🟩 Готовый текст для отчёта: «Метрики блока 2: научно-технический анализ»

Блок 2 решает задачу **оценки научного потенциала и токсичности кандидатов**, используя признаки, которые обычно применяются в доклинических исследованиях и in-silico моделировании.
Ниже приведено подробное описание всех параметров, которые используются для построения моделей `potential_label` и `toxicity_label`.

📌 1. indication — терапевтическая область (заболевание)

Что это:
Тип заболевания, против которого направлено действие препарата (lung_cancer, melanoma и т.д.)

В реальности:
Эта информация извлекается из описаний статей, патентов или ClinicalTrials.gov по ключевым словам.

Зачем нужно:
Позволяет учитывать различия между болезнями — некоторые направления имеют высокие требования к эффективности (например, онкология).

📌 2. target — биологическая мишень молекулы

Что это:
Название таргетного белка или сигнального пути (EGFR, PD1, VEGF и т.п.)

В реальности:
Извлекается из абстрактов PubMed, патентов и аннотаций малекул (часто из раздела "Mechanism of Action").

Зачем нужно:
Разные таргеты имеют различную успешность лечения и типичные профили токсичности.
Например, ингибиторы EGFR чаще имеют кожные побочки, PD1 — иммуно-побочные эффекты.

📌 3. has_positive_efficacy_phrase — упоминания об эффективности в текстах

Что это:
Бинарный признак (0/1), показывает встречаются ли в публикации фразы типа:

- “significant tumor reduction”
- “strong in vitro efficacy”
- “improved survival rate”

В реальности:
Вытаскивается через NLP-фильтры, keyword search, weak supervision.

Зачем нужно:
Это прямой индикатор того, что в исследованиях есть позитивные данные.

📌 4. has_severe_toxicity_phrase — упоминания о серьёзной токсичности

Что это:
Бинарный признак (0/1), отражающий негативные фразы:

- “severe cardiotoxicity”
- “dose-limiting toxicity”
- “high hepatotoxic response”

В реальности:
Ищется в доклинических отчётах, статьях, неожиданных сигналах безопасности.

Зачем нужно:
Это прямой индикатор риска, который снижает потенциал препарата.

📌 5. text_embed_score — релевантность исследования (эмбеддинг статьи)

Что это:
Числовой показатель качества/релевантности научной публикации (0–1).

В реальности:
Вычисляется как косинусная близость между эмбеддингом статьи (SPECTER, SciBERT) и статьями-эталонами или патентами.

Зачем нужно:
Позволяет определить, насколько исследование соответствует текущим научным трендам и является ли оно "сильным".

📌 6. molecular_weight — молекулярная масса (в дальтонах)

Что это:
Фундаментальный химический параметр.

В реальности:
Вычисляется из молекулярной структуры (SMILES → RDKit: `Descriptors.MolWt`).

Зачем нужно:
Масса влияет на:

- проникновение в клетку
- метаболизм

- токсичность

Большие (>550 Da) молекулы часто хуже всасываются.

📌 7. logP — липофильность (октанол/вода)

Что это:
Показывает, насколько молекула любит жиры vs воду.

В реальности:
Считается из структуры (RDKit → `Crippen.MolLogP`).

Зачем нужно:

- logP > 4 → повышенная токсичность
- logP < 1 → плохая проницаемость

Это один из важнейших ADMET-параметров.

📌 8. potential_label — целевой ярлык (0,1,2)

Что это:
Классификация научного потенциала:

- 0 — низкий
- 1 — средний
- 2 — высокий

Как формируется:
Из фичей эффективности, токсичности, logP, массы и качества публикаций.

Зачем нужно:
Это главный таргет блока 2 для ML-модели.

📌 9. toxicity_label — бинарная токсичность (0/1)

Что это:
Логическая оценка токсичности:

- 0 — безопасно
- 1 — высокая токсичность

Как формируется:
Зависит от:

- токсичных фраз в тексте
- logP
- молекулярной массы

Зачем нужно:
Это второй таргет для модели ML.

🟩 Финальное резюме для блока 2

Признаки блока 2:

- часть извлекается из текста (indication, target, efficacy/toxicity phrases, text_embed_score)
- часть является химическими (molecular_weight, logP)
- блок 2 использует ML, чтобы предсказать **potential_label** и **toxicity_label**
- эти предсказания потом используются в блоке 3 (бизнес-аналитика)