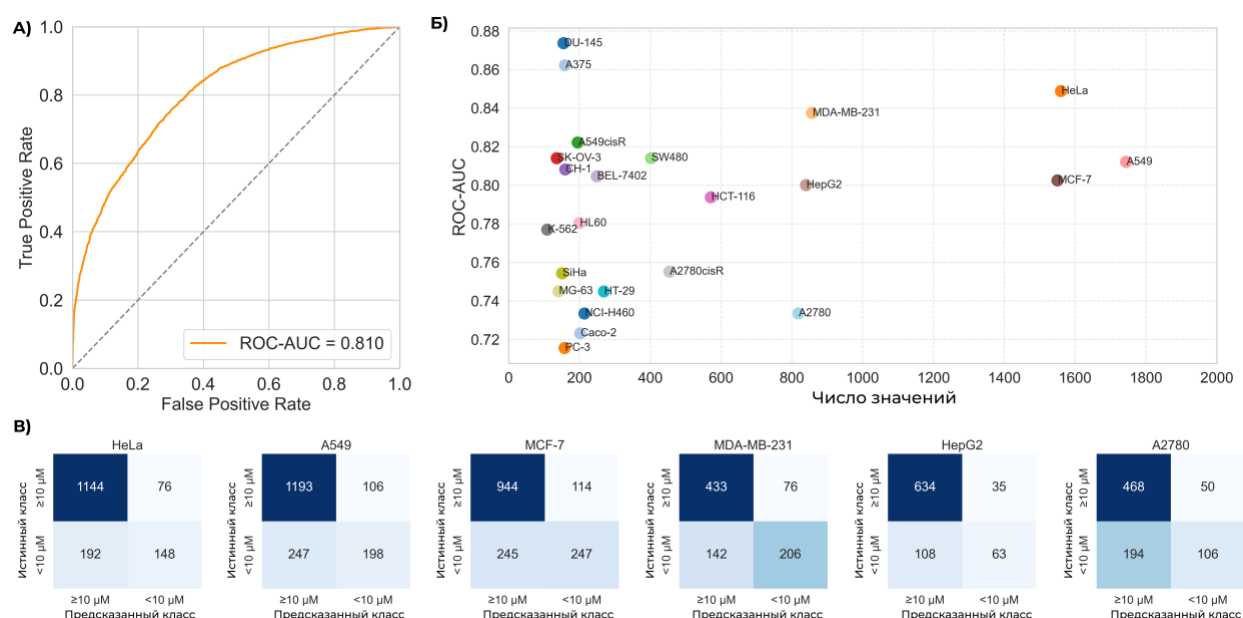


Рисунок 47. Качество классификации комплексов рутения при разбиении по источнику (DOI). А - ROC-кривая для модели по 23 клеточным линиям. Б - величины ROC-AUC для отдельных линий в зависимости от числа записей. В - матрицы ошибок для шести наиболее представленных линий

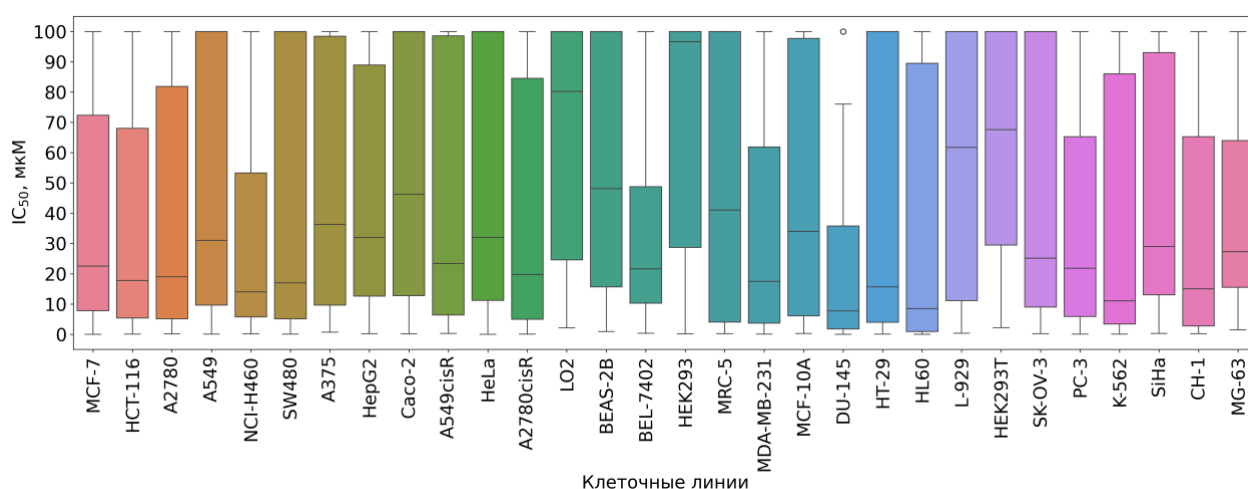


Рисунок 48. Размах измеренных значений IC_{50} по всем комплексам рутения внутри отдельных клеточных линий. Показаны тридцать наиболее представленных линий; распределения объединяют разные соединения и потому отражают как различия их активности, так и разброс условий измерения

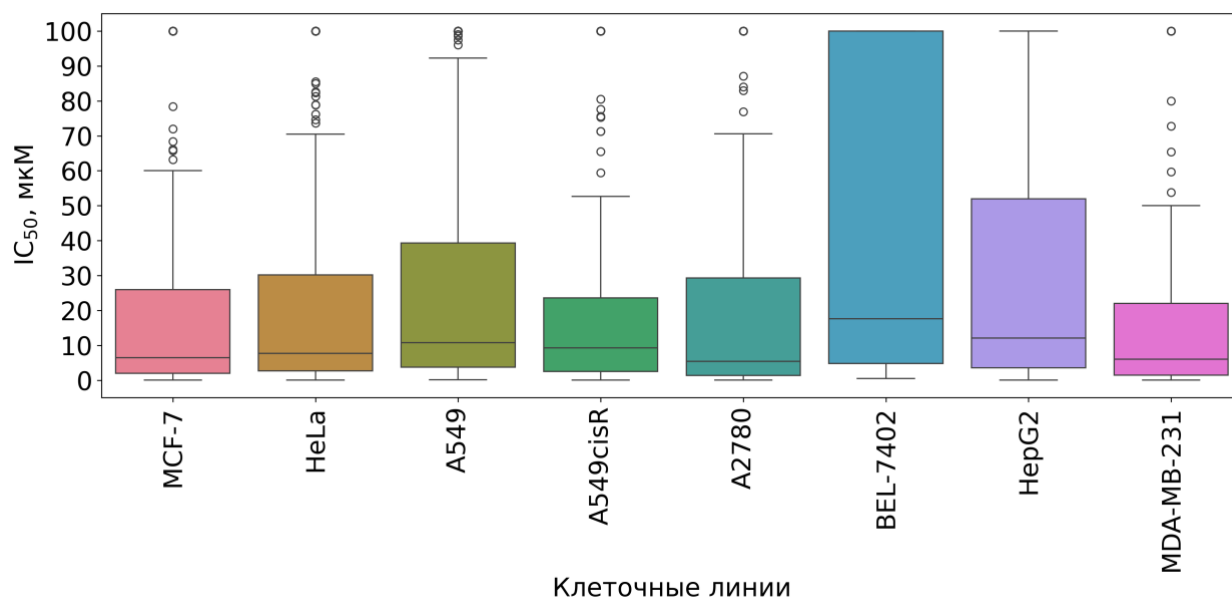


Рисунок 49. Размах измеренных значений IC_{50} по всем комплексам иридия внутри отдельных клеточных линий. Показаны восемь наиболее представленных линий; распределения объединяют разные соединения и потому отражают как различия их активности, так и разброс условий измерения

Для комплексов иридия обучение проводилось по тому же порядку на существенно меньшей выборке, и достигнутая величина ROC-AUC при разбиении по источнику (DOI) составила 0.730 (рис. 50). Снижение относительно рутения отвечает четырехкратной разнице в объеме доступных данных.

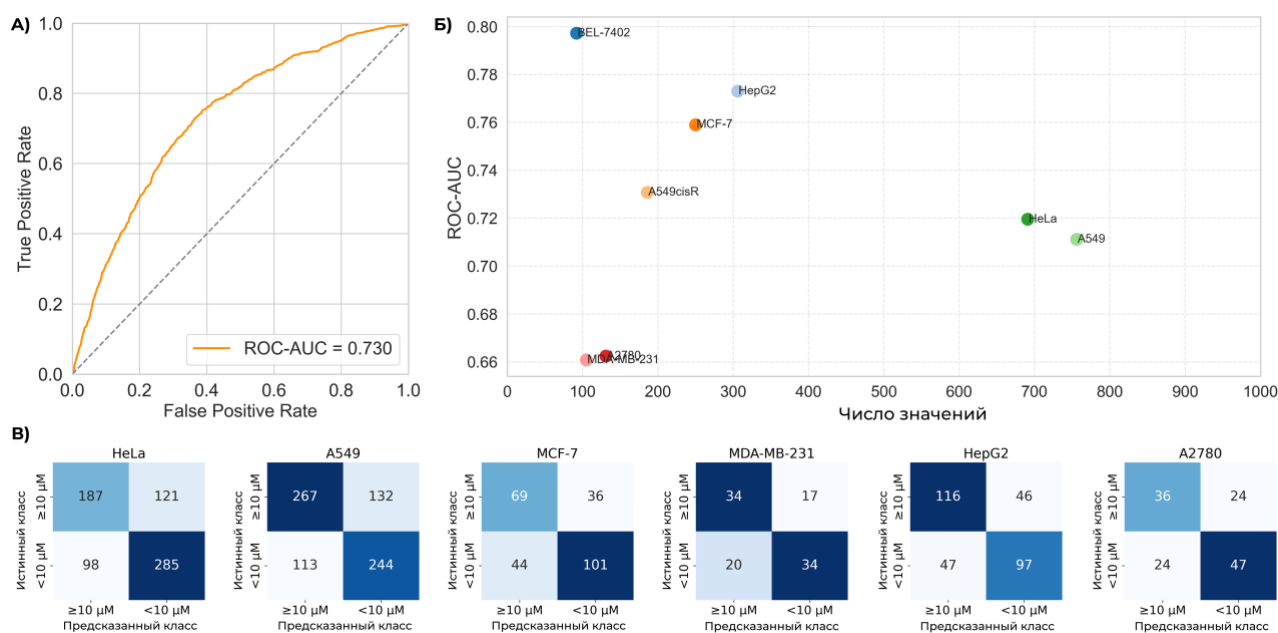


Рисунок 50. Качество классификации комплексов иридия при разбиении по источнику (DOI)

При разборе построенных моделей методом SHAP наибольшее влияние оказывают, как и следовало ожидать, признаки, определенные для всех соединений без исключения: рассчитанный по Криппену коэффициент распределения, суммарный заряд комплекса, молекулярная масса и площадь топологической полярной поверхности. Способ вычисления трех последних величин требует пояснения. Они определяются не для комплекса как целого и не для каждого лиганда порознь, а для лигандной системы, взятой как одна несвязная молекула вида L1.L2.L3. Отказ от вычисления по отдельным лигандам с последующим объединением в общий вектор признаков вызван двумя обстоятельствами. Во-первых, такое описание зависит от порядка перечисления лигандов, поскольку набор величин для L1 и L2 не совпадает с набором для L2 и L1, а установить единый порядок для всего многообразия встречающихся в базе лигандов невозможно. Во-вторых, число лигандов у комплексов различно и меняется от двух до шести, хотя и с явным преобладанием трех (рис. 51), отчего потребовалось бы дополнять вектор до заданной длины либо усекать его, что вносит в описание произвол. Полученные таким образом величины следует понимать как приближенную замену липофильности и полярности самого комплекса, а не как их измеренные значения. Необходимость подобных обходных приемов есть прямое

следствие затруднения, разобранный в разделе 2.3: общепринятого способа описания координационных соединений не существует. С этой оговоркой возрастание коэффициента распределения повышает прогнозируемую цитотоксичность, тогда как возрастание полярной поверхности действует противоположным образом, что согласуется с известной связью липофильности и проницаемости клеточной мембраны. Зависимости эти, однако, немонотонны. Увеличение суммарной молекулярной массы лигандов повышает прогнозируемую вероятность отнесения к цитотоксичному классу лишь до величины около 1000, после чего вклад заметно убывает: начиная с некоторой массы свойства определяются, по-видимому, уже не металлическим центром, а иным структурным фрагментом. Заряд комплекса действует еще сложнее: нейтральные соединения относятся моделью и к цитотоксичным, и к нецитотоксичным, однокатионные значительно чаще относятся к цитотоксичным, а двухкатионные, напротив, преимущественно к нецитотоксичным. Объяснение этому следует, вероятно, искать в соотношении двух обстоятельств: положительно заряженные комплексы охотнее взаимодействуют с отрицательно заряженными нуклеиновыми кислотами и мембранами митохондрий, но избыточная плотность заряда затрудняет проникновение в клетку. Из структурных признаков наибольшее направленное влияние оказывают арильный заместитель при атоме фосфора и циклометаллированный фрагмент, повышающие прогнозируемую цитотоксичность, тогда как иминная часть пиридинового кольца действует в обратную сторону. Действие последнего признака также нелинейно: единичное вхождение фрагмента на прогноз почти не влияет, тогда как три, четыре и в особенности пять вхождений заметно понижают прогнозируемую цитотоксичность, вследствие чего толковать его следует как присутствие двух и более бипиридильных фрагментов в молекуле (рис. 52 и 53). Для комплексов иридия обнаруживаются те же закономерности, но выраженные слабее. Влияние коэффициента распределения и полярной поверхности остается прежним, тогда как молекулярная масса заметно понижает прогнозируемую цитотоксичность лишь при малых величинах, ниже 500. Существенные структурные признаки при этом иные: N-гетероциклический карбеновый фрагмент и подвешенные фенильные группы повышают прогнозируемую цитотоксичность, тогда как имидазофенантролиновый фрагмент ее понижает. Оба наблюдения согласуются с

известными химическими сведениями: понижение концентрации полумаксимального ингибирования при введении фенильных групп объясняется возрастанием липофильности, а комплексы с протяженным сопряженным вспомогательным лигандом образуют активные формы кислорода лишь при облучении, оставаясь в темноте нецитотоксичными.

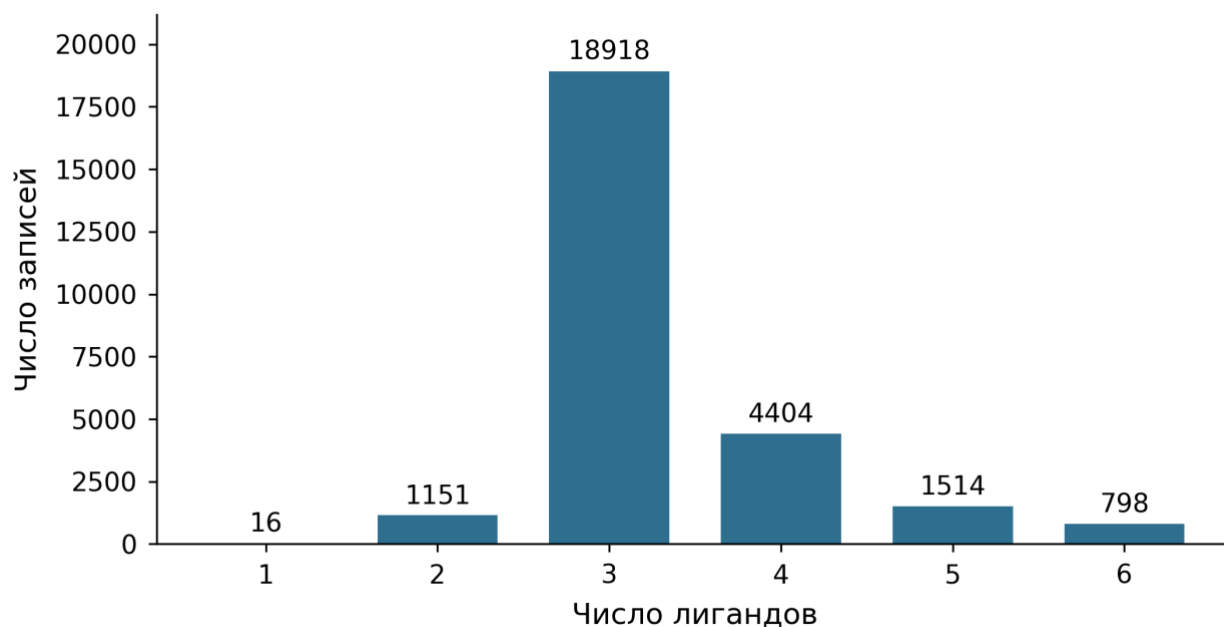


Рисунок 51. Распределение записей по числу лигандов в комплексе в базе данных MetalCytoToxDB

Рисунок 52. Интерпретация модели для комплексов иридия методом SHAP. А - вклад наиболее значимых признаков в прогнозируемую вероятность отнесения к цитотоксичному классу, структурные фрагменты показаны слева. Б, В - зависимость вклада от значения признака для молекулярной массы и заряда комплекса

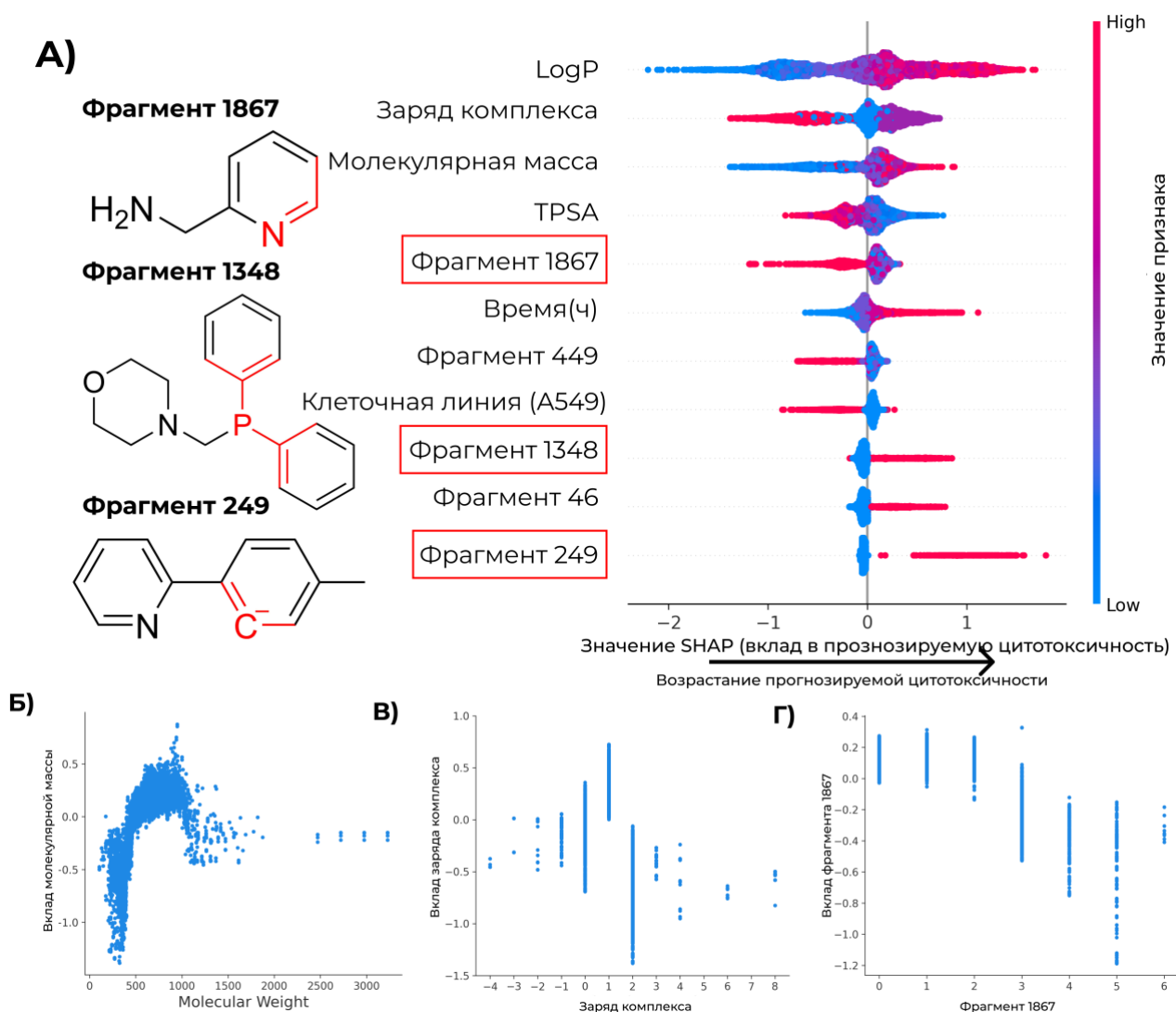


Рисунок 53. Интерпретация модели для комплексов рутения методом SHAP. А - вклад десяти наиболее значимых признаков в прогнозируемую вероятность отнесения к цитотоксичному классу, структурные фрагменты показаны слева. Б, В, Г - зависимость вклада от значения признака для молекулярной массы, заряда комплекса и фрагмента 1867

Прежде чем переходить к проверке на новых данных, следовало убедиться, что достигнутое качество опирается на действительную связь строения с активностью. Достигнутая при перемешивании меток величина ROC-AUC составила 0.512, то есть близка к случайному угадыванию (рис. 54), а матрицы ошибок для отдельных клеточных линий обнаруживают вырождение: для линии HeLa модель относит к цитотоксичному классу лишь 10 соединений из 340 действительно цитотоксичных. Снижение до уровня, близкого к случайному, согласуется с наличием действительной

зависимости между строением и цитотоксичностью, лежащей в основе достигнутых ранее величин порядка 0.81. Сама по себе эта проверка, однако, не исключает систематических смещений выборки, поскольку перемешивание разрушает всякую связь признаков с метками независимо от ее природы.

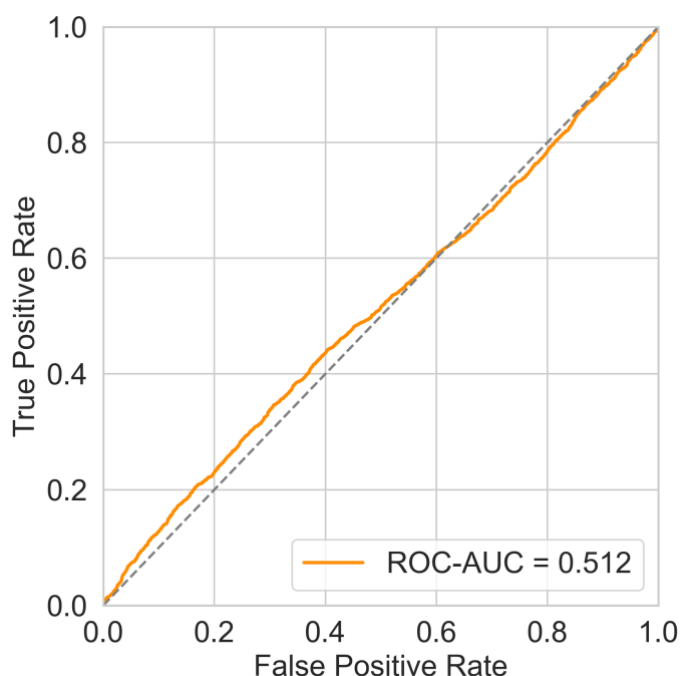


Рисунок 54. Качество классификации при случайном перемешивании меток классов

Наиболее строгой проверкой пригодности модели служит прогнозирование для соединений, опубликованных позже тех, на которых она обучалась, поскольку такая постановка воспроизводит условия действительного применения. Обучение проводилось на работах, вышедших до 2024 года включительно, а прогнозирование - для 325 записей 2025 года. Достигнутая величина ROC-AUC составила 0.74. Доля попаданий и коэффициент обогащения, определенные в разделе 3.3.2, составили 90 % и 2.25 соответственно. Существенно, что при последовательном увеличении проверочной выборки с одновременным сокращением обучающей ни ROC-AUC, ни доля попаданий, ни коэффициент обогащения заметно не ухудшались, что свидетельствует об устойчивости модели, обеспеченной объемом и разнообразием собранных данных.

Наконец, была решена задача, ради которой набор и охватывает пять металлов. Для родия, осмия и рения данных недостаточно, чтобы построить самостоятельную

модель, и потому проверялось, возможно ли переносить сведения между подмножествами разных металлов. К признакам лигандов добавлялись признаки самого металла, перечисленные в разделе 3.2.2, которые в сжатом виде передают и его тождество, и электронное строение, а проверка велась исключением одного металла.

Результаты приведены на рис. 55. При обучении без данных о металле величина ROC-AUC составляет 0.643 для родия, 0.767 для осмия и 0.658 для рения, тогда как при включении их в общую многометалльную модель - 0.738, 0.789 и 0.727 соответственно. Иными словами, даже полное отсутствие сведений о металле в обучающей выборке позволяет прогнозировать цитотоксичность его комплексов заметно лучше случайного угадывания, а добавление даже небольшого их числа доводит качество до уровня, сопоставимого с достигнутым для иридия. Эта возможность имеет прямое практическое значение: она позволяет распространить прогнозирование на металлы, для которых экспериментальных данных накоплено мало, не дожидаясь их накопления.

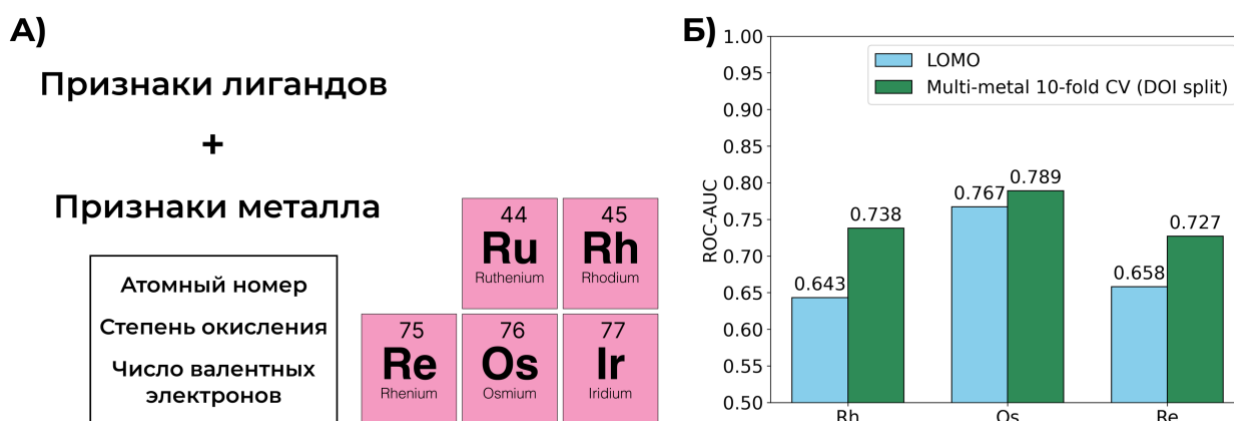


Рисунок 55. Перенос сведений между подмножествами разных металлов. А - признаки металлического центра, добавляемые к признакам лигандов. Б - величины ROC-AUC при исключении одного металла из обучения и при включении его в многометалльную модель

Построенная модель была применена, кроме того, по прямому назначению. Поиском по подструктуре в ресурсе PubChem отобрано 68681 соединение, из которых после отсеивания осталось 49152 кандидата в лиганды. Для отвечающих им гипотетических комплексов рутения спрогнозирован класс цитотоксичности в

отношении линий A549, HeLa и MCF-7, и 20344 соединения отнесены к цитотоксичным. Этот порядок показывает, каким образом собранные данные и построенная на них модель позволяют сузить пространство перебора перед синтезом (рис. 56).

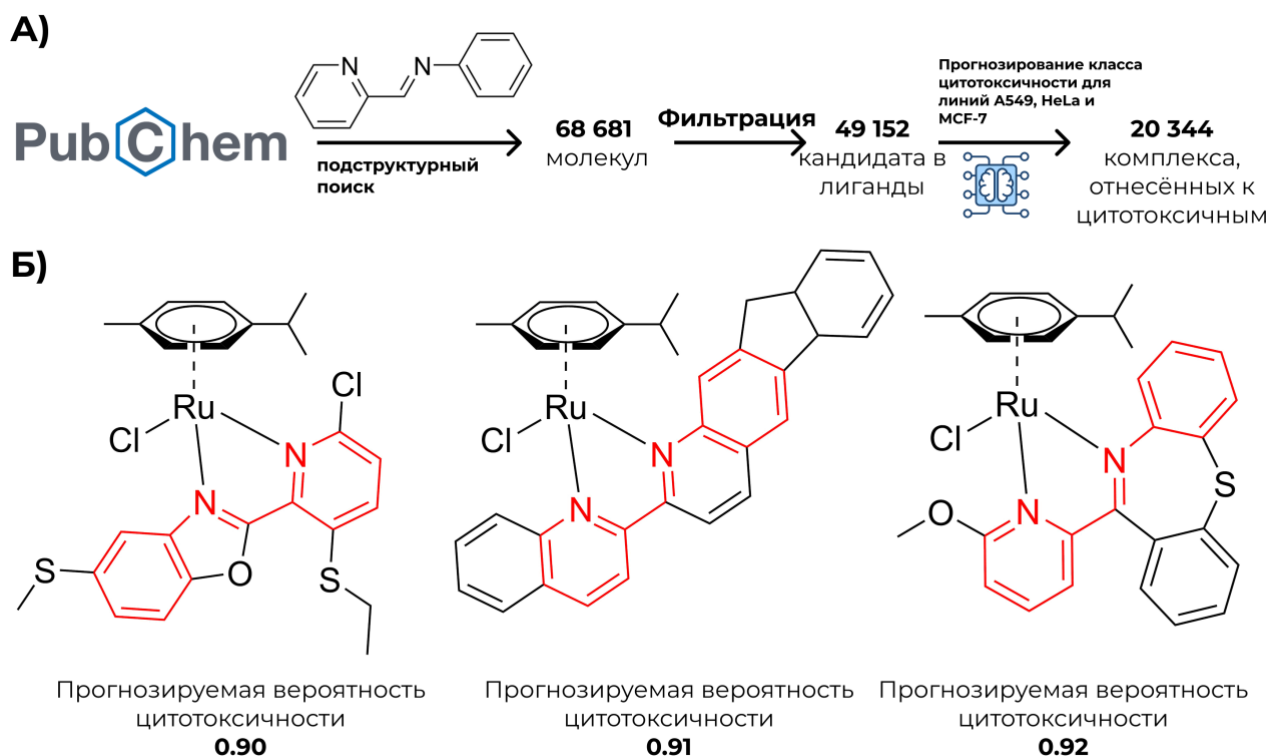


Рисунок 56. Применение модели для отбора соединений. А - порядок обработки: поиск по подструктуре в PubChem, отсеивание и прогнозирование класса цитотоксичности. Б - примеры гипотетических комплексов рутения с наибольшей прогнозируемой вероятностью цитотоксичности

Пределы применимости построенных моделей следует очертить прямо. Первое и наиболее существенное ограничение вытекает из самого способа представления: комплекс задается перечислением лигандов, а взаимное расположение их вокруг металлического центра и природа противоиона не учитываются. Следовательно, влияние положения координационных мест, геометрических изомеров и энантиомеров, равно как и процессов с участием противоиона, такими моделями не улавливается. Между тем эти различия бывают существенны: в литературе описаны случаи, когда геометрические изомеры и энантиомеры заметно расходились по цитотоксичности, что объяснялось либо различием липофильности вследствие изменения полярности

комплекса при ином расположении лигандов, либо неодинаковой реакционной способностью по отношению к хиральным биомолекулам среды. Это ограничение есть прямое следствие того же затруднения с представлением координационных соединений, которому посвящен раздел 2.3.

Второе ограничение состоит в том, что подход применим лишь к соединениям с одним металлическим центром. Лиганды, сами содержащие металл, - например, производные ферроцена, - интерпретированы быть не могут.

Третье касается охвата клеточных линий. Для неопухолевых линий устойчивое отнесение к классам требует смещения порога, поскольку при принятом пороге распределение между классами оказывается сильно несбалансированным. Для линий же, представленных малым числом записей, прогнозирование невозможно вовсе.

Четвертое относится к избирательности действия. Как отмечено выше, индексы селективности вычислены для различных сочетаний опухолевой и неопухолевой линий и потому строго между собой несопоставимы, а для 35 % сочетаний точное значение недоступно вследствие предельных оценок в первоисточниках, причем для 20 % отношение не определяется вовсе, а остальные дают лишь одностороннюю его границу. Ни объем, ни устройство этих данных не позволяют в настоящее время построить для избирательности надежную модель.

Качество многометалльной модели для металлов, представленных малым числом записей, остается ограниченным, что видно и из приведенных выше величин. Перечисленные ограничения очерчивают направления дальнейшей работы: расширение данных об избирательности, привлечение алгоритмов, приспособленных к малым выборкам, и разработка способов представления, учитывающих взаимное расположение лигандов.

4.3.3. Представленность комплексов металлов в ресурсах данных о биологической активности

Необходимость самостоятельного сбора данных о цитотоксичности комплексов, положенного в основу описанных выше наборов, требует обоснования, поскольку крупнейший ресурс сведений о биологической активности содержит миллионы

измерений и мог бы, казалось, служить источником и для комплексов. Изложенное в разделе 2.5.2 положение о том, что металлодержащие соединения исключаются из стандартизации, было поэтому проверено количественно.

Проверка проводилась обращением к программному интерфейсу ресурса (выпуск ChEMBL 37, обращение 1 сентября 2026 года). Из 2921148 молекул, содержащихся в нем, 23329 не имеют структурного представления вовсе, причем 5193 из них отнесены к малым молекулам, то есть по типу записи ничем не отличаются от прочих соединений. Для оценки состава этого подмножества была взята случайная выборка из 2000 записей, набранная с произвольных смещений. Название приведено у 1276 из них, и в 670 случаях, то есть у 52.5 % названных, название прямо указывает на присутствие металла. Среди них встречаются как простые соли, вроде хлорида железа(II) и цитрата железа(III), так и координационные соединения, обозначенные общими наименованиями вида «комплекс платины» либо развернутыми названиями изонитрильных комплексов технеция.

Наиболее показательны, однако, частные случаи. Структурного представления в ресурсе лишены цисплатин и карбоплатин - препараты, применяемые повсеместно и рассмотренные в разделе 2.5.1 как единственные соединения металлов, чья клиническая состоятельность подтверждена практикой. При этом с записью о цисплатине связано 4119 измерений биологической активности, а с записью о карбоплатине - 2677. Эти сведения в ресурсе присутствуют и доступны, но соотнести их со строением соединения средствами самого ресурса невозможно, вследствие чего для построения моделей структура-свойство они непригодны.

Из изложенного следует, что дело не в недостаточной изученности комплексов и не в малом объеме относящихся к ним сведений. Дело в том, что принятый формат хранения структурных данных не передает координационной связи, отчего соединения этого класса выпадают из обихода на самом раннем этапе - при внесении в ресурс. Систематический сбор из первичной литературы оказывается поэтому единственно возможным путем получения данных, пригодных для построения моделей.

4.4. Публикация наборов данных и интерактивные онлайн-инструменты

Все пять созданных наборов данных депонированы в общедоступном репозитории Zenodo, каждый под собственным постоянным идентификатором и под лицензией Creative Commons Attribution 4.0, допускающей любое повторное использование при условии указания авторства. Данные размещены в виде таблиц в формате с разделителями, пригодном для машинного чтения без предварительной обработки, и снабжены описанием состава столбцов. Тем самым выполнены требования находимости, доступности и пригодности к повторному использованию, сформулированные в разделе 2.1.4, причем идентификатор присвоен набору как самостоятельному объекту, а не только описывающей его публикации.

Каждый набор снабжен двумя идентификаторами. Идентификатор версии указывает на неизменяемое состояние набора и обеспечивает воспроизводимость опубликованных на его основе результатов, тогда как общий идентификатор относится к набору в целом и всегда приводит к его последней редакции. Такое устройство существенно, поскольку все наборы пополняются и исправляются, и к настоящему времени каждый из них имеет более одной версии. В табл. 15 приведены оба идентификатора. В тексте настоящей работы описано то состояние наборов, которое отвечает указанным версиям.

Таблица 15. Постоянные идентификаторы созданных наборов данных в репозитории Zenodo, адреса онлайн-инструментов и репозиторий с исходным кодом

Набор данных	Идентификатор DOI	Онлайн-инструмент	Репозиторий с кодом
BigSolDB 2.0	10.5281/zenodo.15094978	bigsolddb.streamlit.app	github.com/levakrasnovs/BigSolDBv2.0
MixtureSolDB	10.5281/zenodo.17846306	mixturesolddb.streamlit.app	github.com/levakrasnovs/MixtureSolDB
IrLumDB	10.5281/zenodo.13868675	irlumddb.streamlit.app	github.com/levakrasnovs/irlumddb
IrCytoToxDB	10.5281/zenodo.10902988	biometalddb.streamlit.app	github.com/levakrasnovs/RuCytoToxDB
MetalCytoToxDB	10.5281/zenodo.15853577	biometalddb.streamlit.app	github.com/levakrasnovs/RuCytoToxDB

Программный код, использованный для сбора, обработки и моделирования, размещен в общедоступных репозиториях системы контроля версий: BigSolDBv2.0, MixtureSolDB, irlumdb и RuCytoToxDB в учетной записи levakrasnovs. Это обеспечивает прослеживаемость каждого этапа обработки от исходной публикации до величины, поданной на вход модели. Для IrCytoToxDB такого репозитория нет: этот набор, будучи первым по времени, собирался и выверялся целиком вручную, без применения написанных для этого программных средств.

Набор данных, выложенный файлом, пригоден для машинной обработки, но не для беглого обращения человека. Чтобы посмотреть, что известно об одном веществе, требуется загрузить десятки мегабайт и написать несколько строк кода. Требование находимости, рассмотренное в разделе 2.1.4, этим формально удовлетворяется, но на деле круг пользователей сужается до тех, кто владеет соответствующими средствами. Для устранения этого несоответствия к четырем наборам созданы общедоступные интерактивные приложения, доступные по адресам bigsolddb.streamlit.app, mixturesolddb.streamlit.app, irlumdb.streamlit.app и biometalddb.streamlit.app. Для IrCytoToxDB отдельного приложения не создавалось, поскольку содержащиеся в нем сведения вошли в MetalCytoToxDB.

Приложения для BigSolDB и MixtureSolDB устроены одинаково. Они дают обзор состава набора, поиск вещества по названию и по структурной формуле, а также выдачу всех имеющихся для выбранного вещества измерений в табличном и графическом виде, с показом температурной зависимости растворимости в каждом из растворителей. Отдельно приведено приглашение к пополнению набора, обращенное как к тем, кто ведет извлечение данных из литературы, так и к лабораториям, готовым передать собственные измерения.

Приложение для IrLumDB устроено иначе. Наряду с обзором набора оно содержит встроенные модели машинного обучения. Пользователь задает структурные формулы лигандов, и приложение возвращает прогнозируемые длину волны максимума испускания и квантовый выход, не требуя ни установки программ, ни трехмерных структур, ни навыков работы с моделями. Тем самым построенные модели доступны не только в виде описания в статье, но и как действующее средство.

Наиболее развитым оказалось приложение для набора о цитотоксичности комплексов, названное OpenBioMetalDB. Поиск по структуре ведется в нем четырьмя способами: по полному совпадению молекулы, по остову, по подструктуре и по мере сходства, причем структуру можно не только ввести в виде строки SMILES, но и нарисовать. Отбор возможен по металлу, по типу ткани, по клеточной линии, по времени экспозиции и по диапазону значений концентрации полумаксимального ингибирования. Отдельные разделы отведены источникам, фототоксичности и избирательности действия, а результат любого отбора выгружается в виде таблицы. Это приложение, в отличие от прочих, продолжает пополняться и к настоящему времени содержит 35567 значений для 9406 комплексов, испытанных на 946 клеточных линиях, извлеченных из 2463 публикаций, причем к пяти металлам, рассмотренным выше, добавлено золото, на которое приходится 23.8 % записей. Приложение тем самым переросло описанный в разделе 4.3.1 набор и представляет собой его развитие.

Следует оговорить, что приложения обращаются к текущим редакциям наборов, тогда как в настоящей работе описаны те их состояния, которым отвечают идентификаторы версий из табл. 15. Приводимые приложениями числа поэтому превышают приведенные выше.

4.5. Использование созданных наборов данных сторонними исследователями

Значимость набора данных определяется тем, что именно сторонние исследователи с этими данными делают. Поэтому все работы, цитирующие BigSolDB, были разобраны по полным текстам, а не по контексту цитирования из библиометрических баз. К августу 2026 года таких работ, выполненных посторонними коллективами, насчитывается 23. В 14 из них набор служит обучающей или проверочной выборкой либо основой всего эксперимента, в четырех входит в сводный корпус наряду с другими источниками или применяется для внешней проверки, в пяти назван как существующий ресурс без работы с данными.

Примеры первого рода разнообразны по постановке задачи. В компании Sanofi на BigSolDB 2.0 в сочетании с расчетным набором обучена модель Solvaformer[194], предназначенная для выбора растворителя при планировании автоматизированного

синтеза. В Университетском колледже Лондона на этих данных разработан способ, при котором температура встраивается в само молекулярное представление, а не подаётся отдельным признаком. Для модели Chemprop это понизило ошибку прогнозирования на величину до 38 %, а среднеквадратичную ошибку при температурах выше 330 К — до 51 %, тогда как для модели на фиксированных дескрипторах выигрыш оказался незначительным[195]. В Лос-Аламосской национальной лаборатории набор послужил основой для мета-обучения линейных моделей, где каждый растворитель рассматривался как отдельная задача[196]. Он же выбран для обучения сети, вырабатывающей температурно-зависимые представления молекул[197,198], и вошел в набор MolMerger как основной по объему источник, дав 4964 значения против 1093 и 972 из двух других[198]. Показательно и обращение к нему химиков-теоретиков: при сопоставлении четырех способов создания начальной геометрии конформеров для расчетов методом COSMO-RS данные брались преимущественно из BigSolDB 2.0, откуда взято 926 значений для 185 молекул в 36 растворителях[95].

В 2015 году в обзоре, посвященном качеству экспериментальных данных о растворимости, отмечалось, что приведение значений к единой температуре позволило бы существенно снизить межлабораторный разброс, однако публикаций, в которых температурная зависимость растворимости прогнозировалась бы непосредственно по двумерной структуре молекулы, на тот момент не существовало[11,66]. Спустя десятилетие модели FastSolv, обученные на данных BigSolDB, решают именно эту задачу. Они прогнозируют растворимость при температурах из указанного в базе данных интервала в широком круге органических растворителей, экстраполируют на не встречавшиеся при обучении растворенные вещества в два-три раза точнее ранее применявшихся подходов и приближаются к алеаторному пределу используемых данных, оцениваемому в 0.5-1 логарифмическую единицу[66]. Тем самым задача, сформулированная как нерешенная именно вследствие нехватки и разнородности экспериментальных данных, была решена после того, как эти данные были собраны и стандартизованы, — при том, что архитектуры моделей за прошедшее десятилетие принципиально не изменились.

Отдельного разбора заслуживает работа, выполненная в Массачусетском технологическом институте[96]. Растворимость прогнозируется в ней через термодинамический цикл плавления, сочетающий отдельно прогнозируемые коэффициент активности, энтальпию плавления и температуру плавления, - тот самый подход, который рассмотрен в разделе 2.2.5 обзора. BigSolDB 2.0 вошла в число наборов, на которых метод испытывался.

Существеннее, однако, что на собранных данных поставлен опыт, очерчивающий область их применимости. Для сопоставления термодинамического подхода с моделью FastSolv, обученной на BigSolDB, отобраны только те вещества, которые в BigSolDB отсутствуют: 47648 значений для 9091 вещества в 213 растворителях. Оказалось, что на неводных растворителях FastSolv термодинамическому подходу не уступает, а на водных данных заметно ему проигрывает. Причину авторы называют прямо: модель, обучаемая непосредственно на значениях растворимости, ограничена распределением обучающей выборки, тогда как распределение BigSolDB смещено в сторону неводных сред. Это заключение для настоящей работы важно и принимается полностью. Собранные данные пригодны для той области, ради восполнения которой они и собирались, но переноса на растворимость в воде сами по себе не обеспечивают.

Показательно и предложенное авторами средство. Чтобы восстановить устойчивость прогнозов за пределами обучающего распределения, они усредняют оценки, полученные при использовании пятнадцати наиболее представленных в BigSolDB растворителей в качестве опорных, и получают заметное улучшение на водных данных. Этот прием примечателен тем, что превращает неравномерность набора из недостатка в опору: растворители, представленные в нем наиболее полно, служат надежными точками отсчета для прогнозирования в тех средах, которые представлены плохо.

К MixtureSolDB, опубликованному позже, обращаются те же коллективы. В работе Левенского католического университета, Массачусетского технологического института и Делфтского технического университета подход через термодинамический цикл, упомянутый в разделе 2.2.4, распространен на бинарные смеси[199]. В работе тех

же авторов совместно с Рейнско-Вестфальским техническим университетом Ахена MixtureSolDB вместе с BigSolDB 2.0 составили блок данных по растворимости в испытательном наборе из семи свойств, для которых имеются согласованные измерения и в чистых растворителях, и в смесях. Предложенный способ оценки разделяет ошибку прогнозирования на вклад чистых компонентов и вклад отклонений от идеального смешения[200].

В первой из этих работ высказано и замечание к набору: более половины записей относится к смесям, одним из компонентов которых является вода, откуда делается вывод о смещении набора в сторону вполне определенных смесей растворителей. Замечание справедливо в том, что касается распределения записей: доля таких записей составляет 56.9 %. Следует, однако, уточнить, что это смещение относится к числу измерений, а не к охвату: смеси, содержащие воду, составляют лишь 6.9 % от 750 различных бинарных смесей набора. Иными словами, водно-органические системы измеряют существенно чаще прочих, тогда как разнообразие самих смесей этим не ограничено.

Обращает на себя внимание и набор областей, из которых приходят пользователи данных: теоретическая химия, химическая технология, фармацевтическая промышленность, где к набору обращались в компаниях GSK[73] и Sanofi, и, наконец, машинное обучение общего назначения. Показатели обращения к опубликованным материалам с момента выхода версии 2.0 составляют более 16000 обращений к статье и около 4700 скачиваний самих данных.

Существенны и высказанные замечания. Независимо друг от друга в нескольких работах отмечается одно и то же: при большом числе измерений набор содержит сравнительно немного различных пар вещество - растворитель, а распределения молекулярных характеристик заметно уже, чем у наборов растворимости в воде[201,202]. Это замечание справедливо и прямо указывает на направление дальнейшей работы: наращивать следует число различных пар и охватываемое химическое пространство.

Не закрыт и вопрос о пределе достижимой точности. В нескольких работах утверждается, что модели уперлись в шум данных при средней абсолютной ошибке около 0.75 логарифмической единицы[66,201]. Оценка, полученная в разделе 4.1.4 попарным сопоставлением независимых измерений, заметно ниже: 0.146 логарифмической единицы по средней абсолютной разности при среднеквадратичном отклонении 0.391. Величины эти, однако, характеризуют разные вещи. Первая относится к ошибке моделей на всем наборе, вторая - к расхождению между лабораториями на том его подмножестве, где имеются независимые повторные измерения. Насколько второе ограничивает первое, зависит от того, в какой мере это подмножество представительно, и окончательного ответа на этот вопрос в настоящее время не существует.

Наборы по комплексам переходных металлов опубликованы позже и цитируются пока реже, однако три случая их использования заслуживают отдельного разбора, поскольку во всех данные послужили основанием самой работы.

Первый относится к IrLumDB. В работе из Nature Nanotechnology, выполненной в университете Йонсей и Гонконгском политехническом университете, на его основе построен набор PhosIrDB, о котором прямо сказано, что он получен расширением и повторным курированием IrLumDB[203]. Сохранены и устройство набора, и описание комплекса через строение его лигандов, а обученные на нем модели дали для длины волны максимума испускания среднюю абсолютную ошибку 19.95 нм при коэффициенте детерминации 0.84, то есть практически те же величины, что получены в настоящей работе. Существеннее, однако, продолжение. Отобрав по прогнозируемым свойствам соединения из составленной перебором лигандов виртуальной библиотеки в 272104 комплекса, авторы синтезировали три ранее не описанных комплекса иридия(III) с квантовыми выходами до 0.86 и изготовили на их основе рассасывающийся имплантат, светящийся сквозь кожу и вырабатывающий электричество под действием ультразвука. Работоспособность его прослежена на мышах в течение 38 недель. Собранные в настоящей работе данные о люминесценции комплексов иридия оказались тем самым доведены сторонним коллективом до действующего устройства биомедицинского назначения.

Второй случай относится к IrCytoToxDB и показателен тем, что набор употреблен не как источник сведений о цитотоксичности, а как испытательный материал для средства построения трехмерных структур. В работе, описывающей новую редакцию программного пакета molSimplify, разрабатываемого в Массачусетском технологическом институте, именно на этом наборе показана работоспособность усовершенствованной процедуры порождения структур комплексов[204]. Авторы прямо отмечают, что IrCytoToxDB, как и другие наборы этого рода, описывает комплекс строками SMILES отдельных лигандов, вследствие чего умение быстро строить по такому описанию трехмерную структуру необходимо для последующих квантово-химических расчетов и для моделей, требующих трехмерной структуры. Для проверки способа автоматического распознавания координирующих атомов вручную размечены 312 лигандов из набора, на которых достигнуты 96.4 % правильных определений дентатности и 94.9 % верно размеченных лигандов. Из полученных описаний построено 346 комплексов иридия, из которых лишь пять оказались нереалистичными, что отвечает доле успеха 98.6 %.

Этот случай заслуживает внимания по существу. Описание комплекса перечислением лигандов было принято в настоящей работе как обходной путь, по причине отсутствия общепринятой машиночитаемой записи координационных соединений, о чем говорится в разделе 2.3.1. Оказалось, что этот путь принят и другими: собранный таким образом датасет без каких-либо преобразований подошел на вход развиваемого сторонней группой средства построения структур и послужил для него мерилom. Тем самым обходной путь обнаружил свойство, на которое при его выборе не рассчитывали, - совместимость с существующими вычислительными средствами.

Третий случай принадлежит коллективу Имперского колледжа Лондона, Массачусетского технологического института, Рурского университета в Бохуме и Городского университета Гонконга[123]. Используются оба набора, причем поразному. По IrCytoToxDB проведен разбор охватываемого химического пространства, для чего из него выделено подмножество в 244 различных лиганда, образующих 355 комплексов. Разбор обнаружил преобладание 2-фенилпиридина в качестве

циклометаллирующего лиганда и, что существеннее, показал, что отбор соединений, испытанных в сопоставимых биологических условиях, резко сокращает пригодный для сравнения объем данных. Это наблюдение послужило авторам обоснованием для получения собственной библиотеки в единых условиях. Из MetalCytoToxDB же извлечен перечень 187 циклометаллирующих лигандов, на основе которого построена расширенная виртуальная библиотека для отбора. Оба случая показательны тем, что собранные данные употреблены как средство описания предметной области и как источник строительных блоков.

Соотнося изложенное с задачами, поставленными в разделе 2.6, следует заключить следующее. Дефицит систематически собранных экспериментальных данных, названный там вторым из трех препятствий, для рассмотренных предметных областей в значительной мере преодолен, и созданные наборы получили признание сторонних коллективов, о чем свидетельствует не столько число ссылок, сколько характер использования данных.

5. Выводы

1. Разработана и апробирована на массиве более 4000 рецензируемых публикаций унифицированная методология систематического сбора, стандартизации и верификации экспериментальных данных, применяемая как к органическим соединениям, так и к комплексам переходных металлов.

2. На основе базы данных IrLumDB разработаны модели прогнозирования длины волны максимума люминесценции и квантового выхода бис-циклометаллированных комплексов иридия(III). При десятикратной перекрестной проверке средняя абсолютная ошибка лучших моделей составила 18.3 нм и 0.129 соответственно. Для длины волны максимума люминесценции проведена независимая экспериментальная проверка на выборке из 33 спектров, включающей двенадцать впервые полученных комплексов; средняя абсолютная ошибка модели CatBoost составила 16.9 нм.

3. Показана возможность применения методов машинного обучения для классификации комплексов переходных металлов (ROC-AUC = 0.810 (Ru) и 0.730 (Ir)) по цитотоксичности на основе базы данных MetalCytoToxDB.

4. Показано, что объединение разрозненных литературных данных о растворимости органических соединений в органических растворителях (BigSolDB 2.0) и их бинарных смесях (MixtureSolDB) существенно расширяет доступный охват по числу измерений, кругу растворителей и температурному интервалу, но не устраняет разброса, вносимого различиями протоколов измерения и неполнотой сведений о твердой фазе. Этот разброс ограничивает достижимую точность моделей, однако его величина требует отдельной оценки: ошибка модели на всем наборе и расхождение независимых измерений на том его подмножестве, где они имеются, характеризуют разные величины.

6. Список цитируемой литературы

1. Muratov E. N. и др. QSAR without borders // Chem. Soc. Rev. The Royal Society of Chemistry, 2020. Т. 49, № 11. С. 3525–3564.
2. Graff D. E. и др. Chemprop v2: An Efficient, Modular Machine Learning Package for Chemical Property Prediction // J. Chem. Inf. Model. 2026. Т. 66, № 1. С. 28–33.
3. Artrith N. и др. Best practices in machine learning for chemistry // Nature Chemistry 2021 13:6. Nature Publishing Group, 2021. Т. 13, № 6. С. 505–508.
4. Janela T., Bajorath J. Simple nearest-neighbour analysis meets the accuracy of compound potency predictions using complex machine learning models // Nat. Mach. Intell. 2022. Т. 4, № 12. С. 1246–1255.
5. Boobier S. и др. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water // Nat. Commun. 2020. Т. 11, № 1. С. 5753.
6. Guo J., Ding S. Do Larger Models Really Win in Drug Discovery? A Benchmark Assessment of Model Scaling in AI-Driven Molecular Property and Activity Prediction. 2026.
7. Jablonka K. M. и др. Leveraging large language models for predictive chemistry // Nat. Mach. Intell. 2024. Т. 6, № 2. С. 161–169.
8. Burns J. W. и др. Deep Learning Foundation Models for Low-Data Regimes from Classical Molecular Descriptors // J. Chem. Inf. Model. 2026.
9. Meng J. и др. Boosting the predictive performance with aqueous solubility dataset curation // Sci. Data. 2022. Т. 9, № 1. С. 71.
10. Llompart P. и др. Will we ever be able to accurately predict solubility? // Sci. Data. 2024. Т. 11, № 1. С. 303.
11. Avdeef A. Suggested Improvements for Measurement of Equilibrium Solubility-pH of Ionizable Drugs // ADMET DMPK. 2015. Т. 3, № 2.
12. Landrum G. A., Riniker S. Combining IC₅₀ or K_i Values from Different Sources Is a Source of Significant Noise // J. Chem. Inf. Model. 2024. Т. 64, № 5. С. 1560–1567.
13. Mansouri K. и др. An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling // SAR QSAR Environ. Res. 2016. Т. 27, № 11. С. 911–937.
14. Swain M. C., Cole J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature // J. Chem. Inf. Model. 2016. Т. 56, № 10. С. 1894–1904.
15. Dagdelen J. и др. Structured information extraction from scientific text with large language models // Nat. Commun. 2024. Т. 15, № 1. С. 1418.

16. Zheng J. W., Lafontant-Joseph O., Green W. H. Digitized dataset of aqueous acid dissociation constants // RSC Adv. 2026. T. 16, № 23. C. 21397–21412.
17. Dannenfelser R., Yalkowsky S. H. Database for aqueous solubility of nonelectrolytes // Bioinformatics. 1989. T. 5, № 3. C. 235–236.
18. Cirino T. и др. SangsterLogP - the largest publicly available dataset of logP values // Sci. Data. 2026. T. 13, № 1. C. 1039.
19. Perlovich G. L. Two-component molecular crystals: evaluation of the formation thermodynamics based on melting points and sublimation data // CrystEngComm. 2017. T. 19, № 21. C. 2870–2883.
20. Fourches D., Muratov E., Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research // J. Chem. Inf. Model. 2010. T. 50, № 7. C. 1189–1204.
21. Fourches D., Muratov E., Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation // J. Chem. Inf. Model. 2016. T. 56, № 7. C. 1243–1252.
22. Lee K., Moldagulov G., Grzybowski B. A. The Fragility of Bioactivity Prediction: Rigorous Dataset Splits Expose the Illusion of ML Accuracy // Chemistry – A European Journal. 2026.
23. Sheridan R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. // J. Chem. Inf. Model. 2013. T. 53, № 4. C. 783–790.
24. Wilkinson M. D. и др. The FAIR Guiding Principles for scientific data management and stewardship // Sci. Data. 2016. T. 3, № 1. C. 160018.
25. Herres-Pawlis S. и др. Minimum Information Standards in Chemistry: A Call for Better Research Data Management Practices // Angew. Chem. Int. Ed. 2022. T. 61, № 51.
26. Gauthier A. и др. A FAIR research data infrastructure for high-throughput digital chemistry // Digital Discovery. 2025. T. 4, № 12. C. 3502–3514.
27. Kim S. и др. PubChem 2025 update // Nucleic Acids Res. 2025. T. 53, № D1. C. D1516–D1525.
28. Zdrazil B. и др. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods // Nucleic Acids Res. 2024. T. 52, № D1. C. D1180–D1192.
29. Groom C. R. и др. The Cambridge Structural Database // Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater. 2016. T. 72, № 2. C. 171–179.

30. Tingle B. I. и др. ZINC-22—A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery // *J. Chem. Inf. Model.* 2023. Т. 63, № 4. С. 1166–1176.
31. Ramakrishnan R. и др. Quantum chemistry structures and properties of 134 kilo molecules // *Sci. Data.* 2014. Т. 1, № 1. С. 140022.
32. Wu H. и др. QuantumPioneer: Scalable Generation of Quantum Chemical Data for Solution-Phase Hydrogen Transfer Reactions // *J. Am. Chem. Soc.* 2026.
33. Diorazio L. J., Hose D. R. J., Adlington N. K. Toward a More Holistic Framework for Solvent Selection // *Org. Process Res. Dev.* 2016. Т. 20, № 4. С. 760–773.
34. Sheikholeslamzadeh E., Chen C.-C., Rohani S. Optimal Solvent Screening for the Crystallization of Pharmaceutical Compounds from Multisolvent Systems // *Ind. Eng. Chem. Res.* 2012. Т. 51, № 42. С. 13792–13802.
35. Brouwers J., Brewster M. E., Augustijns P. Supersaturating Drug Delivery Systems: The Answer to Solubility-Limited Oral Bioavailability? // *J. Pharm. Sci.* 2009. Т. 98, № 8. С. 2549–2572.
36. Sou T., Bergström C. A. S. Automated assays for thermodynamic (equilibrium) solubility determination // *Drug Discov. Today Technol.* 2018. Т. 27. С. 11–19.
37. Fowles D. J. и др. Physics-Based Solubility Prediction for Organic Molecules // *Chem. Rev.* 2025. Т. 125, № 15. С. 7057–7098.
38. Cruz-Cabeza A. J., Reutzel-Edens S. M., Bernstein J. Facts and fictions about polymorphism // *Chem. Soc. Rev.* 2015. Т. 44, № 23. С. 8619–8635.
39. Pudipeddi M., Serajuddin A. T. M. Trends in Solubility of Polymorphs // *J. Pharm. Sci.* 2005. Т. 94, № 5. С. 929–939.
40. Nyman J., Day G. M. Static and lattice vibrational energy differences between polymorphs // *CrystEngComm.* 2015. Т. 17, № 28. С. 5154–5165.
41. Nicoud L., Licordari F., Myerson A. S. Estimation of the Solubility of Metastable Polymorphs: A Critical Review // *Cryst. Growth Des.* 2018. Т. 18, № 11. С. 7228–7237.
42. Glomme A., März J., Dressman J. B. Comparison of a Miniaturized Shake-Flask Solubility Method with Automated Potentiometric Acid/Base Titrations and Calculated Solubilities // *J. Pharm. Sci.* 2005. Т. 94, № 1. С. 1–16.
43. Stuart M., Box K. Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases // *Anal. Chem.* 2005. Т. 77, № 4. С. 983–990.
44. Avdeef A., Berger C. M., Brownell C. pH-Metric Solubility. 2: Correlation Between the Acid-Base Titration and the Saturation Shake-Flask Solubility-pH Methods // *Pharm. Res.* 2000. Т. 17, № 1. С. 85–89.

45. Hou P. и др. Solubility Measurement, Correlation, and Solvent Effect of 3,5-Dihydroxybenzoic Acid in Nine Pure Solvents and Binary Solvent Mixture (Ethanol + Acetonitrile) at Temperatures from 293.15 to 333.15 K // J. Chem. Eng. Data. 2026. T. 71, № 3. C. 1496–1508.
46. Apelblat A., Manzurola E. Solubilities of o-acetylsalicylic, 4-aminosalicylic, 3,5-dinitrosalicylic, and p-toluic acid, and magnesium- aspartate in water from T=(278 to 348) K // J. Chem. Thermodyn. 1999. T. 31, № 1. C. 85–91.
47. Buchowski H., Ksiazczak A., Pietrzyk S. Solvent activity along a saturation line and solubility of hydrogen-bonding solids // J. Phys. Chem. 1980. T. 84, № 9. C. 975–979.
48. McGovern S. L. и др. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening // J. Med. Chem. 2002. T. 45, № 8. C. 1712–1722.
49. Llinas A., Avdeef A. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets // J. Chem. Inf. Model. 2019. T. 59, № 6. C. 3036–3040.
50. Yalkowsky S. H., Valvani S. C. Solubility and partitioning I: Solubility of nonelectrolytes in water // J. Pharm. Sci. 1980. T. 69, № 8. C. 912–922.
51. Jain N., Yalkowsky S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes // J. Pharm. Sci. 2001. T. 90, № 2. C. 234–252.
52. Bertsch-Aguilar E. и др. Data-Driven Critical Evaluation of the General Solubility Equation // J. Chem. Inf. Model. 2026.
53. Zhu X. и др. Building Machine Learning Small Molecule Melting Points and Solubility Models Using CCDC Melting Points Dataset // J. Chem. Inf. Model. 2023. T. 63, № 10. C. 2948–2959.
54. Huuskonen J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology // J. Chem. Inf. Comput. Sci. 2000. T. 40, № 3. C. 773–777.
55. Tetko I. V. и др. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices // J. Chem. Inf. Comput. Sci. 2001. T. 41, № 6. C. 1488–1493.
56. Delaney J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure // J. Chem. Inf. Comput. Sci. 2004. T. 44, № 3. C. 1000–1005.
57. Bloch D. Computer Software Review. Review of PHYSPROP Database (Version 1.0) // J. Chem. Inf. Comput. Sci. 1995. T. 35, № 2. C. 328–329.
58. Llinàs A., Glen R. C., Goodman J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? // J. Chem. Inf. Model. 2008. T. 48, № 7. C. 1289–1303.

59. Llinas A., Oprisiu I., Avdeef A. Findings of the Second Challenge to Predict Aqueous Solubility // *J. Chem. Inf. Model.* 2020. T. 60, № 10. C. 4791–4803.
60. Lusci A., Pollastri G., Baldi P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules // *J. Chem. Inf. Model.* 2013. T. 53, № 7. C. 1563–1575.
61. Francoeur P. G., Koes D. R. SolTranNet—A Machine Learning Tool for Fast Aqueous Solubility Prediction // *J. Chem. Inf. Model.* 2021. T. 61, № 6. C. 2530–2536.
62. Panapitiya G. и др. Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction // *ACS Omega.* 2022. T. 7, № 18. C. 15695–15710.
63. Sorkun M. C., Khetan A., Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds // *Scientific Data* 2019 6:1. Nature Publishing Group, 2019. T. 6, № 1. C. 1–8.
64. Sushko I. и др. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information // *J. Comput. Aided. Mol. Des.* 2011. T. 25, № 6. C. 533–554.
65. Palmer D. S., Mitchell J. B. O. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules? // *Mol. Pharm.* 2014. T. 11, № 8. C. 2962–2972.
66. Attia L. и др. Data-driven organic solubility prediction at the limit of aleatoric uncertainty // *Nat. Commun.* 2025. T. 16, № 1. C. 7497.
67. Ye Z., Ouyang D. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms // *J. Cheminform.* 2021. T. 13, № 1. C. 98.
68. Lee S. и др. Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks // *ACS Omega.* 2022. T. 7, № 14. C. 12268–12277.
69. Vermeire F. H., Chung Y., Green W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures // *J. Am. Chem. Soc.* 2022. T. 144, № 24. C. 10785–10797.
70. Vassileiou A. D. и др. A unified ML framework for solubility prediction across organic solvents // *Digital Discovery.* 2023. T. 2, № 2. C. 356–367.
71. Tetko I. V. и др. Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions // *J. Chem. Inf. Model.* 2013. T. 53, № 8. C. 1990–2000.
72. Makarov D. M. и др. Improved Solubility Predictions in scCO₂ Using Thermodynamics-Informed Machine Learning Models // *J. Chem. Inf. Model.* 2025. T. 65, № 8. C. 4043–4056.

73. Da Vià L. и др. Enabling Data-Driven Solubility Modeling at GSK: Enhancing Purge Predictions for Mutagenic Impurities // *Org. Process Res. Dev.* 2024. Т. 28, № 12. С. 4215–4224.
74. Krzyżanowski M. и др. Predicting aqueous and organic solubilities with machine learning: a workflow for identifying organic cosolvents // *Digital Discovery*. 2025. Т. 4, № 10. С. 3031–3042.
75. Qiu J., Albrecht J., Janey J. Synergistic Solvation Effects: Enhanced Compound Solubility Using Binary Solvent Mixtures // *Org. Process Res. Dev.* 2019. Т. 23, № 7. С. 1343–1351.
76. Leenhouts R. J. и др. Pooling solvent mixtures for solvation free energy predictions // *Chemical Engineering Journal*. 2025. Т. 513. С. 162232.
77. Jung H. и др. Enhancing predictive models for solubility in multicomponent solvent systems using semi-supervised graph neural networks // *Digital Discovery*. 2025. Т. 4, № 6. С. 1492–1504.
78. Jouyban-Gharamaleki A., Acree W. E. Comparison of models for describing multiple peaks in solubility profiles // *Int. J. Pharm.* 1998. Т. 167, № 1–2. С. 177–182.
79. Zhao Y. и др. Data-Driven Point Adjusted Jouyban-Acree-Artificial Neural Network Hybrid Model for Predicting Solubility of Active Pharmaceutical Ingredients in Binary Solvent Mixtures // *Ind. Eng. Chem. Res.* 2024. Т. 63, № 38. С. 16529–16544.
80. Cenci F. и др. Predicting drug solubility in organic solvents mixtures: A machine-learning approach supported by high-throughput experimentation // *Int. J. Pharm.* 2024. Т. 660. С. 124233.
81. Aprile G. и др. Automation in Pharmaceutical Crystallization // *Org. Process Res. Dev.* 2026. Т. 30, № 7. С. 1746–1780.
82. Alqarni M., Alqarni A. Development of several machine learning based models for determination of small molecule pharmaceutical solubility in binary solvents at different temperatures // *Sci. Rep.* 2025. Т. 15, № 1. С. 28880.
83. Bao Z. и др. Towards the prediction of drug solubility in binary solvent mixtures at various temperatures using machine learning // *J. Cheminform.* 2024. Т. 16, № 1. С. 117.
84. Amiri M., Khaleseh F. Predicting drug solubility in binary solvent mixtures using graph convolutional networks: a comprehensive deep learning approach // *Sci. Rep.* 2025. Т. 15, № 1. С. 45711.
85. Hansen C. M. Hansen Solubility Parameters. CRC Press, 2007.

86. Abraham M. H. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes // Chem. Soc. Rev. 1993. T. 22, № 2. C. 73.
87. Fredenslund A., Jones R. L., Prausnitz J. M. Group-contribution estimation of activity coefficients in nonideal liquid mixtures // AIChE Journal. 1975. T. 21, № 6. C. 1086–1099.
88. Perlovich G. L. Formation Thermodynamics of Two-Component Molecular Crystals: Polymorphism, Stoichiometry, and Impact of Enantiomers // Cryst. Growth Des. 2020. T. 20, № 8. C. 5526–5537.
89. Perlovich G. L. Two-Component Molecular Crystals: What Is the Difference between Drug–Drug, Drug–GRAS, and CF–CF Databases? Evaluation of Melting Points and Ideal Solubility of Unknown Co-crystals // Cryst. Growth Des. 2021. T. 21, № 9. C. 5058–5071.
90. Perlovich G. L. Co-Crystallization Thermodynamics of Two-Component Crystals: Databases, Polymorphism, and Stoichiometry Differences // Cryst. Growth Des. 2025. T. 25, № 14. C. 5445–5456.
91. Perlovich G. L. New Approach for Estimation of Two-component Crystals Cocrystallization Thermodynamics // Cryst. Growth Des. 2026. T. 26, № 9. C. 3489–3500.
92. Li L., Totton T., Frenkel D. Computational methodology for solubility prediction: Application to the sparingly soluble solutes // J. Chem. Phys. 2017. T. 146, № 21.
93. Palmer D. S. и др. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules // J. Chem. Theory Comput. 2012. T. 8, № 9. C. 3322–3337.
94. Klamt A. и др. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS // J. Comput. Chem. 2002. T. 23, № 2. C. 275–281.
95. Gomes A. M. M., Granjo J. F. O., Costa P. J. Conformer Generation Workflows for COSMO-RS Calculations: Are They All the Same? // J. Chem. Inf. Model. 2026. T. 66, № 16. C. 10004–10018.
96. Al Ibrahim E. и др. Accurately Predicting Solubility Curves via a Thermodynamic Cycle, Machine Learning, and Solvent Ensembles // J. Am. Chem. Soc. 2025. T. 147, № 49. C. 45057–45069.
97. Wu D. и др. An Interpretable Solute–Solvent Interactive Attention Module Intensified Graph-Learning Architecture toward Enhancing the Prediction Accuracy of an Infinite Dilution Activity Coefficient // Ind. Eng. Chem. Res. 2024. T. 63, № 19. C. 8741–8750.

98. Wang J. и др. Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis // J. Chem. Inf. Model. 2007. Т. 47, № 4. С. 1395–1404.
99. Avdeef A. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database // ADMET DMPK. 2020. Т. 8, № 1. С. 29–77.
100. Bradley J.-C. и др. Open Notebook Science Challenge: Solubilities of Organic Compounds in Organic Solvents // Nature Precedings. 2010.
101. Mousa N. и др. Online OCHEM multi-task model for solubility and lipophilicity prediction of platinum complexes // J. Inorg. Biochem. Elsevier, 2025. Т. 269. С. 112890.
102. Kulik H. J. Using experimental data in computationally guided rational design of inorganic materials with machine learning // J. Mater. Res. 2025. Т. 40, № 6. С. 833–848.
103. St Michel R. G. и др. The BOS-Lig Data Set: Accurate Ligand Charges from a Consensus Approach for 66,810 Experimentally Synthesized Ligands // J. Chem. Inf. Model. 2026. Т. 66, № 14. С. 8336–8348.
104. Moldagulov G. и др. Hybrid Computational Strategy for Predicting Complex Ligand–Metal Architectures // Angew. Chem. Int. Ed. 2026. Т. 65, № 14.
105. Toney J. W. и др. Identifying Dynamic Metal–Ligand Coordination Modes with Ensemble Learning // J. Am. Chem. Soc. 2025. Т. 147, № 52. С. 48218–48234.
106. Bento A. P. и др. An open source chemical structure curation pipeline using RDKit // J. Cheminform. 2020. Т. 12, № 1. С. 51.
107. Orsi M., Frei A. ELECTRUM: an electron configuration-based universal metal fingerprint for transition metal compounds // Digital Discovery. 2025. Т. 4, № 12. С. 3567–3577.
108. Rasmussen M. H. и др. SMILES all around: structure to SMILES conversion for transition metal complexes // J. Cheminform. 2025. Т. 17, № 1. С. 63.
109. Kevlishvili I., Dorabawila D. Taming T-REX: a canonical language for geometry-aware generative design of transition-metal complexes // Digital Discovery. 2026. Т. 5, № 7. С. 2845–2860.
110. Toney J. W. и др. Graph neural networks for predicting metal–ligand coordination of transition metal complexes // Proceedings of the National Academy of Sciences. 2025. Т. 122, № 41.

111. Zhao X. и др. High-Throughput Prediction of Metal-Embedded Complex Properties with a New GNN-Based Metal Attention Framework // J. Chem. Inf. Model. 2025. T. 65, № 5. С. 2350–2360.
112. Taylor M. G. и др. Architector for high-throughput cross-periodic table 3D complex building // Nat. Commun. 2023. T. 14, № 1. С. 2786.
113. Lee K. и др. MetalloGen: Automated 3D Conformer Generation for Diverse Coordination Complexes // J. Chem. Inf. Model. 2025. T. 65, № 21. С. 11878–11891.
114. Jin H., Merz K. M. Partial to Total Generation of 3D Transition-Metal Complexes // J. Chem. Theory Comput. 2024.
115. Balcells D., Skjelstad B. B. TmQM Dataset - Quantum Geometries and Properties of 86k Transition Metal Complexes // J. Chem. Inf. Model. American Chemical Society, 2020. T. 60, № 12. С. 6135–6146.
116. Kneiding H., Balcells D. tmQMg* Data Set: Excited State Properties of 74k Transition Metal Complexes // J. Chem. Inf. Model. 2025. T. 65, № 21. С. 11766–11777.
117. Garrison A. G. и др. BOS-TMC Data Set: DFT Properties of 159k Experimentally Characterized Transition Metal Complexes Spanning Multiple Charge and Spin States // J. Chem. Inf. Model. 2026.
118. Karpov K. V. и др. A universal model for predicting the stability of complexes with rare earths and trivalent actinides // J. Chem. Phys. 2026. T. 165, № 4.
119. Darouich S. и др. Robust generative transition-state models for unseen chemistry // Nat. Comput. Sci. 2026.
120. Vigna V. и др. Prediction of Pt, Ir, Ru, and Rh complexes light absorption in the therapeutic window for phototherapy using machine learning // J. Cheminform. BioMed Central Ltd, 2025. T. 17, № 1. С. 1–13.
121. Welsh A., Husbands D., Frei A. High-Throughput Combinatorial Metal Complex Synthesis // Angew. Chem. Int. Ed. 2025. T. 64, № 6.
122. Kench T. и др. Discovery of Phototoxic Metal Complexes with Antibacterial Properties via a Combinatorial Approach // Inorg. Chem. American Chemical Society, 2025. T. 64. С. 5121.
123. Kench T. и др. Exploring chemical space for iridium(iii) complexes: a direct-to-biology (D2B) approach to identifying anticancer and antibacterial agents // Chem. Sci. 2026.
124. Niemeier F. и др. Combinatorial Synthesis toward the Discovery of Highly Cytotoxic Fe(III) Complexes // J. Med. Chem. American Chemical Society, 2024.

125. Husbands D. R. и др. High-throughput synthesis and screening of cyclometallated iridium complexes for antibacterial applications. 2026.
126. Özsan Ç. и др. Unlocking Iridium Chemical Space for Antibiotic Discovery through One-Pot Combinatorial Synthesis in a Direct-to-Biology Approach. 2026.
127. Kench T. и др. A Semi-Automated, High-Throughput Approach for the Synthesis and Identification of Highly Photo-Cytotoxic Iridium Complexes // *Angew. Chem. Int. Ed.* John Wiley & Sons, Ltd, 2024. T. 63, № 18. C. e202401808.
128. Husbands D. R. и др. High-throughput triazole-based combinatorial click chemistry for the synthesis and identification of functional metal complexes // *Nat. Commun.* 2025. T. 16, № 1. C. 11195.
129. Diluzio S. и др. High-Throughput Screening and Automated Data-Driven Analysis of the Triplet Photophysical Properties of Structurally Diverse, Heteroleptic Iridium(III) Complexes // *J. Am. Chem. Soc.* American Chemical Society, 2021. T. 143, № 2. C. 1179–1194.
130. Yersin H. и др. The triplet state of organo-transition metal compounds. Triplet harvesting and singlet harvesting for efficient OLEDs // *Coord. Chem. Rev.* 2011. T. 255, № 21–22. C. 2622–2652.
131. Wilde A. P., Watts R. J. Solvent effects on metal-to-ligand charge-transfer bands in ortho-metalated complexes of iridium(III): estimates of transition dipole moments // *J. Phys. Chem.* 1991. T. 95, № 2. C. 622–629.
132. Henwood A. F., Zysman-Colman E. Lessons learned in tuning the optoelectronic properties of phosphorescent iridium(iii) complexes // *Chem. Commun.* 2017. T. 53, № 5. C. 807–826.
133. Englman R., Jortner J. The energy gap law for radiationless transitions in large molecules // *Mol. Phys.* 1970. T. 18, № 2. C. 145–164.
134. Baranoff E. и др. Cyclometallated Iridium Complexes as Sensitizers for Dye-Sensitized Solar Cells // *Chem. Asian J.* 2010. T. 5, № 3. C. 496–499.
135. Terrones G. G. и др. Low-cost machine learning prediction of excited state properties of iridium-centered phosphors // *Chem. Sci.* The Royal Society of Chemistry, 2023. T. 14, № 6. C. 1419–1433.
136. Ksenofontov A. A., Lukanov M. M., Bocharov P. S. Can machine learning methods accurately predict the molar absorption coefficient of different classes of dyes? // *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 2022. T. 279. C. 121442.
137. Wang S. и др. Predictions of photophysical properties of phosphorescent platinum(II) complexes based on ensemble machine learning approach // *J. Comput. Chem.* 2024. T. 45, № 6. C. 321–330.

138. Wang S. и др. A general protocol for phosphorescent platinum(II) complexes: generation, high throughput virtual screening and highly accurate predictions // *Materials Futures*. 2025. Т. 4, № 2. С. 025601.
139. Karuth A. и др. Combined Machine Learning, Computational, and Experimental Analysis of the Iridium(III) Complexes with Red to Near-Infrared Emission // *J. Phys. Chem. Lett.* 2024. Т. 15, № 2. С. 471–480.
140. Pal Y. и др. Predicting Emission Spectra of Heteroleptic Iridium Complexes Using Artificial Chemical Intelligence // *ChemPhysChem*. 2024. Т. 25, № 16.
141. Peng D. и др. Stacking machine learning models for predicting photophysical properties of iridium complexes // *J. Photochem. Photobiol. A Chem. Elsevier*, 2025. Т. 466. С. 116374.
142. Cheng Z. и др. Automatic Screen-out of Ir(III) Complex Emitters by Combined Machine Learning and Computational Analysis // *Adv. Opt. Mater. John Wiley & Sons, Ltd*, 2023. Т. 11, № 18. С. 2301093.
143. Wang C. и др. Discovery of Novel Iridium(III) Complex-Based Photofunctional Materials Using Excited-State Descriptors // *Adv. Opt. Mater.* 2025. Т. 13, № 7.
144. Zhang C. и др. Platinum-based drugs for cancer therapy and anti-tumor strategies // *Theranostics*. 2022. Т. 12, № 5. С. 2115–2132.
145. Drees M. и др. Cycloplatin: A novel platinum compound exhibiting a different spectrum of anti-tumour activity to cisplatin // *Eur. J. Cancer*. 1995. Т. 31, № 3. С. 356–361.
146. Горбунова В. А.; Оборотова Н. А.; Чельцов П. А.; Барышников А. Ю.; Матвеев Б. П.; Бухаркин Б. В.; Багрова С. Г.; Орел Н. Ф.; Бяхов М. Ю.; Грицун Н. Н. Новый препарат платины второго поколения циклоплатам лиофилизированный // *Вестник РОНЦ им. Н. Н. Блохина РАМН*. 2001. Т. 12, № 4. С. 53–57.
147. Rademaker-Lakhai J. M. и др. A Phase I and Pharmacological Study with Imidazolium- *trans*- DMSO-imidazole-tetrachlororuthenate, a Novel Ruthenium Anticancer Agent // *Clinical Cancer Research*. 2004. Т. 10, № 11. С. 3717–3727.
148. Leijen S. и др. Phase I/II study with ruthenium compound NAMI-A and gemcitabine in patients with non-small cell lung cancer after first line therapy // *Invest. New Drugs*. 2015. Т. 33, № 1. С. 201–214.
149. Burris H. A. и др. Safety and activity of IT-139, a ruthenium-based compound, in patients with advanced solid tumours: a first-in-human, open-label, dose-escalation phase I study with expansion cohort // *ESMO Open*. 2016. Т. 1, № 6. С. e000154.
150. O’Kane G. M. и др. 44P BOLD-100-001: A phase II study of BOLD-100 in combination with FOLFOX in advanced mCRC patients that have failed at least two prior lines of therapy // *Annals of Oncology*. 2024. Т. 35. С. S22.

151. O’Kane G. M. и др. Bold-100-001: A randomized phase 2 study of bold-100 plus FOLFOX vs FOLFOX in patients with refractory metastatic colorectal cancer. // *Journal of Clinical Oncology*. 2026. Т. 44, № 2_suppl.
152. Monro S. и др. Transition Metal Complexes and Photodynamic Therapy from a Tumor-Centered Approach: Challenges, Opportunities, and Highlights from the Development of TLD1433 // *Chem. Rev.* 2019. Т. 119, № 2. С. 797–828.
153. Kulkarni G. S. и др. A phase II clinical study of intravesical photo dynamic therapy in patients with BCG-unresponsive NMIBC (interim analysis). // *Journal of Clinical Oncology*. 2023. Т. 41, № 6_suppl. С. 528–528.
154. Kulkarni G. и др. PD12-13 INTERIM ANALYSIS OF LIGHT-ACTIVATED TLD-1433 IN A PHASE II CLINICAL STUDY OF BCG-UNRESPONSIVE NON-MUSCLE INVASIVE BLADDER CANCER CARCINOMA IN-SITU // *Journal of Urology*. 2025. Т. 213, № 5S.
155. BERNAL G., AQUEA G., RAMÍREZ-RIVERA S. Metal-based molecules in the treatment of cancer: From bench to bedside // *Oncol. Res.* 2025. Т. 33, № 4. С. 759–779.
156. León I. E. Transition metal complexes: a new generation of anticancer drugs // *Future Med. Chem.* 2024. Т. 16, № 17. С. 1727–1730.
157. Konkankit C. C. и др. Anticancer activity of complexes of the third row transition metals, rhenium, osmium, and iridium // *Dalton Transactions*. The Royal Society of Chemistry, 2018. Т. 47, № 30. С. 9934–9974.
158. Kostova I. Cytotoxic Organometallic Iridium(III) Complexes // *Molecules* 2025, Vol. 30, Page 801. Multidisciplinary Digital Publishing Institute, 2025. Т. 30, № 4. С. 801.
159. Szymaszek P., Tyszka-Czochara M., Ortyl J. Iridium(III) complexes as novel theranostic small molecules for medical diagnostics, precise imaging at a single cell level and targeted anticancer therapy // *Eur. J. Med. Chem.* 2024. Т. 276. С. 116648.
160. Egorova K. S., Ananikov V. P. Toxicity of Metal Compounds: Knowledge and Myths // *Organometallics*. 2017. Т. 36, № 21. С. 4071–4090.
161. Lind A. P., Anderson P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties // *PLoS One* / под ред. Olier I. 2019. Т. 14, № 7. С. e0219774.
162. Webel H. E. и др. Revealing cytotoxic substructures in molecules using deep learning // *J. Comput. Aided. Mol. Des.* 2020. Т. 34, № 7. С. 731–746.
163. Feitosa F. L. и др. Cyto-Safe: A Machine Learning Tool for Early Identification of Cytotoxic Compounds in Drug Discovery // *J. Chem. Inf. Model.* 2024. Т. 64, № 24. С. 9056–9062.

164. Liu Q. и др. Prediction and Interpretation Microglia Cytotoxicity by Machine Learning // J. Chem. Inf. Model. 2024. Т. 64, № 24. С. 9306–9326.
165. Wu Y. и др. Establishment of interpretable cytotoxicity prediction models using machine learning analysis of transcriptome features // Acta Pharm. Sin. B. 2025. Т. 15, № 3. С. 1344–1358.
166. Lee O. W. и др. Cytotoxic Profiling of Annotated and Diverse Chemical Libraries Using Quantitative High-Throughput Screening // SLAS Discovery. 2020. Т. 25, № 1. С. 9–20.
167. Orsi M. и др. Using Machine Learning to Predict the Antibacterial Activity of Ruthenium Complexes** // Angew. Chem. Int. Ed. John Wiley & Sons, Ltd, 2024. Т. 63, № 10. С. e202317901.
168. Lukanov M. M. и др. Machine learning-aided algorithm for predicting spectral properties and cytotoxicity of BODIPY-appended platinum complexes // Opt. Mater. (Amst). 2025. Т. 160. С. 116703.
169. Rusanov D. A. и др. Overcoming cisplatin resistance via deep learning-assisted de novo design of platinum complexes. 2025.
170. Cobalt [Электронный ресурс]. URL: <https://cobalt.colab.ws/> (дата обращения: 03.09.2026).
171. Landrum G. и др. rdkit/rdkit: 2026_03_6 (Q1 2026) Release.
172. Rogers D., Hahn M. Extended-connectivity fingerprints // J. Chem. Inf. Model. American Chemical Society, 2010. Т. 50, № 5. С. 742–754.
173. Noutahi E. и др. datamol-io/molfeat: 0.9.4.
174. Wildman S. A., Crippen G. M. Prediction of Physicochemical Parameters by Atomic Contributions // J. Chem. Inf. Comput. Sci. 1999. Т. 39, № 5. С. 868–873.
175. Ertl P., Rohde B., Selzer P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties // J. Med. Chem. 2000. Т. 43, № 20. С. 3714–3717.
176. Pedregosa F. и др. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2012. Т. 12. С. 2825–2830.
177. Cover T., Hart P. Nearest neighbor pattern classification // IEEE Trans. Inf. Theory. 1967. Т. 13, № 1. С. 21–27.
178. Smola A. J., Schölkopf B. A tutorial on support vector regression // Stat. Comput. 2004. Т. 14, № 3. С. 199–222.
179. Ke G. и др. LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Adv. Neural Inf. Process. Syst. 2017. Т. 30.

180. Prokhorenkova L. и др. CatBoost: unbiased boosting with categorical features // Adv. Neural Inf. Process. Syst. Neural information processing systems foundation, 2017. T. 2018-December. C. 6638–6648.
181. Chen T., Guestrin C. XGBoost // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2016. C. 785–794.
182. Akiba T. и др. Optuna // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: ACM, 2019. C. 2623–2631.
183. Lundberg S. M., Lee S. I. A Unified Approach to Interpreting Model Predictions // Adv. Neural Inf. Process. Syst. Neural information processing systems foundation, 2017. T. 2017-December. C. 4766–4775.
184. Harris C. R. и др. Array programming with NumPy // Nature. 2020. T. 585, № 7825. C. 357–362.
185. Hunter J. D. Matplotlib: A 2D Graphics Environment // Comput. Sci. Eng. 2007. T. 9, № 3. C. 90–95.
186. Neese F. и др. The ORCA quantum chemistry program package // J. Chem. Phys. 2020. T. 152, № 22.
187. Krasnov L. и др. BigSolDB: Solubility Dataset of Compounds in Organic Solvents and Water in a Wide Range of Temperatures. 2023.
188. Ju C. W. и др. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields // J. Chem. Inf. Model. American Chemical Society, 2021. T. 61, № 3. C. 1053–1065.
189. Ćwiklińska-Jurkowska M. и др. Disparities in Cisplatin-Induced Cytotoxicity—A Meta-Analysis of Selected Cancer Cell Lines // Molecules. 2023. T. 28, № 15. C. 5761.
190. Mervin L. H. и др. Understanding Cytotoxicity and Cytostaticity in a High-Throughput Screening Collection // ACS Chem. Biol. 2016. T. 11, № 11. C. 3007–3023.
191. Koutsoukas A. и др. In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window // J. Chem. Inf. Model. 2013. T. 53, № 8. C. 1957–1966.
192. Mervin L. H. и др. Target prediction utilising negative bioactivity data covering large chemical space // J. Cheminform. 2015. T. 7, № 1. C. 51.

193. Müller A. T. и др. Machine Learning-Assisted Iterative Screening for Efficient Detection of Drug Discovery Starting Points // J. Med. Chem. 2025. Т. 68, № 14. С. 14806–14817.
194. Broadbent J. и др. Solvaformer: Unified Geometric Learning for Solubility-Aware Automated Synthesis.
195. Joo C., Lee L. Y. S. Temperature-conditioned molecular representation for organic solubility prediction // Chemical Engineering Journal. 2026. Т. 546. С. 180019.
196. Pimonova Y. и др. Meta-learning linear models for molecular property prediction // Digital Discovery. 2026. Т. 5, № 7. С. 2831–2844.
197. Ren G. и др. Learning temperature-dependent molecular representations for physicochemical property prediction // AIChE Journal. 2026. Т. 72, № 6.
198. Ramani V., Karmakar T. Graph Neural Networks for Predicting Solubility in Diverse Solvents Using MolMerger Incorporating Solute–Solvent Interactions // J. Chem. Theory Comput. 2024. Т. 20, № 15. С. 6549–6558.
199. Buzzi S. и др. Predicting solubility curves in solvent mixtures using thermodynamic cycles and machine learning. 2026.
200. Ku R. J. L. и др. A Systematic Evaluation of Molecular Mixture Behavior Prediction. 2026.
201. Fan Y. и др. Hierarchical attention graph learning with LLM enhancement for molecular solubility prediction // Digital Discovery. 2026. Т. 5, № 2. С. 603–616.
202. Gopichand B. и др. Deciphering Molecular and Solvent Effects on Aqueous and Organic Solubility through Interpretable Machine Learning Approaches // ACS Omega. 2026. Т. 11, № 20. С. 29757–29771.
203. Meng X. и др. Bioresorbable and optically readable triboelectric implants enabled by nanoscale iridophosphors // Nat. Nanotechnol. 2026.
204. Terrones G. G. и др. molSimplify 2.0: Improved Structure Generation for Automating Discovery in Inorganic Molecular and Reticular Chemistry // J. Chem. Inf. Model. 2026. Т. 66, № 5. С. 2753–2767.

7. Список публикаций, в которых изложены основные результаты диссертации

1. **L.V. Krasnov**, S.V. Tatarin, D.E. Smirnov, S.I. Bezzubov. IrCytoToxDB: a dataset of iridium(III) complexes cytotoxicities against various cell lines. // Sci. Data, 2024, v. 11, p. 870. DOI: 10.1038/s41597-024-03735-w

2. S.V. Tatarin, **L.V. Krasnov**, E.V. Nykhrikova, M.M. Minin, D.E. Smirnov, A.V. Churakov, S.I. Bezzubov. Towards accelerating the discovery of efficient iridium(III) emitters using a novel database and machine learning based only on structural formulas. // J. Mater. Chem. C, 2025, v. 13, p. 8217-8227. DOI: 10.1039/d5tc00305a

3. **L.V. Krasnov**, D.A. Malikov, M.A. Kiseleva, S.V. Tatarin, S.S. Sosnin, S.I. Bezzubov. BigSolDB 2.0, dataset of solubility values for organic compounds in different solvents at various temperatures. // Sci. Data, 2025, v. 12, p. 1236. DOI: 10.1038/s41597-025-05559-8

4. D.A. Malikov, **L.V. Krasnov**, M.A. Kiseleva, E.A. Meshcheriakova, F. Kuznetsov, V. Elistratov, M. Vasiyarov, S.V. Tatarin, S.I. Bezzubov. Dataset of solubility values for organic compounds in binary mixtures of solvents at various temperatures. // Sci. Data, 2026. DOI: 10.1038/s41597-026-07047-z

5. **L.V. Krasnov**, D.A. Malikov, M.A. Kiseleva, E.V. Nykhrikova, S.V. Tatarin, S.I. Bezzubov. Machine Learning Approach to Anticancer Activity Prediction of Transition-Metal Complexes Based on a Large-Scale Experimental Database. // J. Med. Chem., 2026, v. 69, p. 8838-8851. DOI: 10.1021/acs.jmedchem.5c02755

8. Благодарности

Автор искренне благодарен научному руководителю, к.х.н. С.И. Беззубову за руководство работой, помощь и поддержку на всех этапах работ, а также всему коллективу Лаборатории кристаллохимии и рентгеноструктурного анализа за участие в обсуждении результатов и ценные рекомендации.

Автор выражает огромную благодарность: С.В. Татарину, М.А. Киселевой, Д.И. Маликову, Е.А. Мещеряковой за участие в сборе данных и плодотворные дискуссии, А.И. Загоруйко и Н.С. Буряку за разработку научного поисковика Cobalt.