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

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата химических наук

Москва - 2026 г.

Работа выполнена в Лаборатории кристаллохимии и рентгеноструктурного анализа ИОНХ РАН.

Научный руководитель:	**Беззубов Станислав Игоревич**, кандидат химических наук, старший научный сотрудник, Лаборатория кристаллохимии и рентгеноструктурного анализа, Федеральное государственное бюджетное учреждение науки Институт общей и неорганической химии им. Н.С. Курнакова Российской академии наук
Официальные оппоненты:	**Кинжалов Михаил Андреевич**, доктор химических наук, профессор, кафедра физической органической химии, Институт Химии, Федеральное государственное бюджетное образовательное учреждение высшего образования Санкт-Петербургский государственный университет
	Митрофанов Артем Александрович, кандидат химических наук, доцент, кафедра радиохимии, Химический факультет, Федеральное государственное бюджетное образовательное учреждение высшего образования Московский государственный университет имени М. В. Ломоносова
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт химии растворов им. Г.А. Крестова Российской академии наук (ИХР РАН)

Защита диссертации состоится «11» декабря 2026 г. в 14:00 часов на заседании диссертационного совета 01.4.005.97 в конференц-зале ИОНХ РАН по адресу: г. Москва, Ленинский проспект, д. 31.

С диссертацией можно ознакомиться у Ученого секретаря ИОНХ РАН (Ленинский проспект, д. 31) и на сайте ИОНХ РАН: http://www.igic.ras.ru/

Автореферат разослан «9» ноября 2026 г.

Ученый секретарь диссертационного совета, кандидат химических наук — Воронина Ю.К.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы и степень разработанности темы исследования.

Растворимость органических соединений является одним из ключевых физико-химических свойств, определяющих их применимость в синтетической, медицинской и промышленной химии[1,2]. Ввиду трудоёмкости экспериментального определения растворимости для больших библиотек соединений, разработка методов её прогнозирования по молекулярной структуре является важной задачей современной хемоинформатики[3]. Наибольшее число исследований посвящено водной растворимости, благодаря наличию больших и аннотированных датасетов, таких как AqSolDB[4]. Прогнозирование растворимости в органических растворителях развивается менее интенсивно, а ключевым ограничением остаётся нехватка крупных и разнородных наборов данных[5]. В ещё большей мере это справедливо для растворимости в смесях растворителей, которые широко используются в технологических процессах: существующие публичные базы данных в этой области невелики по объёму и охватывают ограниченное химическое пространство[6,7]. В рамках настоящей работы созданы базы данных BigSolDB 2.0 и MixtureSolDB, восполняющие указанные пробелы для индивидуальных растворителей и бинарных смесей соответственно.

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[8,9]. Разработка новых эффективных иридиевых комплексов с необходимыми свойствами требует значительных экспериментальных усилий, тогда как прогностическая точность квантово-химических расчётов для таких соединений остаётся недостаточной для задач рационального молекулярного дизайна. Применение методов машинного обучения для прогнозирования фотофизических свойств комплексов иридия ограничивалось либо узкими классами соединений[10], либо использованием расчётных, а не экспериментальных данных в качестве целевых переменных[11]. Систематической экспериментальной базы данных люминесцентных свойств структурно разнообразных бис-циклометаллированных комплексов иридия(III) в машиночитаемом формате до настоящего времени создано не было. В рамках настоящей работы создана база данных IrLumDB и разработаны ML-модели для прогнозирования длины волны испускания и квантового выхода фотолюминесценции непосредственно по SMILES-представлению лигандов.

Другим важным применением комплексов переходных металлов является противоопухолевая активность. Соединения рутения, иридия, осмия, родия и рения активно

исследуются как альтернатива препаратам платины[12,13], а ряд комплексов рутения, таких как NAMI-A, KP1019/KP1339, достиг стадии клинических испытаний[14]. Цитотоксичность является одной из ключевых количественных характеристик при первичном скрининге потенциальных металлолекарств, однако её экспериментальное определение для больших библиотек соединений остаётся трудоёмким и дорогостоящим процессом[15]. Применение методов машинного обучения для прогнозирования цитотоксичности хорошо развито для органических молекул[16], однако комплексы переходных металлов традиционно остаются недостаточно представлены в этой области вследствие трудностей в их машиночитаемом представлении[17] и отсутствия крупных систематизированных баз данных. В рамках настоящей работы последовательно созданы база данных IrCytoToxDB, содержащая данные о цитотоксичности комплексов иридия(III), и расширяющая её база данных MetalCytoToxDB, охватывающая комплексы пяти металлов. На их основе разработаны ML-модели для прогнозирования класса цитотоксичности по металлическому иону и структуре лигандов.

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, содержащая 1454 экспериментальных значения для 1287 соединений с разнообразным лигандным окружением, извлечённых из 340 публикаций, и на ее основе разработаны ML-модели прогнозирования длины волны максимума эмиссии (λmax) 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 7_{50} 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 расчётов.

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 применимая 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, обеспечивается разработанной автором методологией формирования баз данных, включающей независимое двойное извлечение данных из рецензируемых публикаций с последующей сверкой, стандартизацию структур и проверку физической допустимости значений, а также статистической проверкой построенных моделей. Для минимизации ошибок извлечения применялась процедура перекрёстной 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, ^{1}H ЯМР-спектроскопия, масс-спектрометрия высокого разрешения), данные которых хорошо воспроизводятся и согласуются между собой.

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 конференция “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.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы исследования, поставлены цель и задачи работы, а также обоснована научная новизна исследования и его практическая значимость.

Глава 1 (Литературный обзор). Рассмотрены физико-химические основы растворимости органических соединений и её роль в фармацевтической разработке и химической технологии. Представлена историческая справка по развитию методов

прогнозирования растворимости: от эмпирических правил и групповых вкладов до современных подходов на основе машинного обучения и глубоких нейронных сетей. Обсуждены ключевые общедоступные базы данных растворимости органических соединений: ESOL, AqSolDB и другие — с анализом их объёма, химического пространства и ограничений применимости. Рассмотрены стандартные метрики качества моделей прогнозирования растворимости (RMSE, MAE, R^2), а также проблемы воспроизводимости результатов, связанные с различиями в условиях измерения и неоднородностью экспериментальных данных. Отдельно рассмотрены особенности прогнозирования растворимости в смешанных растворителях и дефицит структурированных данных в этой области. Обсуждены фотофизические свойства и структурные факторы, определяющие люминесцентные характеристики циклометаллированных комплексов иридия(III), а также примеры их практического применения. Проанализированы литературные данные о противоопухолевой активности комплексов переходных металлов, механизмах их действия и существующих базах данных биологической активности металлоорганических соединений.

Глава 2 (Экспериментальная часть). Описан протокол систематического сбора экспериментальных данных из научной литературы, включая критерии отбора источников, правила извлечения числовых значений и контроль качества данных. Представлены схемы разработанных баз данных. Рассмотрены методы стандартизации химических структур путем приведения SMILES к каноническому виду с использованием библиотеки RDKit. Рассмотрены способы проверки и чистки созданных баз данных. Описано молекулярное представление соединений на основе молекулярных отпечатков ECFP4, использованных в качестве входных признаков для моделей машинного обучения. В качестве основных алгоритмов прогнозирования применялись методы градиентного бустинга (XGBoost, LightGBM, Catboost). Описаны процедуры подбора гиперпараметров и оценки важности признаков. Представлены метрики валидации регрессионных и классификационных моделей (RMSE, MAE, R^2, ROC-AUC), а также способы кросс-валидации с избежанием утечки данных между обучающей и тестовой выборками. Описан синтез 12 новых бис-циклометаллированных комплексов иридия(III) на основе арилбензимидазольного скаффолда и их характеризация комплексом физико-химических методов анализа (рентгеноструктурный анализ, рентгенофазовый анализ, ^{1}H ЯМР-спектроскопия, масс-спектрометрия высокого разрешения, спектроскопия люминесценции).

Глава 3 (Обсуждение результатов).

3.1. База данных растворимости органических соединений BigSolDB 2.0

Несмотря на то, что растворимость является одним из ключевых физико-химических свойств, определяющих применимость органических соединений в фармацевтике, химическом синтезе и материаловедении, подавляющее большинство существующих баз данных и моделей прогнозирования ограничено водными системами. Для восполнения этого пробела сформирована крупнейшая на сегодняшний день база данных экспериментальных значений растворимости органических соединений в индивидуальных растворителях — BigSolDB 2.0, содержащая 103944 записи для 1448 уникальных соединений в 213 растворителях в диапазоне температур 243-425 K, извлечённые из 1595 рецензируемых публикаций (рис. 1).

По сравнению с промежуточной версией BigSolDB 1.0 объём данных увеличен вдвое за счёт систематической проработки журнальных источников: *Journal of Chemical & Engineering Data, Chinese Journal of Chemical Engineering и Fluid Phase Equilibria* с использованием поисковой системы Cobalt с последующим ручным добавлением в базу. В базу данных включены исключительно значения растворимости, выраженные в мольной доле, что исключает ошибки пересчёта единиц. Молекулярные структуры стандартизированы в формате канонических SMILES с использованием RDKit. Хранение значений растворимости реализовано как в единицах мольной доли, так и в единицах LogS (моль/л), которые приняты в хемоинформатике для прогнозирования растворимости с помощью машинного обучения.

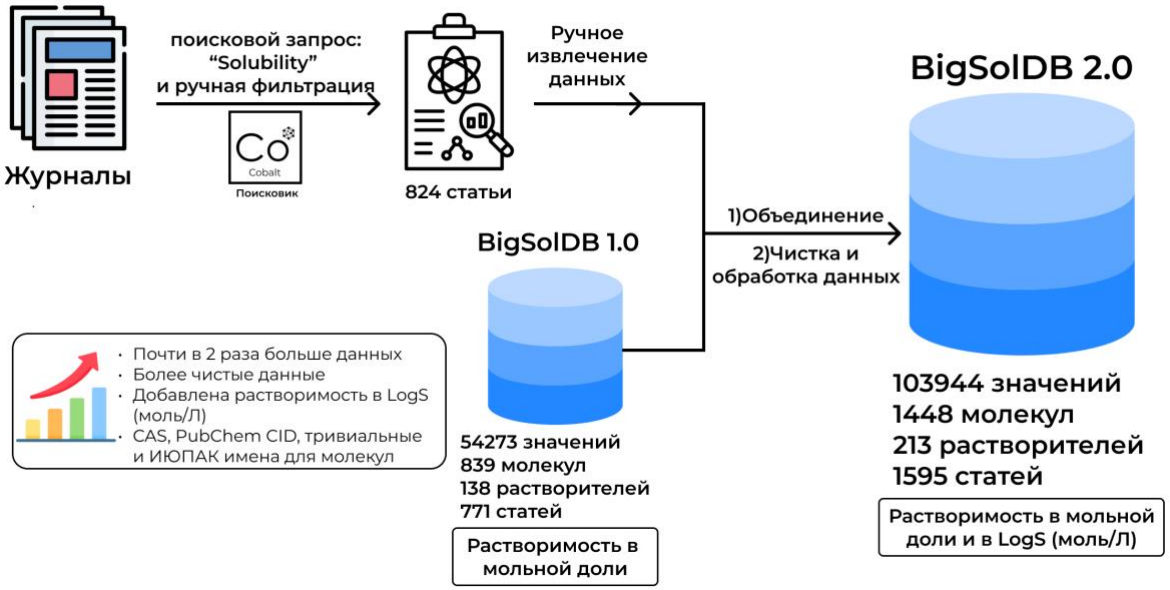


Рис.1. Общая схема создания BigSolDB 2.0.

Для пересчёта единиц из мольной доли в LogS(моль/л) были самостоятельно собраны экспериментальные значения плотностей растворителей (с зависимостью от температуры) из рецензируемых публикаций для всех растворителей, представленных более чем 100 записями

в базе. Для промежуточных температур построены линейные аппроксимации ($R^2 > 0.995$ для каждого растворителя, не менее двух независимых источников на аппроксимацию). Каждое соединение в BigSolDB 2.0 представлено идентификаторами CAS и PubChem CID, тривиальными и ИЮПАК-названиями, а также признаком одобрения FDA, полученного через ChEMBL API. Многоступенчатый контроль качества датасета включал:

- устранение дублирующихся записей, проверку физически невозможных значений мольной доли
- унификацию названий растворителей. Например, “2-butanol”, “butan-2-ol” и “sec-butyl alcohol”
- фильтрацию аномальных температурных диапазонов
- валидацию всех статей по DOI через CrossRef API.

Внутренняя согласованность базы данных подтверждена на 6939 записях с повторными независимыми измерениями в идентичных условиях (RMSE = 0.39 LogS, R^2 = 0.90), что соответствует уровню экспериментальной воспроизводимости, ранее установленному в литературе.

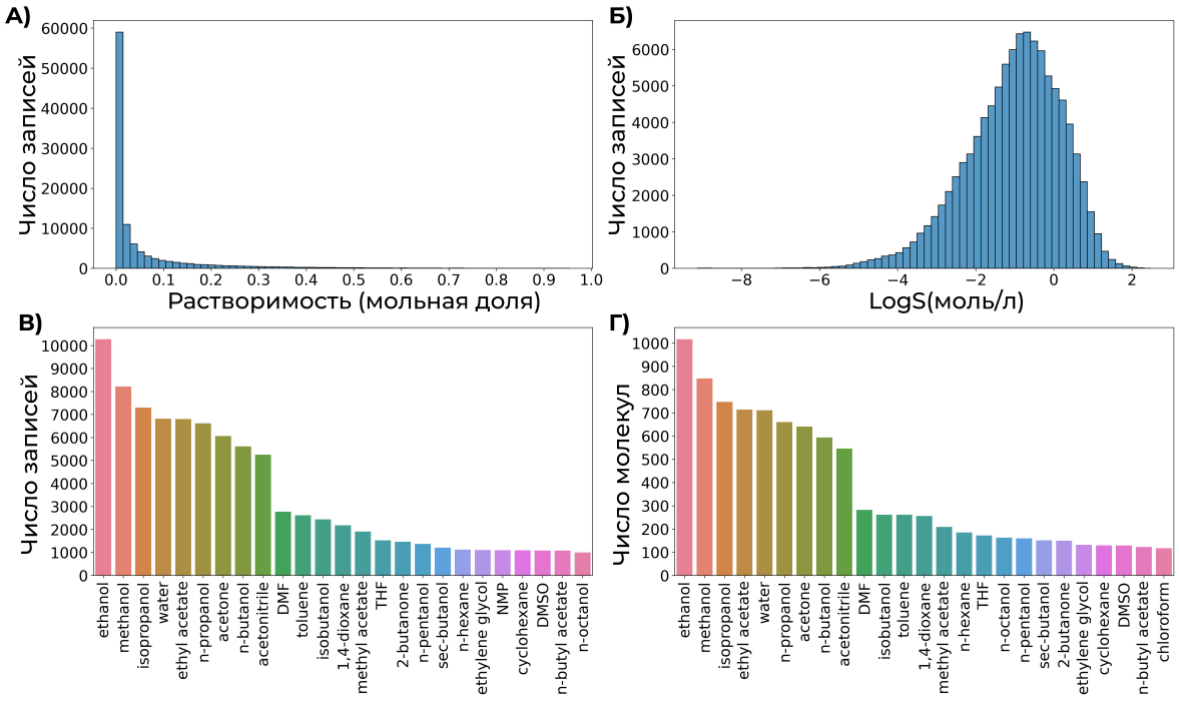


Рис.2. Статистика BigSolDB 2.0. **А)** Распределение записей по значениям растворимости в единицах мольной доли. **Б)** Распределение записей на основе значений растворимости в логарифмическом масштабе (моль/л). **В)** Распределение растворителей по числу записей (представлены 25 наиболее распространенных растворителей). **Г)** Распределение

растворителей по числу уникальных растворенных веществ (представлены 25 наиболее распространенных растворителей).

Анализ распределения данных показал, что большинство соединений обладает низкой растворимостью. Свыше половины записей соответствуют значениям мольной доли менее 0.01, что отражает типичную картину для твёрдых органических соединений в органических растворителях (рис. 2А). Логарифмическое представление (LogS, моль/л) даёт значительно более равномерное распределение с максимумом около -1 и диапазоном от -9 до +2 (рис. 2Б), что удобно для задач машинного обучения. Среди растворителей по числу записей и числу уникальных растворённых веществ лидируют низкомолекулярные спирты (этанол, метанол, изопропанол), вода, этилацетат, ацетон и ацетонитрил (рис. 2В-Г).

Для удобства работы с данными разработан веб-инструмент интерактивного поиска и визуализации на основе библиотеки Streamlit: https://bigsoldb.streamlit.app/.

3.2. База данных растворимости органических соединений в бинарных смесях растворителей MixtureSolDB

Использование бинарных смесей растворителей в синтетической химии, кристаллизации и разработке лекарственных форм значительно расширяет доступное пространство условий проведения процессов, однако прогнозирование растворимости в таких системах представляет существенно более сложную задачу по сравнению с индивидуальными растворителями ввиду дополнительной зависимости от состава смеси. Для восполнения дефицита данных в этой области сформирована крупнейшая на сегодняшний день база данных экспериментальных значений растворимости органических соединений в бинарных смесях растворителей MixtureSolDB, содержащая 175166 записей для 810 органических соединений в 750 уникальных бинарных смесях растворителей (3001 уникальная система растворитель-смесь) в диапазоне температур 252-383 К, извлечённые из 1115 рецензируемых публикаций.

По объёму данных, числу растворённых веществ, бинарных смесей и уникальных систем MixtureSolDB значительно превосходит ранее доступные публичные аналогичные датасеты: BaoDB (28098 записей) и MixSolDB (34734 записи). Отбор статей для разметки осуществлялся по поисковым запросам “solubility + binary” и “solubility + mixture” с использованием поисковой системы Cobalt. Из 5775 первично идентифицированных публикаций после удаления дубликатов и исключения нерелевантных статей к извлечению данных были допущены 1115 статей, что отражено на диаграмме в стиле PRISMA (рис. 3).

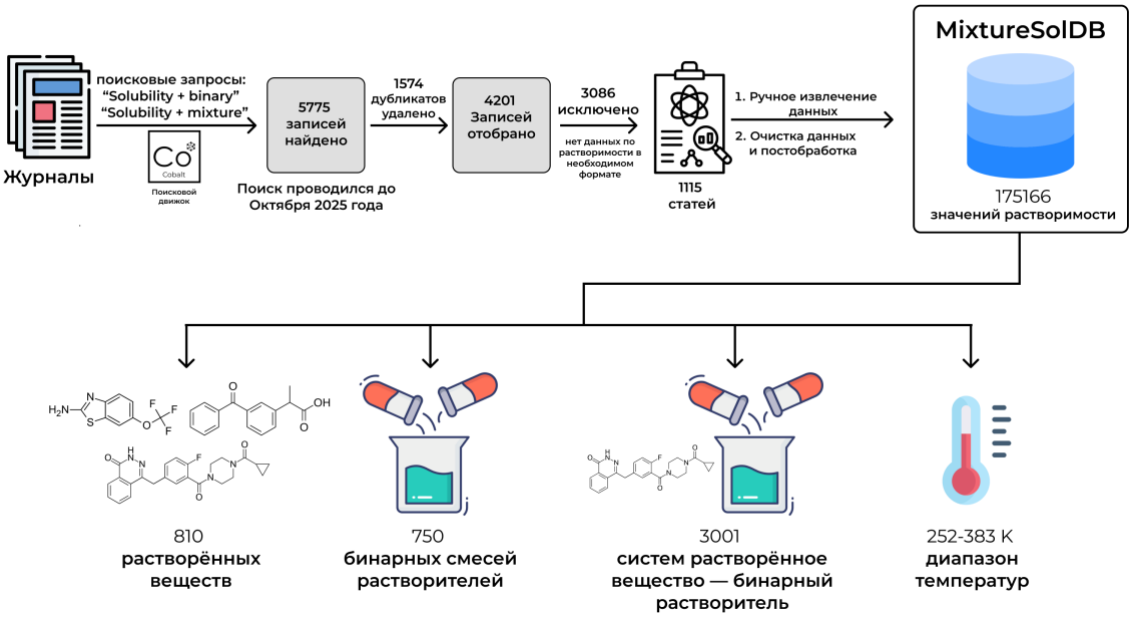


Рис.3. Общая схема создания MixtureSolDB.

Значения растворимости хранятся в единицах мольной доли, LogS(мольная доля) и LogS(г/100 г). Состав бинарной смеси представлен как в единицах мольных, так и массовых долей. Для однозначной интерпретации пары растворителей канонизированы по алфавитному порядку. Многоступенчатый контроль качества данных включал унификацию названий растворителей, проверку физически невозможных значений мольных долей, контроль температурных диапазонов относительно точек кипения растворителей и валидацию статей по DOI через CrossRef API.

Анализ распределения данных показал, что в 60% записей значение мольной доли растворённого вещества составляет менее 0.01 (рис. 4А). Логарифмическое представление даёт близкое к нормальному распределение с максимумом около LogS(мольная доля) = -2 (рис. 4Б). Среди пар растворителей безусловно доминируют водно-органические системы. Вода присутствует в 57% записей, наиболее часто в сочетании с этанолом, метанолом, изопропанолом, этилацетатом, ДМФА и ацетоном (рис. 4Г-Д).

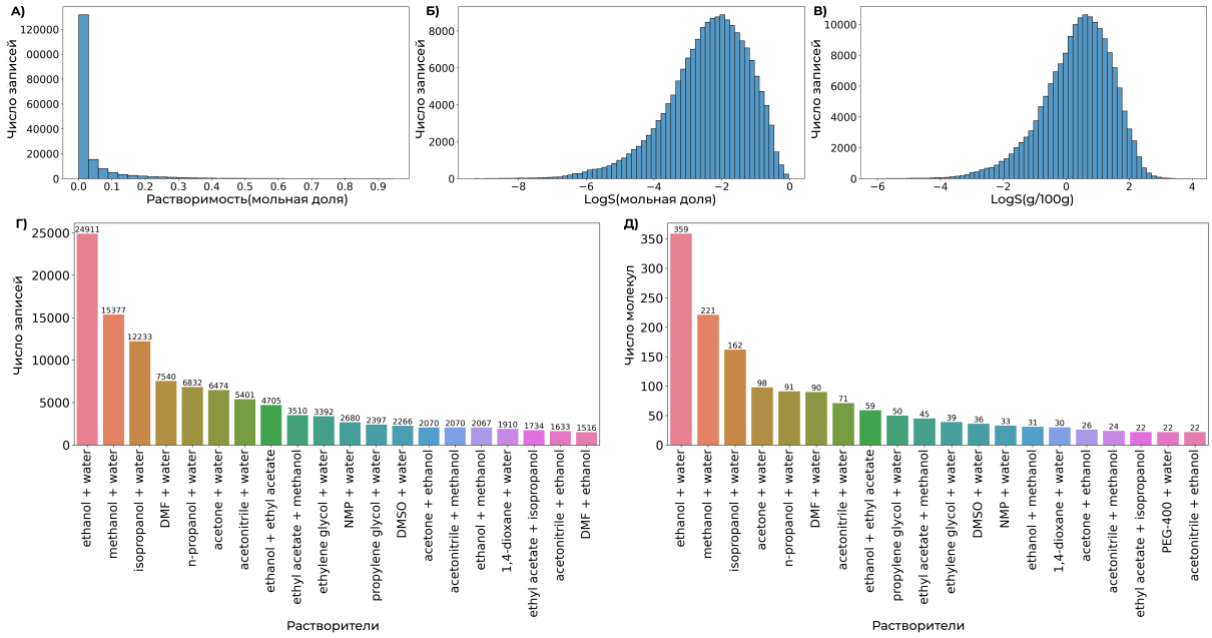


Рис.4. Статистика MixtureSolDB. **А)** Распределение записей по значениям растворимости в единицах мольной доли. **Б)** Распределение записей по основе значений растворимости в логарифмическом масштабе (мольная доля). **В)** Распределение записей на основе значений растворимости в логарифмическом масштабе (г/100 г). **Г)** Распределение пар растворителей по количеству записей (представлены 20 наиболее часто встречающихся пар растворителей). **Д)** Распределение пар растворителей по количеству уникальных растворенных веществ (представлены 20 наиболее часто встречающихся пар растворителей).

Для удобства работы с данными разработан веб-инструмент интерактивного поиска и визуализации на основе библиотеки Streamlit: https://mixturesoldb.streamlit.app/.

3.3. База данных люминесцентных свойств бис-циклометаллированных комплексов иридия(III) IrLumDB и модели машинного обучения для прогнозирования фотофизических свойств

Циклометаллированные комплексы иридия(III) являются одними из наиболее перспективных эмиттеров для фосфоресцентных органических светодиодов (PhOLED), фотосенсибилизаторов и биовизуализирующих агентов, однако рациональный дизайн таких соединений с заданными фотофизическими свойствами требует значительных экспериментальных усилий, а прогностическая точность квантово-химических методов (TD-DFT) остаётся недостаточной. Решение этой задачи методами машинного обучения потребовало формирования базы данных IrLumDB (рис. 5), содержащая 1454 экспериментальных значения люминесцентных свойств для 1287 структурно разнообразных

бис-циклометаллированных комплексов иридия(III), извлечённых из 340 рецензируемых публикаций с использованием поисковой системы Cobalt и последующей ручной экстракцией.

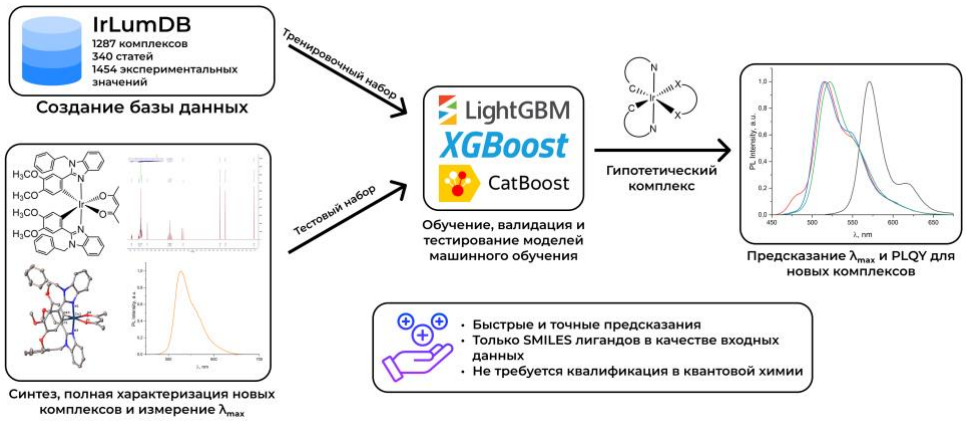


Рис.5. Общая стратегия прогнозирования максимумов люминесценции и квантовых выходов фотолюминесценции (PLQY) бис-циклометаллированных комплексов иридия(III) на основе базы данных IrLumDB и моделей машинного обучения.

В базу данных включены длина волны максимума эмиссии (λ_{max}), квантовый выход фотолюминесценции (PLQY) и время жизни возбуждённого состояния. Для каждой записи указаны растворитель измерения и условия дегазации. База данных охватывает комплексы с эмиссией в диапазоне 450-900 нм и значениями PLQY от близких к нулю до близких к единице, что обеспечивает широкое покрытие пространства люминесцентных свойств в видимой области спектра (рис. 6).

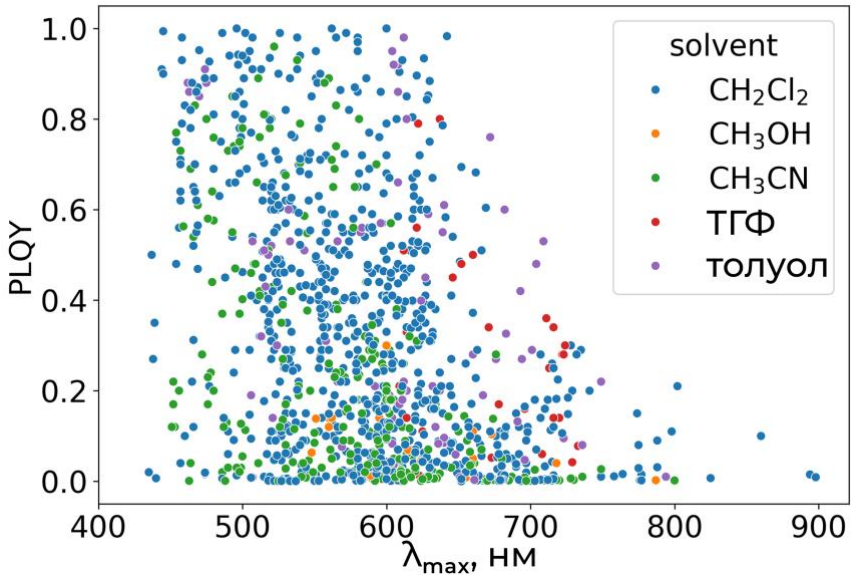


Рис.6. Распределение люминесцентных свойств (квантовый выход фотолюминесценции (PLQY) и длина волны максимума эмиссии (λ_{max}) бис-циклометаллированных комплексов

иридия(III) в базе данных IrLumDB. Показаны данные для пяти наиболее часто используемых растворителей.

Анализ базы данных показал, что большинство комплексов излучает в зелёно-жёлто-оранжевой области (500-650 нм, 1023 из 1454 значений), тогда как в области $\lambda_{max} > 650$ нм наблюдается выраженное смещение в сторону низких значений PLQY. Распределение значений PLQY смещено в сторону низких значений: срединная величина составляет 0.19, а высокоэффективные эмиттеры представлены лишь малой долей записей, что указывает на сохраняющуюся проблему безызлучательной релаксации. Наиболее распространёнными циклометаллированными лигандами являются производные 2-фенилпиридина (ppy, dfppy, pq, piq), а в качестве вспомогательного лиганда с большим отрывом доминирует ацетилацетонат (acac).

Для молекулярного представления комплексов разработан подход, основанный на поэлементном суммировании молекулярных отпечатков лигандов: SMILES лигандов канонизированы с использованием RDKit и преобразованы в отпечатки ECFP4 (2048 бит), а отпечаток комплекса формируется как поэлементная сумма отпечатков его лигандов. Обучение проводилось исключительно на данных в метилене (CH_2Cl_2) как наиболее распространенном растворителе (851 из 1454 записей, 59%). Из обучающей выборки были исключены лиганды, содержащие редко встречающиеся атомы B, Br и Si. Итоговая обучающая выборка для прогнозирования λ_{max} составила 818 записей, для PLQY — 748 записей. Для обучения применялись алгоритмы kNN, SVR, LightGBM, CatBoost и XGBoost с подбором гиперпараметров средствами Optuna (алгоритм TPE) по минимуму RMSE на отложенной выборке.

Наилучшие результаты для прогнозирования λ_{max} показали алгоритмы градиентного бустинга: MAE = 18.3 нм, RMSE = 27.3 нм, $R^2 = 0.84$ на 10-кратной кросс-валидации для LightGBM. В качестве тестовой выборки были синтезированы 12 новых комплексов (рис. 7А-В) и измерены 33 экспериментальных спектра люминесценции на основе арилбензимидазольного скаффолда (рис. 7Г). Наилучший результат на этом тестовом наборе показал CatBoost, MAE = 16.9 нм, что сопоставимо с точностью моделей для органических флуорофоров и превосходит результаты квантово-химических расчётов ΔSCF, которые систематически завышают длину волны эмиссии. Для прогнозирования PLQY достигнуто MAE = 0.129 на 10-кратной кросс-валидации, что сопоставимо с лучшими результатами для органических флуорофоров.

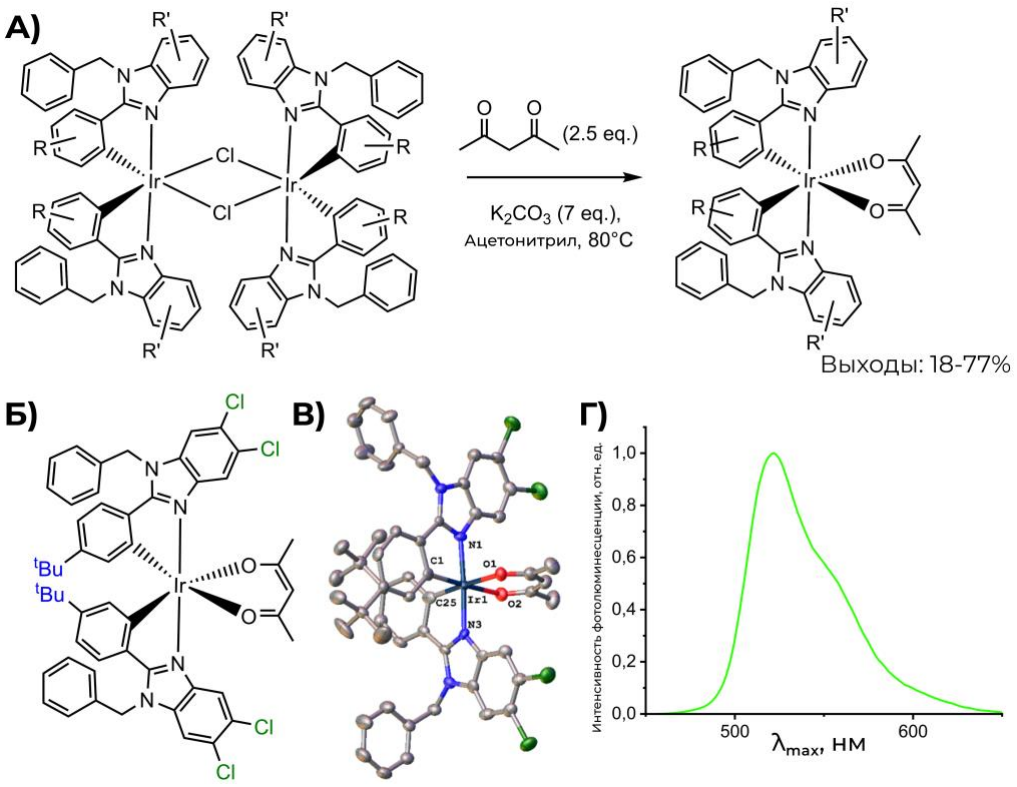


Рис.7. **А)** Общая схема синтеза комплексов для тестовой выборки. **Б)** Молекулярная формула одного из комплексов серии. **В)** Молекулярная структура комплекса по данным РСА. **Г)** Нормированный спектр люминесценции в дегазированном CH_2Cl_2 (25°C).

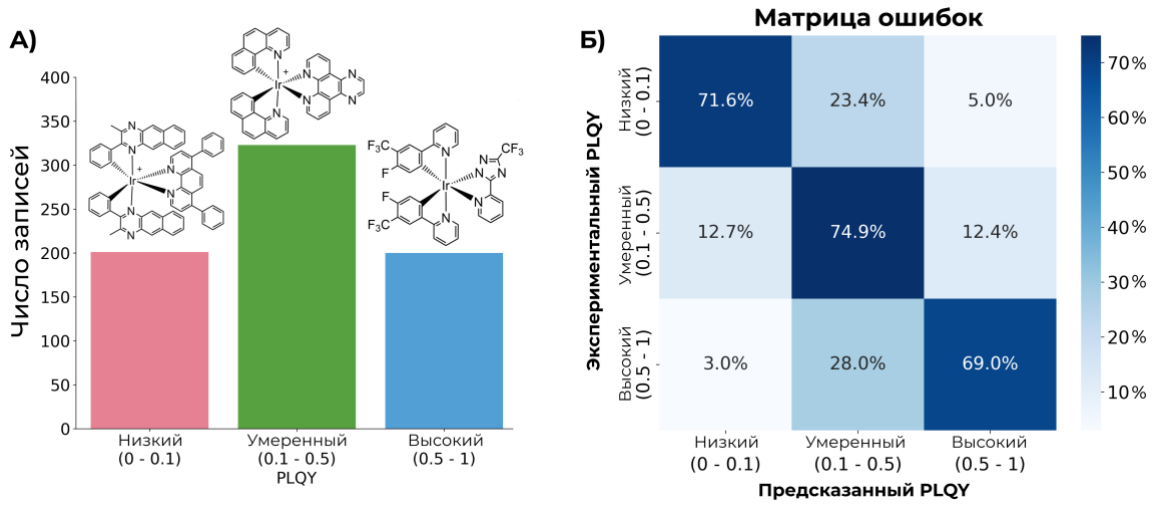


Рис.8. **А)** Распределение комплексов базы данных IrLumDB по классам квантового выхода фотолюминесценции и примеры структур каждого класса. **Б)** Матрица ошибок тернарного классификатора CatBoost на 10-кратной кросс-валидации.

Поскольку для практических применений важнее не точное значение PLQY, а принадлежность комплекса к высоко- или низкоэффективным эмиттерам, был обучен классификатор на основе CatBoost с разбиением на три класса: низкий (0-0.1), умеренный (0.1-0.5) и высокий (0.5-1) PLQY (рис. 8А). Точность (доля верно классифицированных объектов)

классификатора составила 72.4% на 10-кратной кросс-валидации. При этом доля принципиально неверных прогнозов (высокоэффективный комплекс спрогнозирован как низкоэффективный и наоборот) не превышает 5% (рис. 8Б).

Для удобства прогнозирования свойств новых комплексов и навигации по базе данных разработан интерактивный веб-инструмент на основе библиотеки Streamlit: https://irlumdb.streamlit.app/.

3.4. Базы данных цитотоксичности комплексов переходных металлов IrCytoToxDB и MetalCytoToxDB, а также модели машинного обучения для прогнозирования противоопухолевой активности

Комплексы переходных металлов представляют собой перспективный класс потенциальных противоопухолевых агентов, однако экспериментальное определение их цитотоксичности для больших массивов соединений остаётся трудоёмким и дорогостоящим процессом, а данные рассредоточены по сотням публикаций в отсутствие единой структурированной базы.

В качестве первого шага к решению этой проблемы была сформирована база данных IrCytoToxDB — первый систематический датасет о цитотоксичности комплексов переходных металлов в машиночитаемом формате, содержащий 2694 экспериментальных значений IC_{50} для 803 комплексов иридия(III) против 127 клеточных линий, извлечённых из 222 рецензируемых публикаций за период 2008-2022 гг. из 36 ведущих научных журналов.

Химическое пространство базы охватывает два основных структурных типа — бис-циклометаллированные и полусандвичевые комплексы Ir(III). Молекулы представлены в формате SMILES трех лигандов, записанных через точку: L1.L2.L3, с указанием условий эксперимента (клеточная линия, время инкубации, параметры облучения для фотоцитотоксичных соединений).

На основе накопленного методологического опыта и с целью расширения охвата химического пространства была сформирована база данных MetalCytoToxDB, содержащая 26500 значений IC_{50} для 7050 моноядерных комплексов пяти переходных металлов (Ru, Ir, Rh, Re, Os) против 754 клеточных линий, извлечённых из 1921 рецензируемой публикации. Для формирования базы из более чем 50000 статей, первично найденных с помощью поисковой системы Cobalt, по ключевым словам, связанным с цитотоксичностью каждого металла, 1921 статья была отобрана для ручного извлечения данных (рис. 9). Для каждой записи выписаны: значение IC_{50} в темновых условиях, время инкубации, наименование клеточной линии, DOI источника и год публикации. Для фотоцитотоксичных соединений (1873 записи для 757

комплексов) дополнительно извлечены значения IC_{50} под облучением, длина волны, мощность и время облучения. Для части соединений приведены референсные значения IC_{50} для цисплатина как стандартного препарата сравнения.



Рис.9. Общая схема создания MetalCytoToxDB.

Анализ базы данных показал, что доминирующим подмножеством являются комплексы рутения (18895 записей, 71.3%), за которыми следуют комплексы иридия (4500 записей, 17.0%) и значительно менее изученные Rh, Os и Re (1134, 1103 и 868 значений соответственно) (рис. 10А). Среди наиболее изученных клеточных линий выделяются A549, HeLa и MCF-7, каждая из которых представлена более чем 1500 записями (рис. 10Б). Анализ распределения времён инкубации показал, что большинство комплексов инкубировалось с клетками в течение 48 ч (42.5% записей), 72 ч (26.7%) и 24 ч (18.8%) (рис. 10В). Медианная цитотоксичность комплексов Ru и Re заметно выше, чем у Ir, Rh и Os (рис. 10Г).

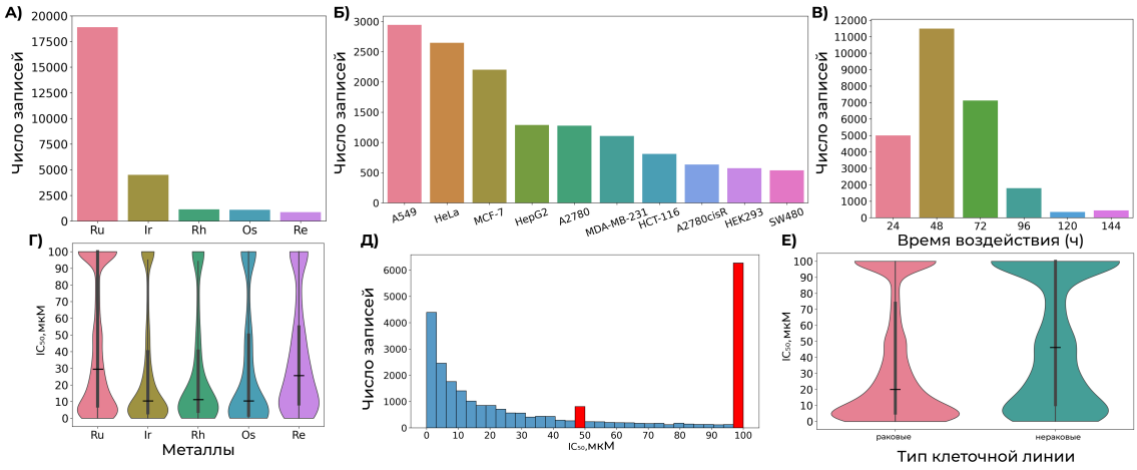


Рис.10. Статистика MetalCytoToxDB. **А)** Количество записей для каждого металла в базе данных. **Б)** Количество записей для 10 наиболее популярных клеточных линий. **В)** Распределение времени воздействия комплексов. **Г)** Скрипичные диаграммы распределения цитотоксичности в рамках каждого подмножества металлов среди топ-50 клеточных линий.

Д) Распределение значений IC_{50}. Столбцы при 50 мкМ и 100 мкМ выделены красным, поскольку они соответствуют дополнительным значениям, представленным в исходных статьях как >50 мкМ и >100 мкМ соответственно. **Е)** Скрипичные диаграммы распределения цитотоксичности на раковых и нераковых клетках среди топ-50 клеточных линий.

Анализ распределения значений IC_{50} среди 50 наиболее представленных клеточных линий показал, что цитотоксичность комплексов против нераковых клеточных линий систематически выше (медиана ~45 мкМ), чем против раковых (медиана ~20 мкМ), что свидетельствует о недостаточной селективности большинства изученных комплексов в отношении опухолевых клеток и создаёт значительный дисбаланс классов при моделировании (рис. 10E). По этой причине нераковые клеточные линии были исключены из обучающей выборки.

Анализ структурных типов выявил, что наиболее распространенным скаффолдом рутениевых комплексов является “piano-stool” геометрия [Ru(η^6-арен)(Cl)(X-Y)] (рис. 11А), тогда как для иридия характерны бис-циклометаллированные [Ir(C-N)$_2$(X-Y)] и циклопентадиенильные [Ir(η^5-Cp)(Cl)(X-Y)] типы (рис. 11Б).

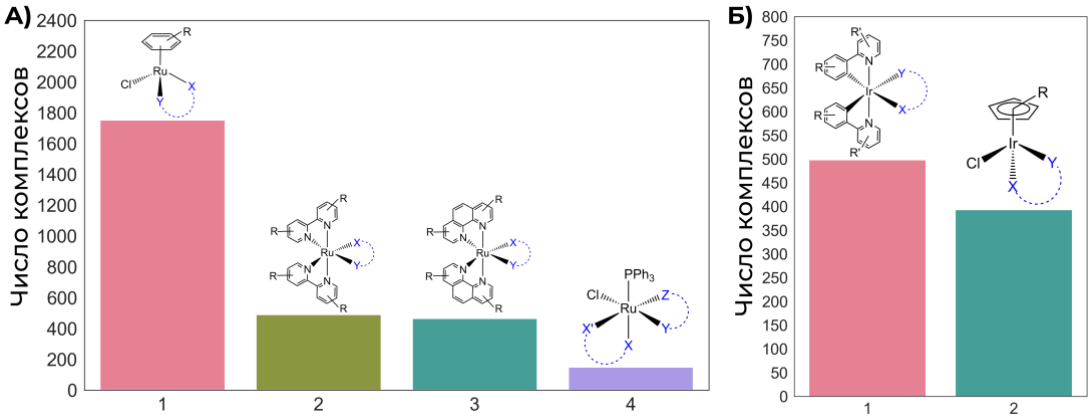


Рис.11. Распределение комплексов по структурным типам: **А)** комплексы рутения. **Б)** комплексы иридия.

Молекулярные структуры комплексов представлены как поэлементная сумма отпечатков ECFP4 (2048 бит) лигандов, что позволяет описывать комплекс через совокупный вклад его составных частей без явного учёта геометрии координационного окружения. В качестве дополнительных признаков использованы заряд комплекса, молекулярная масса, LogP и TPSA суммы лигандов, время инкубации. Клеточная линия кодировалась методом “one-hot” (рис. 12).

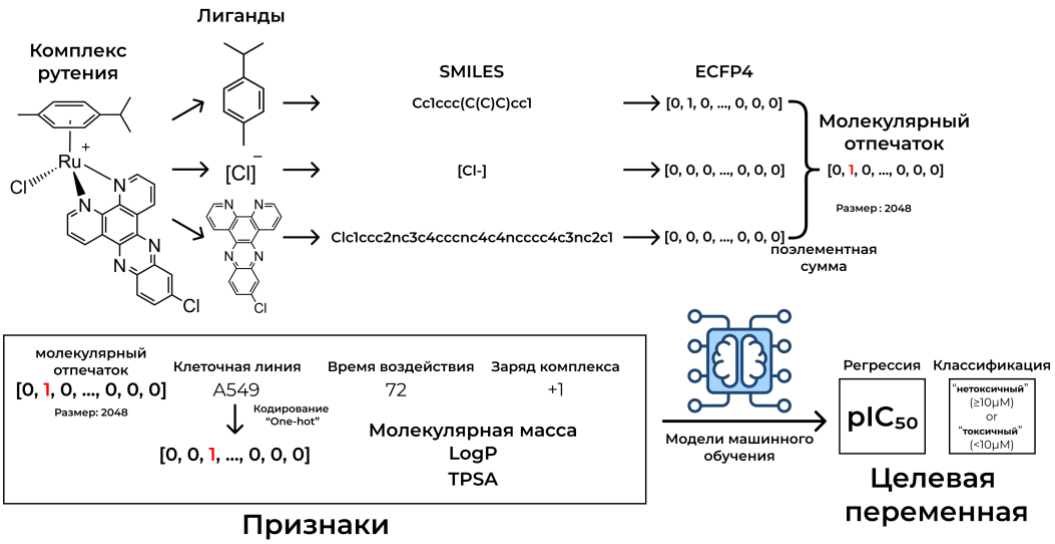


Рис.12. Схема генерации молекулярных отпечатков, признаков и целевой переменной для моделей машинного обучения прогнозирования цитотоксичности комплексов переходных металлов.

Задача прогнозирования цитотоксичности первоначально решалась в режиме регрессии с использованием целевой переменной $pIC_{50} = -\log_{10}(IC_{50}\ [мкМ]) + 6$. Однако регрессия при разбиении по источникам лишь умеренно превосходила предсказание среднего значения (RMSE 0.62 против 0.77). Значительная часть значений приведена в первоисточниках не числом, а предельной оценкой вида “>100 мкМ”, обучение на которых воспроизводило бы условное обозначение отсутствия величины, а достижимая точность ограничена воспроизводимостью самого эксперимента: значения IC_{50} цисплатина (14309 значений из 978 публикаций), измеренные разными научными группами на одной и той же клеточной линии, могут различаться на два-три порядка. В связи с этим основной акцент был сделан на бинарной классификации с порогом IC_{50} = 10 мкМ (“токсичный” / ”нетоксичный”) с применением алгоритмов градиентного бустинга LightGBM, CatBoost и XGBoost с гиперпараметрами по умолчанию. Итоговая обучающая выборка для рутениевой модели составила 11294 записи для 4399 комплексов против 23 раковых клеточных линий.

Для исключения утечки данных между выборками применялась стратегия разбивки по DOI источника, при которой все комплексы из одной публикации попадают исключительно в одну из выборок, обеспечивая реалистичную оценку обобщающей способности модели. Показано, что переход от разбивки по уникальным SMILES к разбивке по DOI источника приводит к снижению ROC-AUC классификатора LightGBM с 0.894 до 0.810 для рутениевого подмножества, что наглядно демонстрирует масштаб завышения метрик при наличии утечки данных.

В качестве негативного контроля модели были обучены на случайно перемешанных метках цитотоксичности. В этом случае ROC-AUC снизился до уровня, близкого к случайному (0.51), что согласуется с наличием действительной зависимости структура-активность в исходных данных; систематических смещений выборки эта проверка сама по себе не исключает. Наилучшие результаты показал классификатор LightGBM: ROC-AUC = 0.810 для рутениевого подмножества и ROC-AUC = 0.730 для иридиевого на 10-кратной кросс-валидации в режиме DOI-разбивки.

Внешняя валидация по временному принципу (обучение на данных до 2024 года, тестирование на данных 2025 года) показала ROC-AUC = 0.74 и долю попаданий 90% в топ-10% соединений после применения фильтра структурной схожести (коэффициент Танимото <0.9), что соответствует более чем двукратному превышению эффективности случайного отбора. Примечательно, что при последовательном увеличении тестовой выборки (охватывая данные с 2021 по 2025 год) ROC-AUC и доля попаданий остаются стабильными, что свидетельствует о высокой стабильности работы модели.

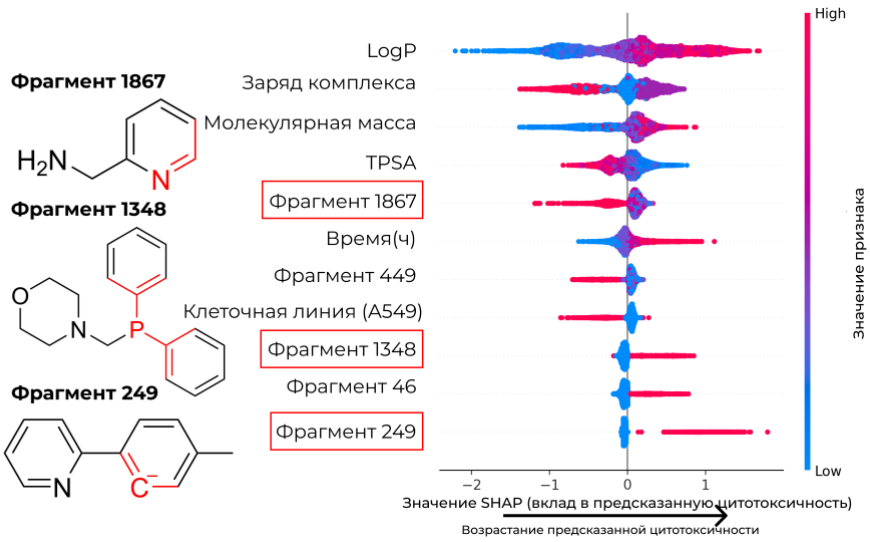


Рис.13. Распределение значений SHAP для классификатора LightGBM на подмножестве рутениевых комплексов (10-кратная кросс-валидация, DOI-разбивка). Красными рамками выделены наиболее значимые структурные фрагменты.

Интерпретируемость моделей исследована методом SHAP. В результате установлено, что наибольший вклад в прогнозирование цитотоксичности рутениевых комплексов вносят LogP, заряд комплекса, молекулярная масса и TPSA. Показано, что рост LogP в целом повышает прогнозируемую цитотоксичность, тогда как рост TPSA снижает, что согласуется с ролью липофильности в клеточной проницаемости. Заряд комплекса влияет нелинейно: катионные комплексы с зарядом (+1) чаще прогнозируются как токсичные, тогда как комплексы с зарядом (+2) — как нетоксичные. Из структурных фрагментов наибольшее

положительное влияние на прогнозируемую цитотоксичность оказывают арилфосфиновые и циклометаллированные фрагменты (рис. 13). Для иридиевой модели аналогичные тенденции сохраняются, при этом наиболее значимыми фрагментами являются NHC-карбеновый и фенильный фрагменты, тогда как имидазофенантролиновый лиганд снижает прогнозируемую цитотоксичность, что согласуется с его склонностью к генерации активных форм кислорода преимущественно под облучением.

Также была разработана мультиметальная модель, включающая металл-специфические признаки (атомный номер, степень окисления, число валентных электронов), что позволяет расширить прогнозирование на подмножества Rh, Os и Re со значительно меньшим размером выборки. Стратегия Leave-One-Metal-Out показала умеренное качество для Rh и Re (ROC-AUC 0.643 и 0.658) и несколько лучшее для Os (0.767) ввиду структурного сходства с Ru. При включении части данных каждого металла в обучающую выборку метрики существенно возрастают до ROC-AUC 0.727-0.789 на 10-кратной кросс-валидации.

Наряду с достигнутыми результатами обозначены ключевые ограничения разработанного подхода. Используемое представление комплексов не учитывает геометрию координационного окружения металла, наличие изомеров и влияние противоиона, что не позволяет напрямую разграничивать структурные изомеры и энантиомеры, известные случаи существенного различия цитотоксичности которых описаны в литературе. Модели применимы исключительно к моноядерным комплексам и не могут корректно обрабатывать как лиганды, содержащие металлические центры (например, ферроценильные производные), так и полиядерные комплексы. Прогнозирование цитотоксичности против нераковых клеточных линий затруднено ввиду систематического дисбаланса классов в данных, следовательно, надежное прогнозирование индексов селективности в рамках текущего набора данных практически невозможно. Для металлов с малым числом записей (Rh, Re) качество мультиметальной модели ограничено. Указанные ограничения определяют перспективы дальнейшего развития подхода: учет геометрии координационного окружения с использованием алгоритмов генерации 3D-структур металлокомплексов, расширение базы данных за счет новых металлов (Pt, Au, Cu и т.д.) и введение механистической информации о действии комплексов.

4. Выводы

1. Разработана и апробирована на массиве более 4000 рецензируемых публикаций унифицированная методология систематического сбора, стандартизации и верификации

экспериментальных данных, применимая 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.

Основное содержание работы изложено в следующих публикациях:

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

5. Список цитируемых источников

1. 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.
2. 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.
3. 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. Т. 48, № 7. С. 1289–1303.
4. Sorkun M. C., Khetan A., Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds // Sci. Data. 2019. Т. 6, № 1. С. 143.
5. Llompart P. и др. Will we ever be able to accurately predict solubility? // Sci. Data. 2024. Т. 11, № 1. С. 303.
6. 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.
7. 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.
8. Jayabharathi J., Thanikachalam V., Thilagavathy S. Phosphorescent organic light-emitting devices: Iridium based emitter materials – An overview // Coord. Chem. Rev. 2023. Т. 483. С. 215100.
9. Ho P.-Y., Ho C.-L., Wong W.-Y. Recent advances of iridium(III) metallophosphors for health-related applications // Coord. Chem. Rev. 2020. Т. 413. С. 213267.
10. Terrones G. G. и др. Low-cost machine learning prediction of excited state properties of iridium-centered phosphors // Chem. Sci. 2023. Т. 14, № 6. С. 1419–1433.
11. Wang S. и др. Multi-level Δ-learning for predicting the radiative decay rate constant of phosphorescent platinum(ii) complexes // J. Mater. Chem. C Mater. 2025. Т. 13, № 10. С. 5189–5199.
12. Abdolmaleki S., Aliabadi A., Khaksar S. Riding the metal wave: A review of the latest developments in metal-based anticancer agents // Coord. Chem. Rev. 2024. Т. 501. С. 215579.
13. Konkankit C. C. и др. Anticancer activity of complexes of the third row transition metals, rhenium, osmium, and iridium // Dalton Transactions. 2018. Т. 47, № 30. С. 9934–9974.
14. Alessio E., Messori L. NAMI-A and KP1019/1339, Two Iconic Ruthenium Anticancer Drug Candidates Face-to-Face: A Case Story in Medicinal Inorganic Chemistry // Molecules. 2019. Т. 24, № 10. С. 1995.
15. 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.

16. 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.

17. 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.

Благодарности.

Автор искренне благодарен научному руководителю, к.х.н. С.И. Беззубову за руководство работой, помощь и поддержку на всех этапах работ.

Автор выражает огромную благодарность заведующему лаборатории кристаллохимии к.х.н. А.В. Чуракову и всему коллективу за помощь в работе.

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.