# Требования к извлечению данных по биологической активности из базы данных ChEMBL v3.0

**Версия:** 2.3+ (2025-10-21) **Репозиторий:** <https://github.com/SatoryKono/bioactivity_data_acquisition/tree/test_refactoring_04>

## Область применения

Выполнить извлечение, нормализацию и контроль качества данных о биоактивности из ChEMBL и вторичных источников; схема звезды; детерминированные артефакты (CSV, QA-отчёты, meta.yaml). **Критерии приемки:** сформированы артефакты согласно §5.2 и §9.2; побайтная идентичность экспорта при повторном прогоне с неизменными входами.

## Статус

Одобрен для тестовой среды. **Критерии приемки:** строки change control заполнены (§0).

## 0. Контроль изменений — CHEMBL-DM02

Вести таблицу изменений протокола с влиянием на данные.

| Этап | Владелец | Дата | Подпись | Влияние на данные |
| --- | --- | --- | --- | --- |
| Подготовлен | Хранитель документации | 2025-10-21 | CHEMBL-DM02 | Уточнены источники, схемы, постобработка, кэш, метрики QA |
| Проверен | Руководитель QA, DataOps | 2025-10-21 | CHEMBL-DM02 | Согласованность, резолюция конфликтов, устойчивость |
| Одобрен | Руководитель управления ДЗ | 2025-10-21 | CHEMBL-DM02 | Одобрено для тестовой среды |

**Критерии приемки:** все поля заполнены; указано влияние на данные.

## 1. Введение

### 1.1 Функциональная цель

Выполнить извлечение, нормализацию и контроль качества данных о биоактивности из ChEMBL с обогащением из PubMed, Semantic Scholar, Crossref, OpenAlex, UniProt, IUPHAR/Guide to Pharmacology, PubChem. Выход: CSV, QA-отчёты, meta.yaml. **Критерии приемки:** перечисленные артефакты созданы по путям §5.2 и §9.2.

### 1.2 Модель данных

Схема звезды: факты activities; измерения documents, targets, assays, test items. Ссылочная целостность обязательна; факты без валидных FK запрещены к публикации. **Критерии приемки:** 0 orphan-строк; все FK валидны.

### 1.3 Источники и контракты

Использовать официальные REST-эндпоинты; соблюдать пагинацию, лимиты, TTL кэша, политику ошибок (§3, §8). **Критерии приемки:** отчёт по API (§3.2) содержит rps\_peak, rps\_95p, бюджеты и статусы ошибок.

### 1.4 Среда выполнения и CI

Зависимости в pyproject.toml, фактические версии в lock-файле. Обязательны джобы CI: lint, unit, integration, schema-validate, export-determinism, thresholds-check. Артефакты CI:

* reports/coverage.xml
* reports/lint.txt
* reports/test\_results.xml
* reports/schema\_validate.json
* reports/export\_determinism.json
* reports/thresholds\_check.json
* reports/ci\_status.json

**Критерии приемки:** все джобы PASSED; при нарушениях порогов §5.1, сортировки (§2.0.2), формата чисел (§5.4), guard-rails (§5.5) — fail CI.

#### 1.4.1 Покрытие тестами

Совокупное покрытие unit+integration ≥ 90%. **Критерии приемки:** при покрытии ниже порога — fail CI (см. reports/ci\_status.json).

## 2. Схемы данных

### 2.0.1 Порядок колонок и hash\_row

Порядок колонок фиксируется спецификацией §2.1–§2.5; новые поля добавлять в конец. Порядок сериализации для hash\_row совпадает с финальным CSV. **Критерии приемки:** стабильная сортировка по PK; побайтная идентичность.

### 2.0.2 Первичные ключи и сортировка

PK: Documents=document\_chembl\_id, Targets=target\_chembl\_id, Assays=assay\_chembl\_id, Test Items=molecule\_chembl\_id, Activities=activity\_id. Экспорт CSV сортировать строго по PK по возрастанию. Нарушения → <stem>\_failure\_cases.csv с sort\_violation. **Критерии приемки:** финальные CSV отсортированы по PK.

### 2.1 Documents

| Столбец | Тип | Ограничения | Описание |
| --- | --- | --- | --- |
| document\_chembl\_id | str | ≠ null, уникальный | Ключ ссылочной целостности |
| doi | str | ∪ ∅, валидный DOI | Нормализация и валидация (§3.4) |
| title | str | ≠ null | Очистка пробелов и кодировки |
| journal | str | ∪ ∅ | Нормализация |
| year | int | ∪ ∅, 1800–(current\_year+1) | Восстановление по §3.4 |
| document\_pubmed\_id | str | ∪ ∅, целое как строка | PMID |

**Критерии приемки:** 100% валидных document\_chembl\_id; year из доверенного источника; doi валиден при наличии.

### 2.2 Targets

| Столбец | Тип | Ограничения | Описание |
| --- | --- | --- | --- |
| target\_chembl\_id | str | ≠ null, уникальный | Связь с assays и activities |
| pref\_name | str | ≠ null | Синонимы унифицированы |
| organism | str | ∪ ∅ | Нормализация таксономии |
| uniprot\_id | str | ∪ ∅, регэксп | UniProt accession без изоформ |

**Регэксп uniprot\_id:**

^(?:[OPQ][0-9][A-Z0-9]{3}[0-9]|[A-NR-Z][0-9][A-Z][A-Z0-9]{2}[0-9])$

Запрещены суффиксы -\d+ (удалить, исходник логировать в outputs/qa/uniprot\_normalization.log). **Критерии приемки:** ≥85% валидных uniprot\_id по паттерну; 0 неоднозначностей (§3.3); 0 суффиксов.

### 2.3 Assays

| Столбец | Тип | Ограничения | Описание |
| --- | --- | --- | --- |
| assay\_chembl\_id | str | ≠ null, уникальный |  |
| target\_chembl\_id | str | ≠ null, FK → targets | Orphan запрещены |
| assay\_type | str | ≠ null | Нормализация словарём ChEMBL |
| bao\_endpoint | str | ∪ ∅ | Мэппинг по словарю |
| description | str | ∪ ∅ | Очистка HTML |
| assay\_parameters | str(JSON) | ∪ ∅, схема §2.3.2 | Биоконтекст |

**Критерии приемки:** bao\_coverage\_pct ≥ 80%; валидный JSON; invalid → failures.

#### 2.3.1 BAO coverage и словари

Справочник config/dictionary/bao\_mapping.csv, контроль bao\_mapping.sha256. **Критерии приемки:** контроль целостности SHA256.

#### 2.3.2 JSON-схема assay\_parameters v1.1

Draft-07, additionalProperties=false; ключи и домены фиксированы. Версия схемы и её SHA256 фиксируются в outputs/qa/assay\_parameters.schema.sha256. Минимальный набор свойств (выдержка, полный список в схеме):

{  
 "$schema": "http://json-schema.org/draft-07/schema#",  
 "title": "assay\_parameters v1.1",  
 "type": "object",  
 "additionalProperties": false,  
 "properties": {  
 "organism\_strain": {"type":"string","minLength":1},  
 "cell\_line": {"type":"string"},  
 "tissue": {"type":"string"},  
 "target\_isoform": {"type":"string","pattern":"^[A-Z0-9]+(?:-[0-9]+)?$"},  
 "variant\_mutation": {"type":"string","pattern":"^[A-Z][0-9]+[A-Z]$"},  
 "temperature\_C": {"type":"number","minimum":0,"maximum":60},  
 "pH": {"type":"number","minimum":0,"maximum":14},  
 "buffer": {"type":"string"},  
 "detection\_technique": {"type":"string"},  
 "readout": {"type":"string"},  
 "replicates\_n": {"type":"integer","minimum":1,"maximum":1000},  
 "concentration\_range": {  
 "type":"object","additionalProperties":false,  
 "properties":{  
 "lower":{"type":["number","null"],"minimum":0},  
 "upper":{"type":["number","null"],"minimum":0},  
 "units":{"type":"string","enum":["pM","nM","uM","mM","M"]}  
 },  
 "required":["units"]  
 }  
 }  
}

**Критерии приемки:** 100% JSON валидны; invalid → assay\_params\_invalid; мягкий KPI полноты: ≥4 заполненных ключа, в отчёте QA.

### 2.4 Test Items

| Столбец | Тип | Ограничения | Описание |
| --- | --- | --- | --- |
| molecule\_chembl\_id | str | ≠ null, уникальный |  |
| canonical\_smiles | str | ∪ ∅ | Синтакс. валидация/канонизация |
| molecule\_type | str | ∪ ∅ |  |
| full\_mwt | float | ∪ ∅, ≥0 | Обновление из PubChem |
| parent\_molecule\_chembl\_id | str | ∪ ∅ | Группировки |

**Критерии приемки:** 0 невалидных SMILES; full\_mwt ≥ 0.

### 2.5 Activities

| Столбец | Тип | Ограничения | Описание |
| --- | --- | --- | --- |
| activity\_id | str | ≠ null, уникальный |  |
| assay\_chembl\_id | str | ≠ null, FK → assays |  |
| molecule\_chembl\_id | str | ≠ null, FK → test items |  |
| target\_chembl\_id | str | ≠ null, FK → targets |  |
| document\_chembl\_id | str | ≠ null, FK → documents |  |
| standard\_value | float | ∪ ∅, ≥0, в nM после конверсии |  |
| standard\_units | str | ∪ ∅, whitelist | См. §2.5.1 |
| pchembl\_value | float | ∪ ∅ | См. §4.6 |
| standard\_relation | str | ∪ ∅, {“=”,“<”,“>”,“≤”,“≥”} | Управляет интервалами |
| interval\_lower | float | ∪ ∅ | Нижняя граница (nM) |
| interval\_upper | float | ∪ ∅ | Верхняя граница (nM) |

#### 2.5.1 Единицы и конверсии к nM

Входной регэксп:

^(pM|nM|uM|µM|μM|mM|M)$

Нормализовать µ/μ → u; итоговый домен: {"pM","nM","uM","mM","M"}. Конверсии: pM→0.001; nM→1; uM→1000; mM→1\_000\_000; M→1\_000\_000\_000. Запрещённые единицы → unit\_unknown, запись исключить из финала. Whitelist endpoints (регистрозависимо): {"IC50","Ki","Kd","EC50","AC50","pIC50","pKi"}; прочее → endpoint\_blacklisted. **Критерии приемки:** 100% единиц в домене; нет сырых µ/μ в финале; интервалы согласованы; pchembl\_value только для "=" и "≤".

## 3. Источники данных и интеграции

### 3.1 Таблица источников

Структура без изменения семантики; поля источников, базовые URL, параметры, ключи аутентификации при необходимости. **Критерии приемки:** таблица присутствует в документации и в отчёте постобработки.

### 3.2 SLA и лимиты

Тайм-ауты: connect=5s, read=20s, total=25s. Ретраи: 5xx/сеть — до 3; 429 — до 5; backoff t\_k=min(2^k,60)s, jitter ±10%. Параллелизм: ≤8 на источник, ≤24 глобально. RPS: ChEMBL≤5; PubMed≤3; UniProt≤3; Crossref≤3; OpenAlex≤3; PubChem≤5; IUPHAR≤2. Бюджеты: ChEMBL≤100k; PubMed≤50k; UniProt≤30k; Crossref≤30k; OpenAlex≤30k; PubChem≤50k; IUPHAR≤10k. Красные зоны: ≥90% warn; >100% fail. 4xx — immediate fail; 5xx/429 — после ретраев fail. Отчёт: outputs/qa/api\_usage.report.json со схемой:

{ "source": "...", "calls\_used": 0, "budget": 0, "rps\_peak": 0.0, "rps\_95p": 0.0,  
 "retries\_total": 0, "http\_2xx": 0, "http\_4xx": 0, "http\_5xx": 0, "http\_429": 0 }

RPS считать на окне 60 s, шаг 1 s. **Критерии приемки:** бюджеты и RPS не превышены; отчёт валиден.

### 3.3 ChEMBL↔UniProt mapping

Приоритет reviewed > unreviewed; обязательный match по organism; при множественных reviewed с одинаковым organism — выбрать запись с максимальным updated\_date (UniProt). Невозможность однозначного выбора → uniprot\_ambiguous. targets.uniprot\_id без -\d+; удалённый суффикс логировать в outputs/qa/uniprot\_normalization.log. **Критерии приемки:** ≥85% targets с валидным uniprot\_id; 0 неоднозначностей.

### 3.4 Конфликт-резолюция документов

Приоритет: PubMed > Crossref > OpenAlex > Semantic Scholar. Ключи сопоставления: DOI → PMID → title по Jaccard ≥ 0.60 (lower-case, без пунктуации). Нормализация DOI: lower-case; удалить doi:, https://doi.org/, пробелы. Валидация DOI:

(?i)^10\.\d{4,9}/[-.\_;()/:A-Z0-9]+$

Отчёт: outputs/qa/document\_conflicts.jsonl (JSON Lines) с полями и chosen\_source. **Критерии приемки:** conflict\_rate ≤ 0.05 (до 4 знаков, half-even); при превышении — fail.

### 3.5 Кэширование

.cache/{source}/, TTL: ChEMBL=1d; PubMed/Crossref/OpenAlex=7d; UniProt/IUPHAR=14d; PubChem=7d. Ключ кэша: SHA256(base\_url + "?" + normalized\_query); для POST — хеш отсортированного JSON. Цель cache\_hit\_rate ≥ 95% при повторном запуске в пределах TTL. **Критерии приемки:** достигнут hit rate; TTL соблюдён.

## 4. Потоки постобработки

### 4.1 Общая цепочка

raw → normalized → final: нормализация, обогащение, валидация, дедуп по hash\_row, производные, логирование изменений. **Критерии приемки:** слои и журналы присутствуют; отчёты собраны.

### 4.2 Documents

Тримминг, кодировка; year → int; проверка DOI; восстановление year (§3.4); фильтр предаторских журналов config/dictionary/blacklist\_journals.txt (SHA256). Финал: hash\_row, дедуп, отчёт. **Критерии приемки:** отчёт отказов; статистика дедупликации; blacklisted\_count и примеры.

### 4.3 Targets

Нормализация имён и таксономии; валидация формата UniProt; обогащение синонимами/классами; приоритет UniProt; 4xx логируются; финал: hash\_row, отчёт покрытия. **Критерии приемки:** пороги §5.1 соблюдены; изоформ-суффиксы отсутствуют.

### 4.4 Test Items

Нормализация, валидация FK; SMILES: синтаксическая валидация и канонизация; очистка соль/гидрат по словарю config/dictionary/salts\_hydrates.csv (SHA256). Невалидность → smiles\_invalid. Финал: hash\_row, дедуп. **Критерии приемки:** 0 smiles\_invalid; протокол дельт присутствует.

### 4.5 Цензура и интервалы

Норма v в nM.

* "=" → [v, v]
* "<" → (NULL, v)
* ">" → (v, NULL)
* "≤" → (NULL, v]
* "≥" → [v, NULL) Бесконечности не записывать; вместо ±∞ использовать NULL. **Контроль предикатами:**

(interval\_lower IS NULL OR interval\_lower >= 0)  
AND (interval\_upper IS NULL OR interval\_upper >= 0)  
AND (interval\_lower IS NULL OR interval\_upper IS NULL OR interval\_lower <= interval\_upper)

Нарушения → interval\_inconsistent. **Критерии приемки:** 100% согласованность relation ↔ интервалы; 0 случаев interval\_lower > interval\_upper.

### 4.6 Расчёт pChEMBL и лог-метрики

Если вход в pIC50/pKi: value\_nM = 10^(9 − pValue). pchembl\_value = −log10(value\_nM × 1e−9); заполнять для "=" и "≤" (для "≤" использовать interval\_upper). Округление half-even до 2 знаков; диапазон [0;20]; вне диапазона → pchembl\_out\_of\_range. **Критерии приемки:** отсутствует pchembl\_value при relation ∈ {“<”,“>”,“≥”}; диапазон соблюдён.

### 4.7 Детерминированный экспорт

Кодировка UTF-8; LF; разделитель ,; кавычки "; NA → "". Числа рендерить как "%.6g"; запрет экспоненты; локаль “C”; до 6 значащих цифр. Сортировка по PK; фиксированный порядок колонок. **Критерии приемки:** побайтная идентичность; отсутствие [eE] в числовых полях; reports/export\_determinism.json = PASSED.

### 4.8 hash\_row

Сериализация значений в порядке финального CSV; разделитель для хеша |; экранирование внутренних | как \|; NULL → ""; числа как в §4.7. hash = sha256(utf8(serialized\_line)) в hex нижним регистром. Коллизии → failures hash\_collision с serialized\_line. **Критерии приемки:** 0 коллизий; протокол коллизий присутствует.

## 5. Обеспечение качества и валидация

### 5.1 Пороговые метрики (единый источник)

configs/qa\_thresholds.yaml (приоритет над текстом).

| Метрика | Documents | Targets | Assays | Test Items | Activities |
| --- | --- | --- | --- | --- | --- |
| Валидные строки после дедупа | >0 | >0 | >0 | >0 | >0 |
| Пропуски в ключевых полях | <10% | <15% | <20% | <5% | <10% |
| Дубликаты (hash\_row) | 0% | 0% | 0% | 0% | 0% |
| Покрытие нормализаций/обогащ. | >90% | >85% | >80% | >95% | >90% |
| BAO coverage (assays) | — | — | ≥80% | — | — |

**Критерии приемки:** reports/thresholds\_check.json = PASSED; при нарушении — fail CI.

### 5.2 Отчётность и пути артефактов

* Финал: outputs/final/{documents|targets|assays|test\_items|activities}.csv
* QA-таблицы: outputs/qa/<stem>\_quality\_report\_table.csv
* Пост-отчёты: outputs/qa/<stem>.postprocess.report.json
* Failures: outputs/failures/<stem>\_failure\_cases.csv
* Метаданные: outputs/meta/meta.yaml, outputs/meta/meta.sha256
* API-отчёт: outputs/qa/api\_usage.report.json
* Конфликты документов: outputs/qa/document\_conflicts.jsonl
* SHA256 ресурсов словарей/схем — рядом с ресурсами **Критерии приемки:** все файлы существуют; JSON валиден; SHA256 совпадает.

### 5.3 Дедупликация

По hash\_row; при необходимости — по бизнес-ключам. Коллизии — см. §4.8. **Критерии приемки:** 0% дубликатов.

### 5.4 Формат чисел

Без экспоненты; ≤6 значащих цифр; локаль “C”; round-half-to-even. **Критерии приемки:** отсутствуют [eE] в CSV.

### 5.5 Guard-rails объёма

Если row\_count финальной таблицы снизился строго более чем на 10.0% относительно ближайшего предыдущего успешного релиза (meta.yaml) — fail CI; отчёт reports/rowcount\_regression.json. Допустимы waiver в configs/waivers.yaml с полями {id, reason, scope, expires\_at, approver}. **Критерии приемки:** либо соблюдено, либо waiver валиден.

### 5.6 Таксономия failure codes

Словарь в configs/failure\_codes.yaml с severity ∈ {blocker, major, minor} и configs/failure\_codes.sha256. Минимальный набор: unit\_blacklisted, unit\_unknown, hash\_collision, orphan\_fk, assay\_params\_invalid, smiles\_invalid, rate\_limited, http\_5xx\_exhausted, http\_4xx, uniprot\_ambiguous, doi\_invalid, bao\_missing, endpoint\_blacklisted, cache\_key\_conflict, sort\_violation, pchembl\_out\_of\_range, interval\_inconsistent. **Критерии приемки:** 100% строк в failures имеют код из словаря; SHA256 совпадает.

## 6. Локализация и документация

docs/en/PROTOCOL\_EN.md, docs/ru/PROTOCOL\_RU.md — источники истины; генерация DOCX из Markdown; бинарники не коммитить. Русские артефакты — UTF-8. **Критерии приемки:** Markdown синхронизированы; DOCX сгенерированы; UTF-8 валиден.

## 7. Журнал изменений

| Версия | Дата | Автор | Ключевые обновления |
| --- | --- | --- | --- |
| 2.2 | 2025-10-01 | Хранитель документации | Усилены схемы и постобработка; локализация |
| 2.3 | 2025-10-21 | Хранитель документации | Среда выполнения, TTL-кэш, валидаторы, raw/normalized, единицы/pChEMBL, BAO, QA, FK |

**Критерии приемки:** строки отражают изменения и влияние на данные; хронология монотонна.

## 8. Операционные параметры

Логирование: JSON Lines {ts, run\_id, stage, source, entity, key, level, message, http\_status, retry\_k, cache\_hit, rps\_bucket}. Seed: 0. SLA/RPS/TTL см. §3.2 и §3.5. **Критерии приемки:** бюджеты и RPS соблюдены; метрики в отчётах; cache\_hit\_rate ≥ 95%.

## 9. Метаданные релиза

### 9.1 Структура meta.yaml

Обязательные поля: pipeline\_version (semver), chembl\_release (строка), row\_count, checksums, run\_id (UUIDv4), started\_at, finished\_at, current\_year, chembl\_release\_source ∈ {cli,status}. Валидировать по JSON Schema v1.0 (преобразуя YAML в JSON), с правилом finished\_at ≥ started\_at. **Критерии приемки:** meta.yaml валиден схеме.

### 9.2 Пути и контроль

outputs/meta/meta.yaml, outputs/meta/meta.sha256 (hex от meta.yaml). **Критерии приемки:** все поля заполнены; checksums соответствуют; схема зелёная.

## Приложение A. Справочные ссылки

* ChEMBL Data Web Services
* NCBI E-utilities
* Crossref REST Works
* OpenAlex Works

**Критерии приемки:** ссылки перечислены; не влияют на CI.