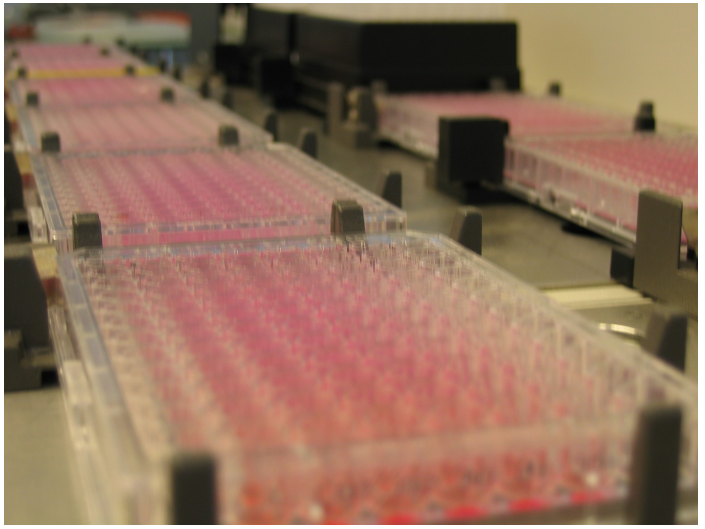
Виртуальный скрининг

Высокопроизводительный скрининг



Поиск активных молекул
в библиотеках реальных
соединений — один из ключевых
методов идентификации ведущих
соединений. Хотя бы каких-нибудь.

Высокопроизводительный скрининг: недостатки...

- ✗ Труднодоступность для академических исследователей;
- ✗ Высокая стоимость и низкая окупаемость, значительные издержки на синтез сотен тысяч соединений;
- ✗ Разработка автоматизированных методик не всегда возможна;
- ✗ Низкая вероятность обнаружения активных соединений в «слепом» варианте метода (без предварительной подготовки выборки) — не более 0,5%.

... и достоинства

- ✓ Надёжность получаемых результатов;
- ✓ Возможность получения данных по соотношениям «структура—активность»;
- ✓ Высокая эффективность при правильном построении библиотек.

Определение

Виртуальный скрининг – процедура *фильтрации* библиотеки соединений с целью нахождения молекул, *могущих* обладать требуемой биологической активностью.

Суть и задача

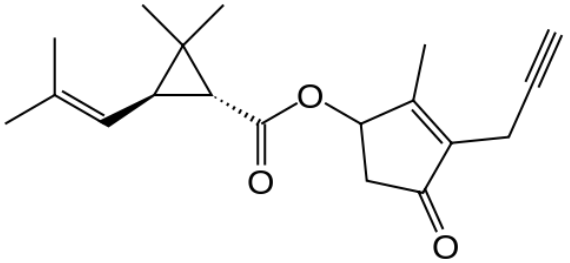
- Ранжировать «хорошие» соединения («хиты») выше, чем «плохие».
- Обогатить библиотеку соединений предположительно активными.

Зачем?

Для сокращения затрат на высокопроизводительный скрининг.

Оценка (Score)

Структура (Structure)



Дескриптор (Descriptor)

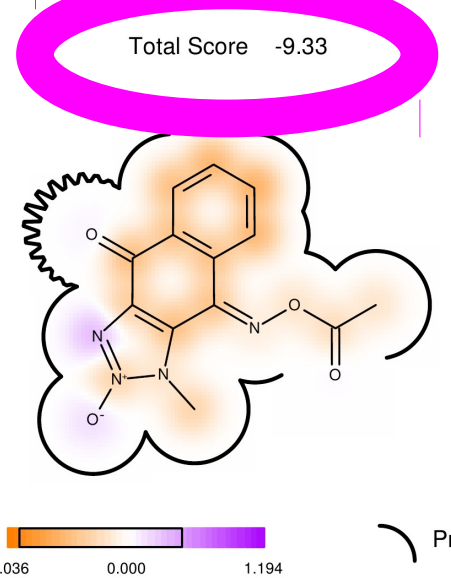
$$descriptor = f(structure)$$

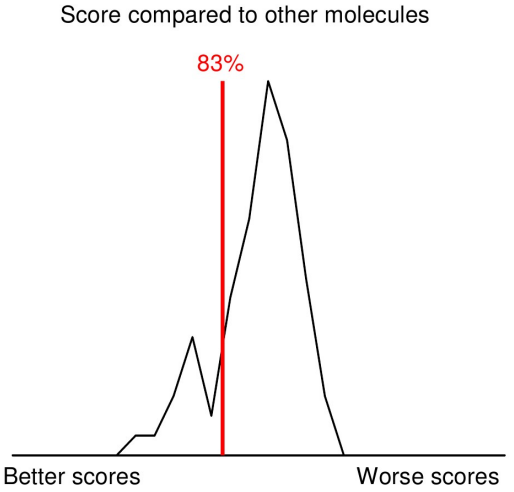
Оценка (Score)

$$score = g(descriptors)$$

Оценка — критерий *ранжирования* соединений

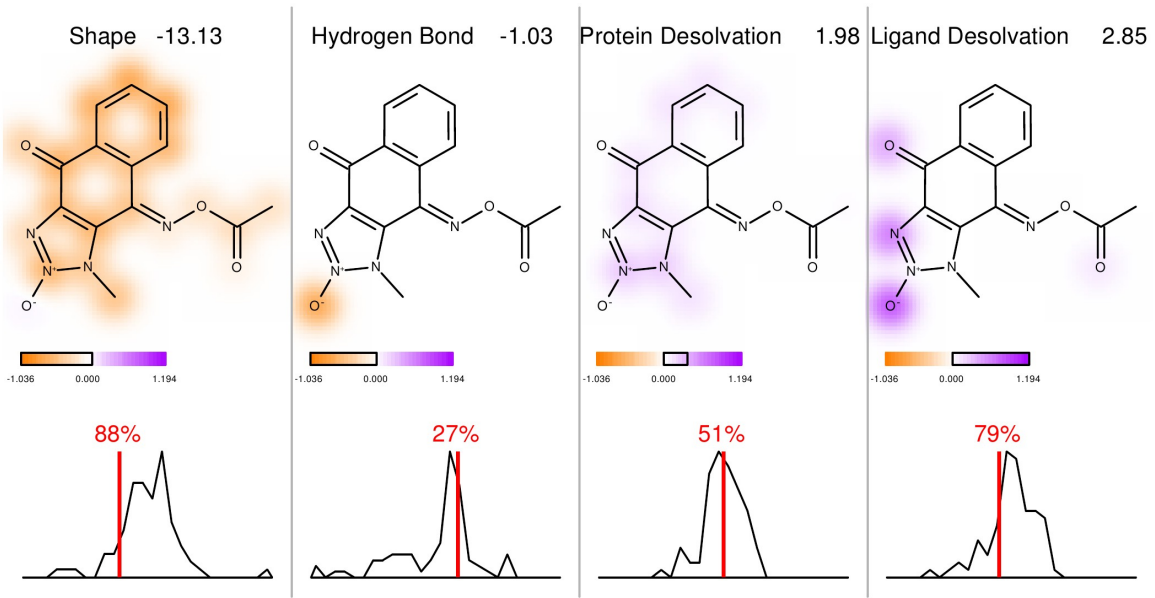
Molecule Name	shtil_20130625
Molecular Weight	286.2
XLogP	2.2
PSA	100.5
Heavy Atoms	21
Acceptor Count	5
Donor Count	0
Chelator Count	1
mmff94s_NoEstat	143.20
Pose Index	11





Residue Fingerprint

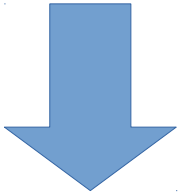
ALA49A	ALA75A
ASN62A	ASP76A
ASP9A	GLN61A
GLU63A	GLY29A
ILE28A	LEU64A
LEU8A	LIG2
LIG3	LYS51A
PHE33A	PHE77A
PRO12A	THR14A
TYR10A	**VAL11A**
VAL36A	VAL76A



Acceptor	Donor
Metal	Contact

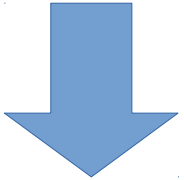
Методология виртуального скрининга

Библиотека доступных соединений



Неспецифическое фильтрование (препроцессинг)

Соединения, обладающие приемлемыми свойствами (нетоксичные, небольшие и пр.)



Специфическое фильтрование

Потенциально активные соединения

Визуальный анализ (постпроцессинг)

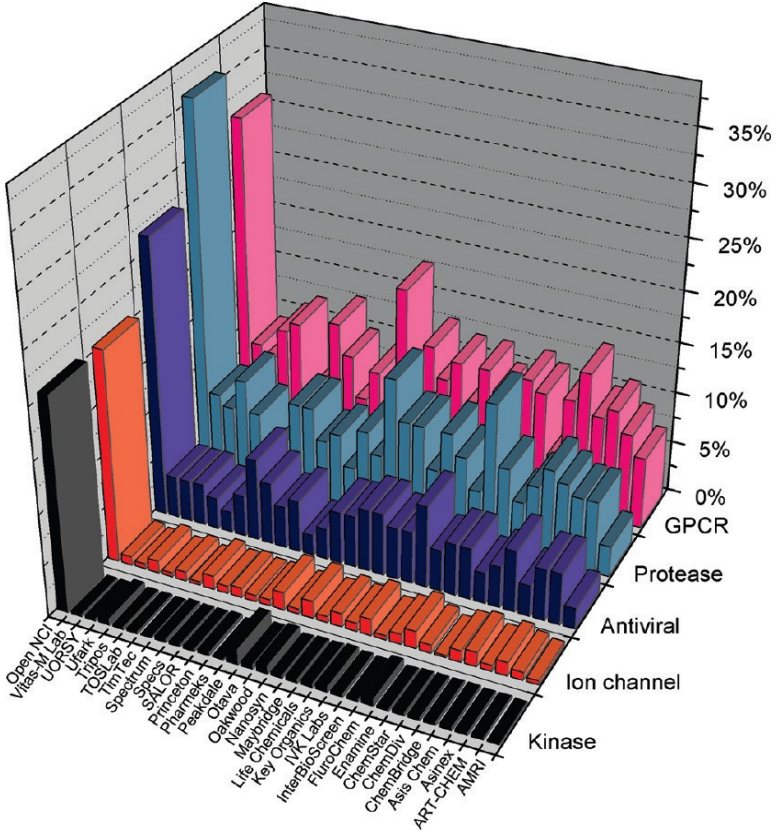
Фокусированная библиотека потенциальных лигандов

Библиотеки

- **ZINC** (http://zinc.docking.org) — 21 млн. соединений, большинство коммерчески доступно;
- **GDB-17** (http://www.gdb.unibe.ch) — 166,4 млрд. молекул, результат генерации;
- Библиотеки поставщиков (ChemDiv, Asinex, etc.) — разное количество, зачастую много билдинг-блоков;
- Библиотеки природных соединений;
- Комбинаторные библиотеки.

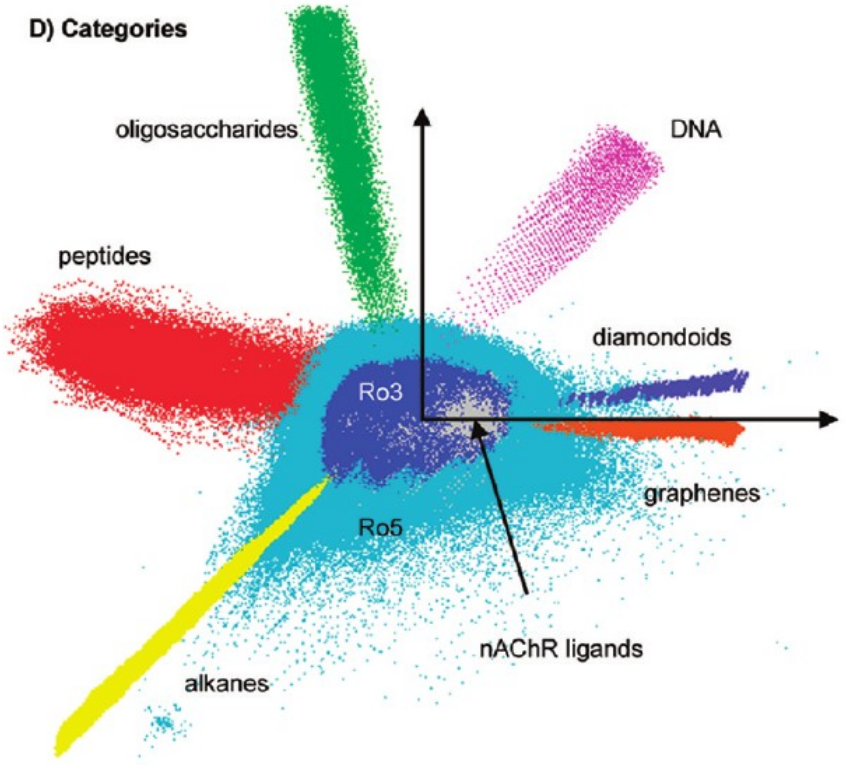
Поскольку соединений в библиотеках много, требуется предварительная подготовка библиотек: ликвидация информационного шума *и* адекватное представление молекул.

Распределение типов активности в библиотеках



Chuprina *et al.*, *JCIM,* **2010,** *50,* 470-479

Распределение типов структур в химическом пространстве



Reymond & Awale, *ACS Chem. Neurosci.,* **2012,** *3,* 649-657

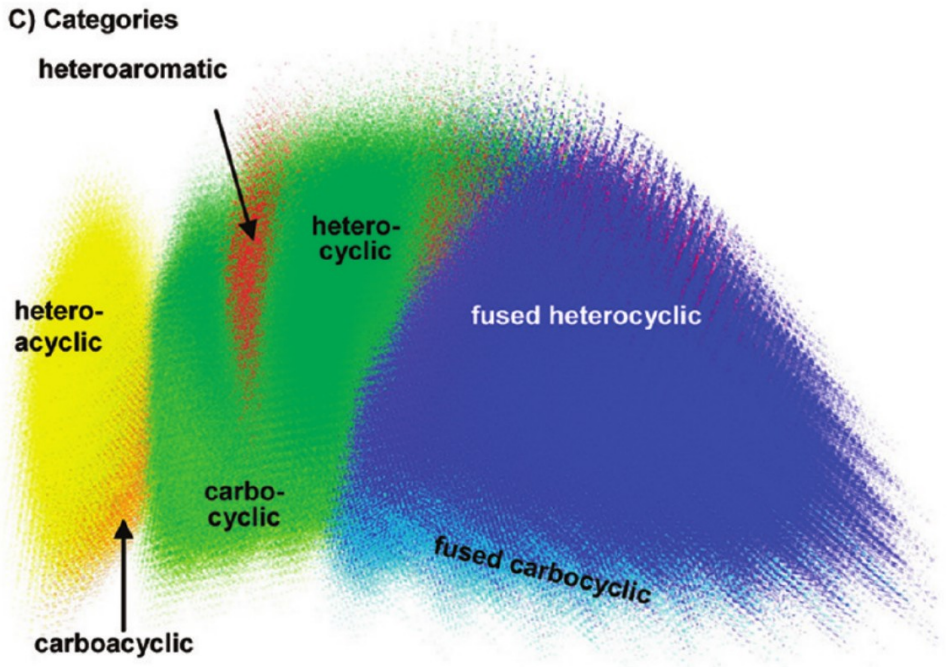
Масштаб бедствия

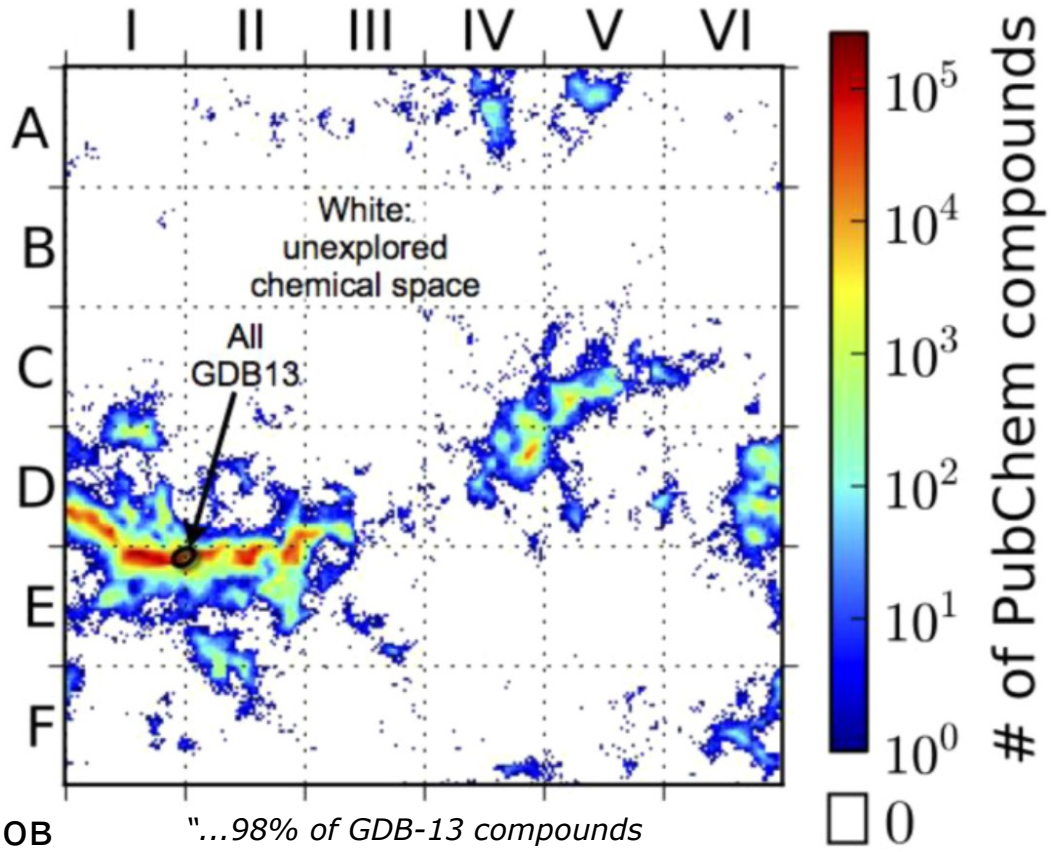
- **DrugBank:** 2200 допущенных лекарств, 8200 записей
- **PubChem:** $19{,}2 \cdot 10^6$ соединений
- **ChemSpider:** $>43 \cdot 10^6$ структур
- **ZINC15:** $100 \cdot 10^6$ структур, готовых для докинга
- **GDB-13:** $977 \cdot 10^6$ структур
- **GDB-17:** $166{,}4 \cdot 10^9$ молекул
- **Удовлетворяют правилам Липински:** $\sim 10^{33}$

Polishchuk *et al.*, *JCAMD,* **2013,** *27,* 675-679

Химическое пространство большое. Очень большое.

GDB-13





10^9 молекул — не больше 13 тяжёлых атомов

"...98% of GDB-13 compounds were assigned to just 10 neurons, and the GDB-13 compounds overall occupy a total of only 61 adjacent neurons, 0.07% of the total."

Reymond & Awale, *ACS Chem. Neurosci.*, **2012,** *3,* 649-657
Virshup *et al., JACS,* **2013,** *135,* 7296-7303

Инструменты предварительного фильтрования

- Физико-химические свойства (правила Липински, правила Опреа — “lead-likeness” и “drug-likeness”)
- Подструктурные фильтры (PAINS, REOS, правила Eli Lilly)
- Прогноз ADMET (QikProp, VolSurf)

Позволяют избавиться от соединений, которые заведомо не пройдут на дальнейшие стадии поиска

См. также: http://www.click2drug.org/directory_ADMET.html

«Правила пяти» Липински

Биодоступность оптимальна, если:

 H-bond donors < 5

 MW < 500

 Log*P* < 5

 H-bond acceptors < 10

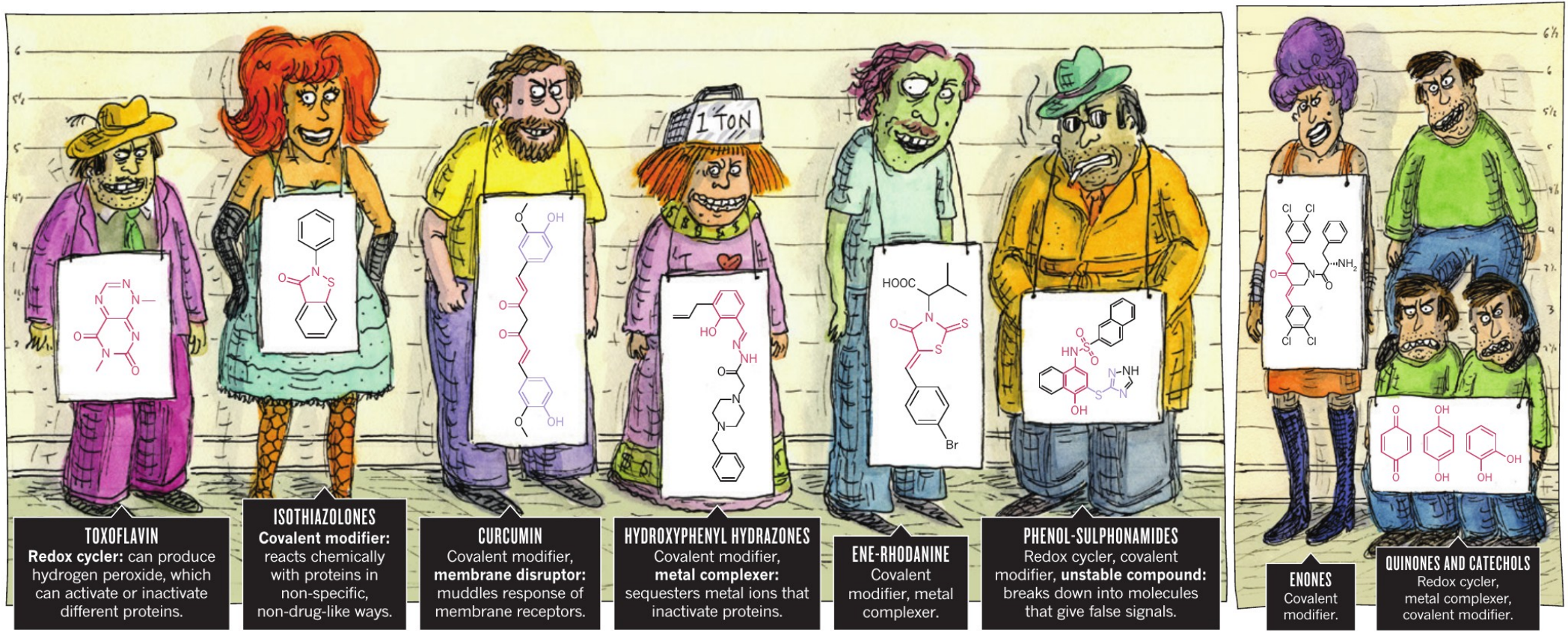
Крис Липински

Критерии «lead-likeness»

- Rotatable bonds < 10
- Rings > 0
- Chiral centers < 3
- Ограничения на число атомов различных типов, число гетероатомов, число связей между циклами и пр. Хватило бы фантазии.

WORST OFFENDERS

Pan-assay interference compounds (PAINS) fall into hundreds of chemical classes, but some groups occur much more frequently than others. Among the most insidious are the eight shown here (reactive portions shown in red and purple). These and related compounds should set off alarm bells if they show up as 'hits' in drug screens.



TOXOFLAVIN
Redox cycler: can produce hydrogen peroxide, which can activate or inactivate different proteins.

ISOTHIAZOLONES
Covalent modifier: reacts chemically with proteins in non-specific, non-drug-like ways.

CURCUMIN
Covalent modifier, **membrane disruptor:** muddles response of membrane receptors.

HYDROXYPHENYL HYDRAZONES
Covalent modifier, **metal complexer:** sequesters metal ions that inactivate proteins.

ENE-RHODANINE
Covalent modifier, metal complexer.

PHENOL-SULPHONAMIDES
Redox cycler, covalent modifier, **unstable compound:** breaks down into molecules that give false signals.

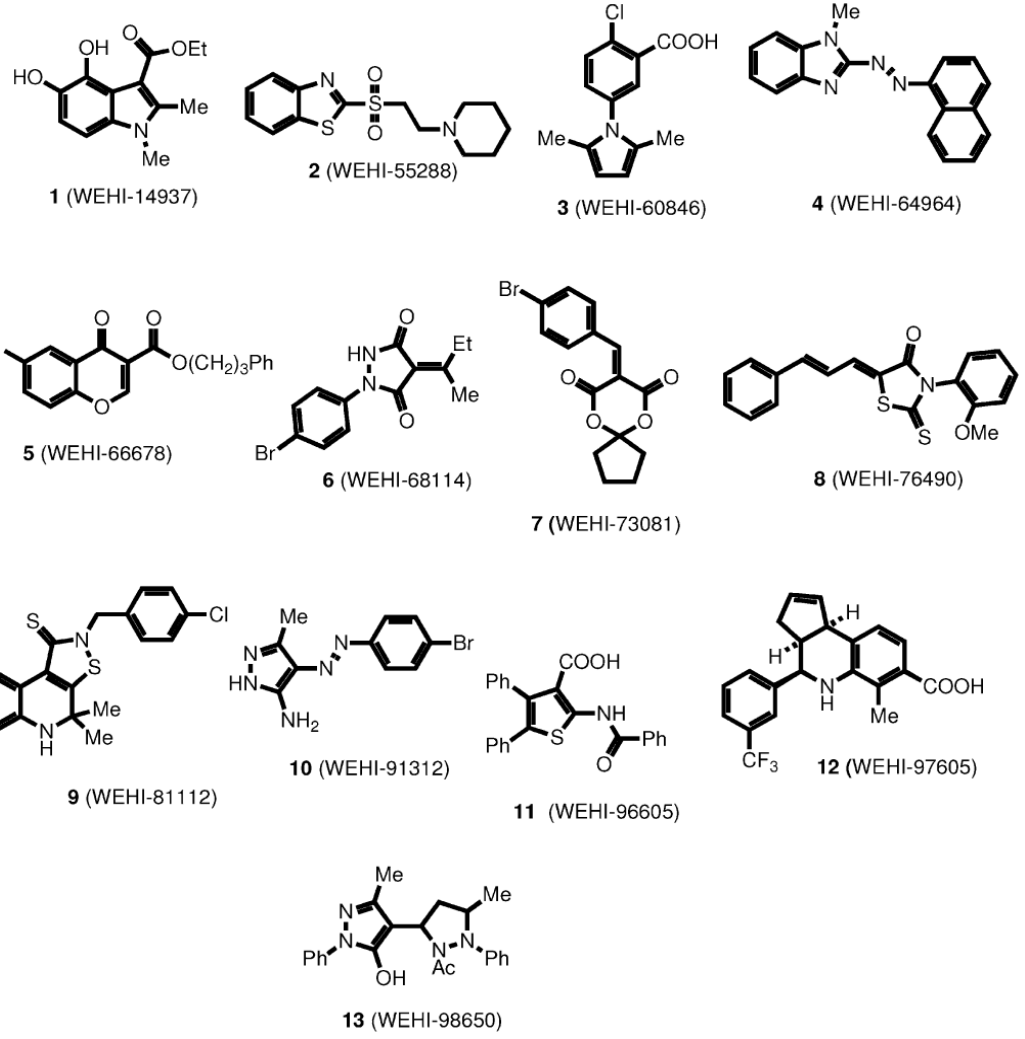
ENONES
Covalent modifier.

QUINONES AND CATECHOLS
Redox cycler, metal complexer, covalent modifier.

Baell & Walters, *Nature,* **2014,** *513,* 481-483

Pan Assay Interference Compounds (PAINS)

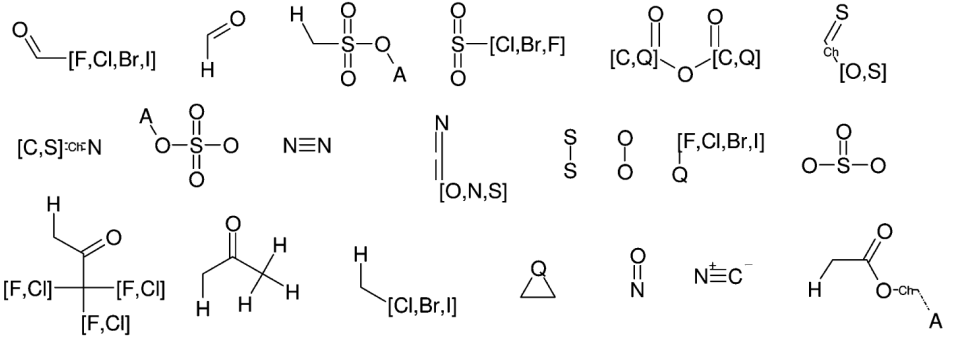
- Красители (в частности, роданиновые);
- Соединения, вступающие в реакции с индикаторами;
- Флуоресцентные соединения;
- Фотоактивируемые соединения



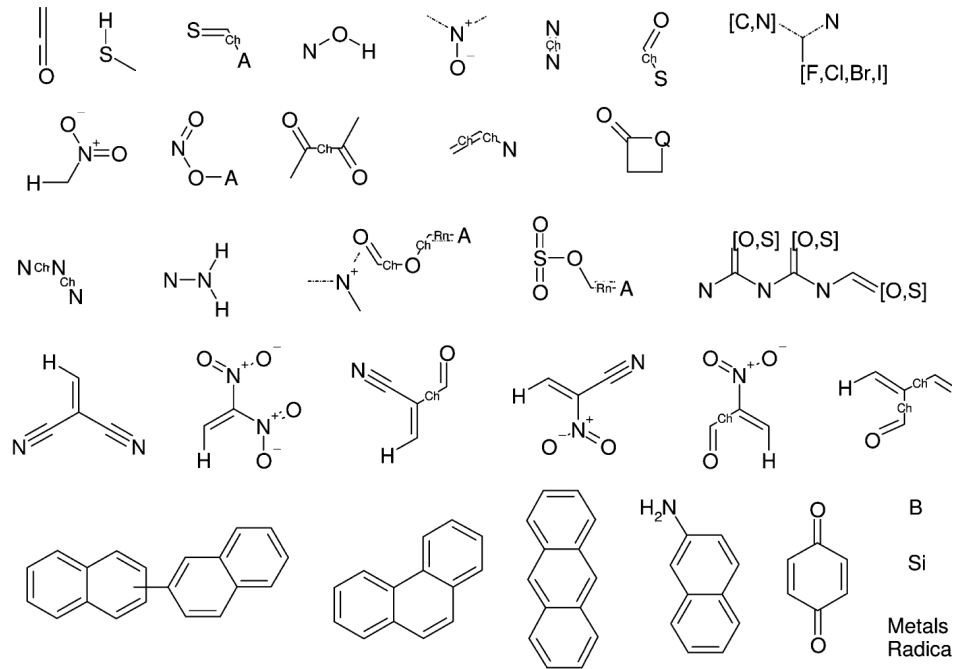
Baell & Holloway, *JMC,* **2010,** *53,* 2719-2740

Примеры фильтров по токсичности

Токсофорные группировки
(*JCIM,* **2010,** *50,* 470-479)

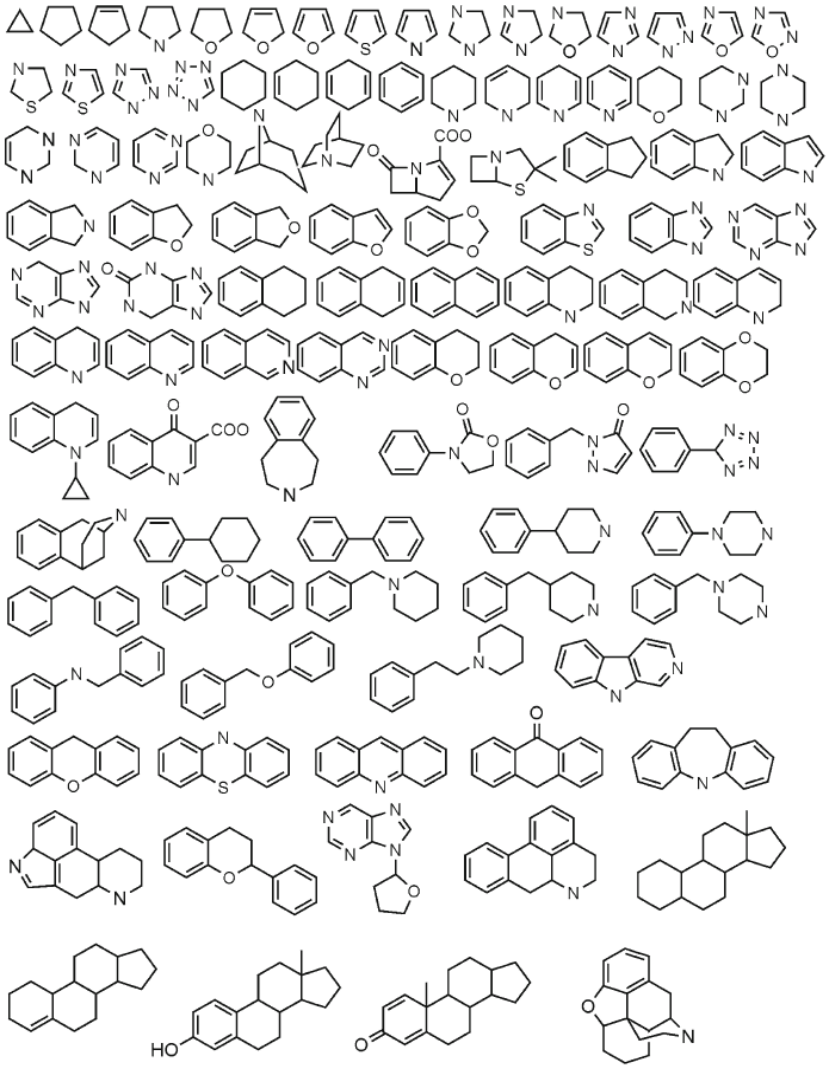


Проекты по предсказанию токсичности — например, OpenTox (http://opentox.org)



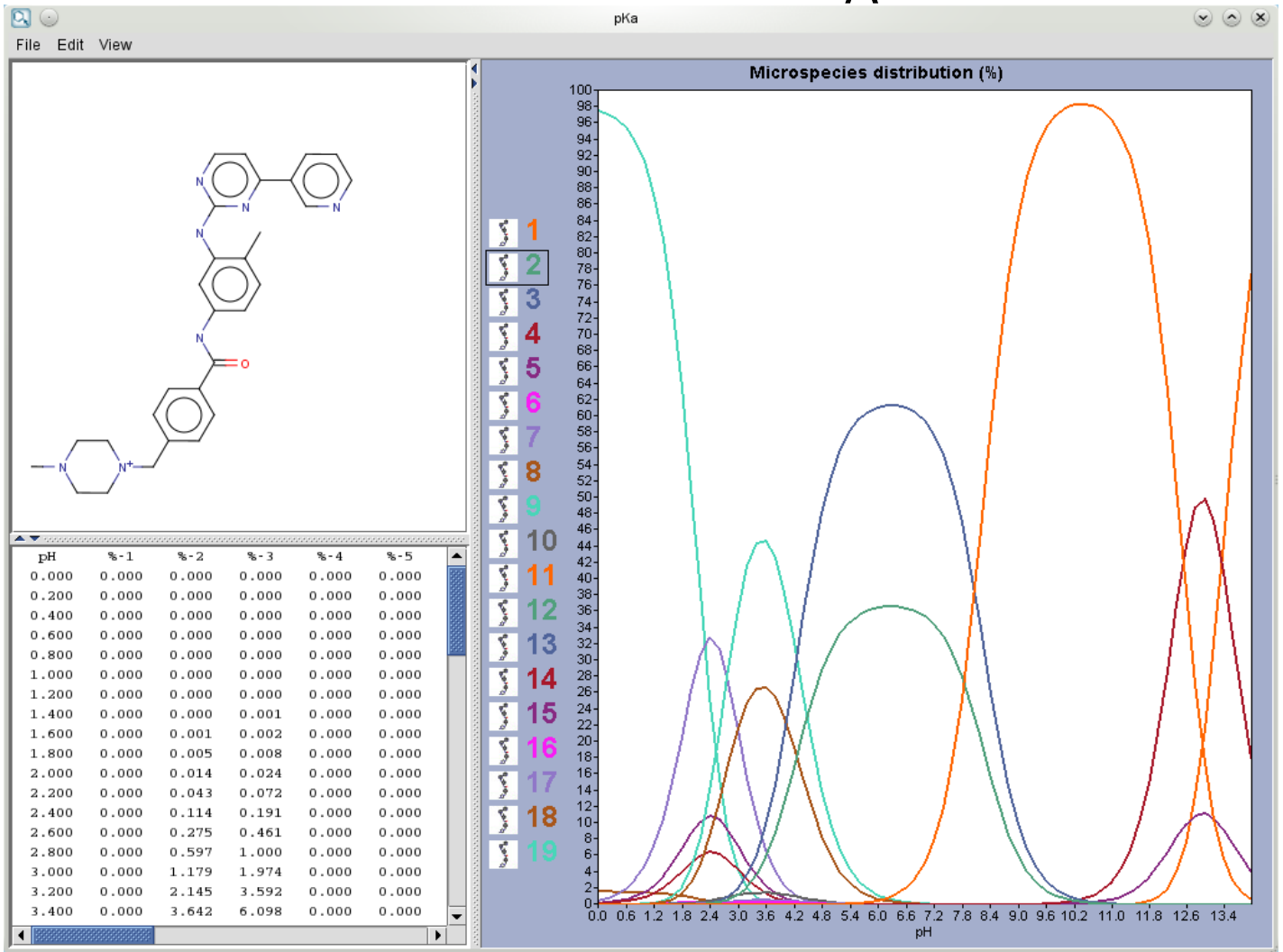
Дизайн библиотеки

Библиотека *ChemBioNet* (http://chembionet.info): построена на основе 561 подструктуры, содержащейся в *World Drug Index*, таким образом, чтобы максимизировать разнообразие и максимально полно покрыть химическое пространство.



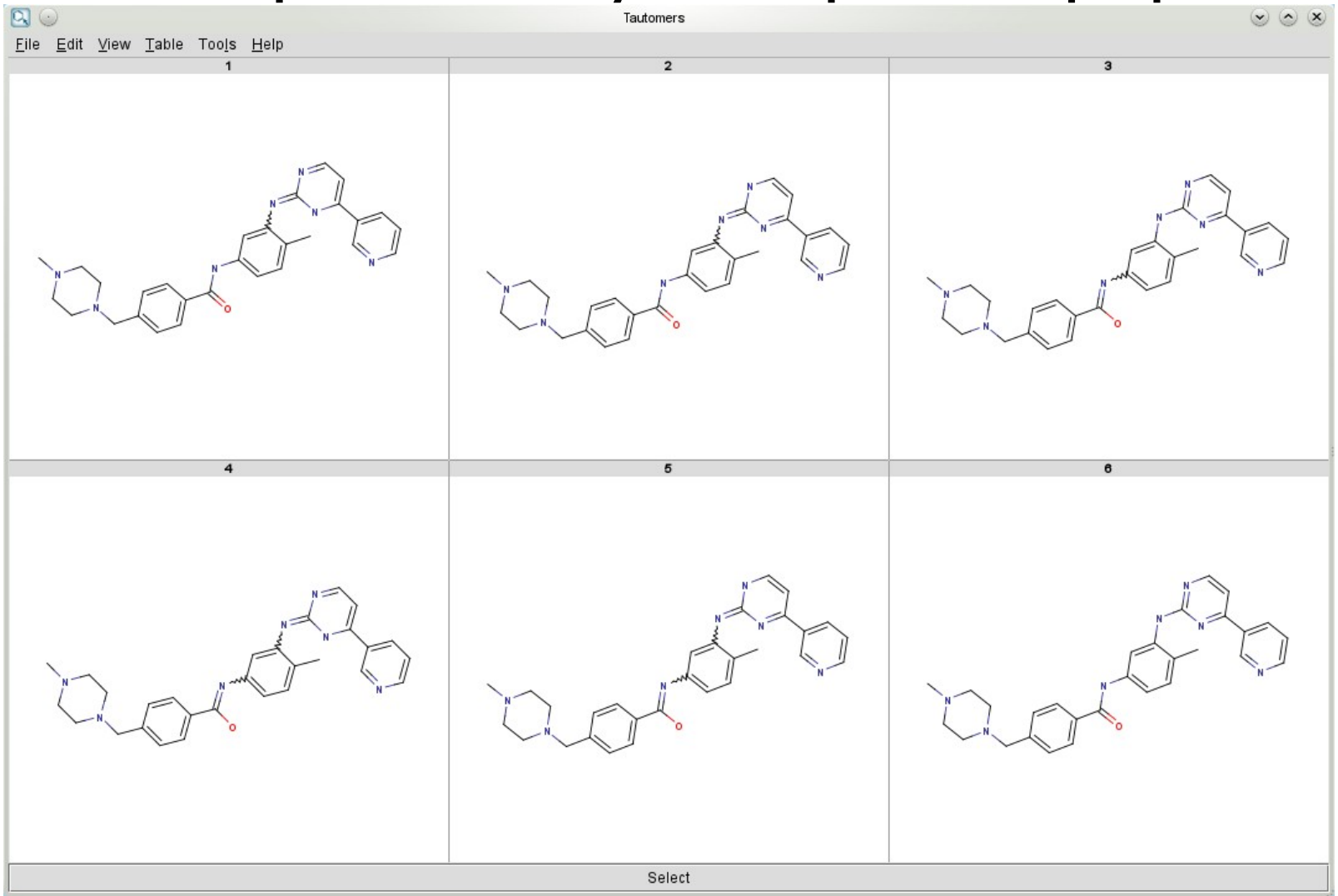
Mol. Divers., **2010, 14,** 401-408

Расчёт pK_A



pH	%-1	%-2	%-3	%-4	%-5
0.000	0.000	0.000	0.000	0.000	0.000
0.200	0.000	0.000	0.000	0.000	0.000
0.400	0.000	0.000	0.000	0.000	0.000
0.600	0.000	0.000	0.000	0.000	0.000
0.800	0.000	0.000	0.000	0.000	0.000
1.000	0.000	0.000	0.000	0.000	0.000
1.200	0.000	0.000	0.000	0.000	0.000
1.400	0.000	0.000	0.001	0.000	0.000
1.600	0.000	0.001	0.002	0.000	0.000
1.800	0.000	0.005	0.008	0.000	0.000
2.000	0.000	0.014	0.024	0.000	0.000
2.200	0.000	0.043	0.072	0.000	0.000
2.400	0.000	0.114	0.191	0.000	0.000
2.600	0.000	0.275	0.461	0.000	0.000
2.800	0.000	0.597	1.000	0.000	0.000
3.000	0.000	1.179	1.974	0.000	0.000
3.200	0.000	2.145	3.592	0.000	0.000
3.400	0.000	3.642	6.098	0.000	0.000

Генерация таутомерных форм



Препроцессинг

Ликвидация шума:

- Липински;
- по токсичности;
- PAINS;
- «lead-like»;
- по разнообразию;

В некоторых случаях приводит к отсеву до 90% молекул

Подготовка структур:

Расчёт pK_A;

Генерация таутомеров

Расчёт частичных атомных зарядов

Виртуальный скрининг агонистов и антагонистов ионотропных глутаматных рецепторов

	Общий размер базы	После препро-цессинга	Докинг в открытую форму рецептора	Докинг в закрытую форму рецептора	3D-QSAR-модели (CoMFA)
Глутаматный сайт AMPA-рецептора	135000	41338	4167	2154	163
Глициновый сайт NMDA-рецептора			2968	1517	89
Глутаматный сайт NMDA-рецептора			1816	1636	—

Russ. Chem. Bull., **2004,** *53,* 1335-1344

Специфическое фильтрование

Методы поиска веществ, действующих на интересующую нас мишень, могут быть основаны:

На структурах известных лигандов (фармакофорный поиск, поиск по подобию, QSAR, другие методы)

Для этого нужна хотя бы одна активная молекула

Встаёт вопрос о важности конкретной конформации

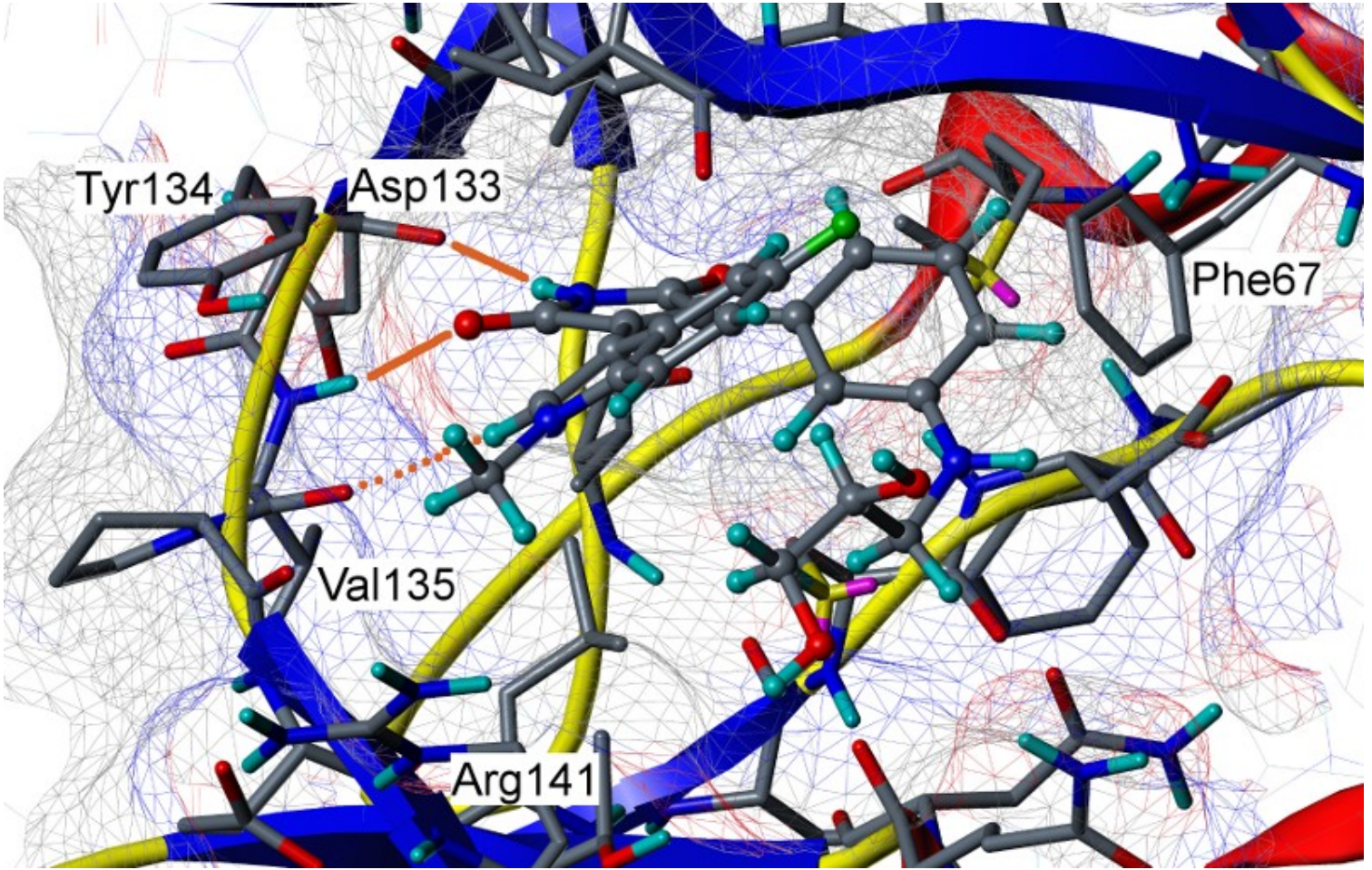
На модели структуры мишени (даже рентгеновская структура — не более чем модель)

Нужна структура белка (не всегда доступна)

Конформации белка тоже могут иметь разный смысл

— Можно ли использовать конформацию модели?
— Ну, в общем, да. Всё равно ничего лучше часто нет.

Виртуальный скрининг по структуре мишени



На самом деле это просто докинг!

Ограничения сильно улучшают качество докинга и виртуального скрининга

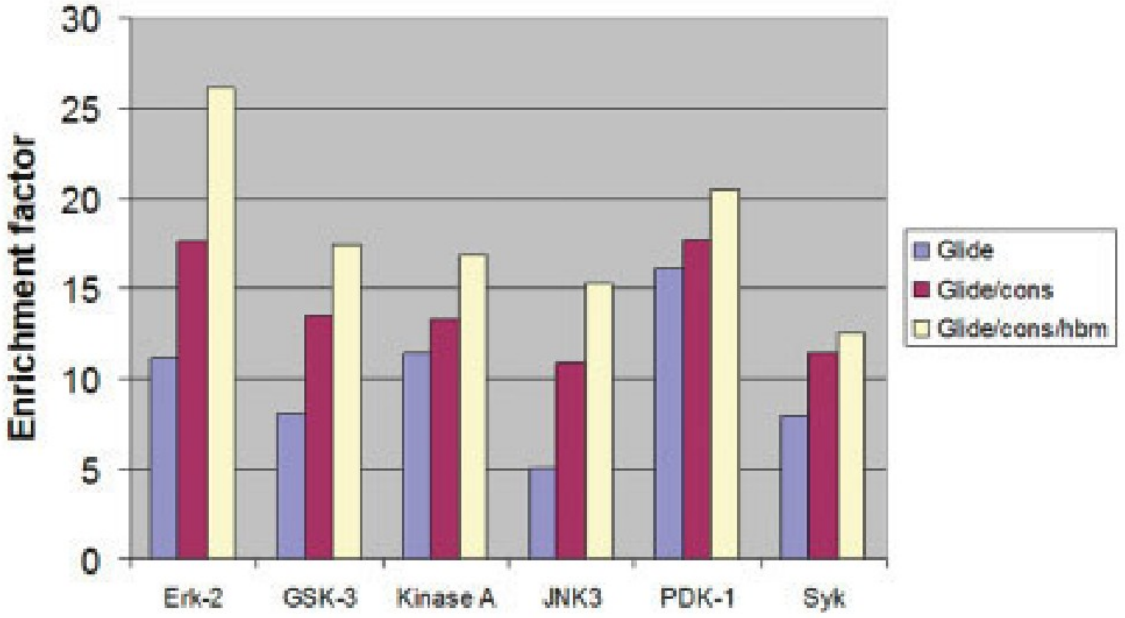


Figure 6. Comparison of enrichment factors obtained with different protocols. 1) unconstrained docking (Glide); 2) constrained docking (Glide/cons); 3) constrained docking followed by hinge-binding motif filter (Glide/cons/hbm). The enrichment factors were calculated on the top 3% of the ranking for protocols 1 and 2. In protocol 3 the hinge-binding motif filter was applied to the top 3% of protocol 2 and the enrichment was calculated on the resulting hit list.

Perola *et al., Proteins,* **2006,** *64,* 422-435

Виртуальный скрининг при отсутствии известных лигандов

Мишень: гликопротеин E вируса денге

Структура мишени: рентген комплекса с детергентом (PDB ID: 1OKE)

Библиотека: подвыборка CMC* (молекулярный вес от 200 до 800, единственный фрагмент) — 5331 структура

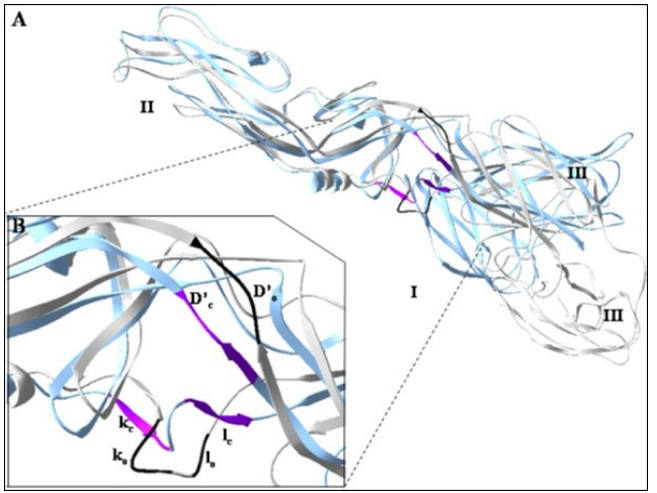
Метод: GEMDOCK

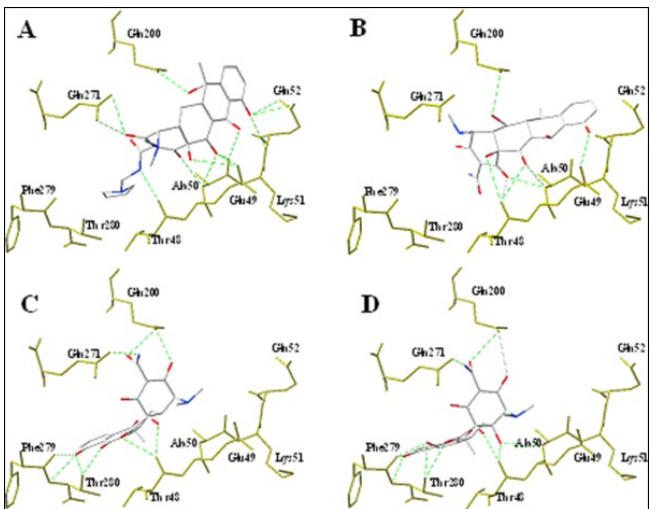
Анализ: значение оценочной функции, визуальный анализ

Отобрано: 10

Активных: **2**

Ссылка: PLoS One, 2007, 2, e428





* Comprehensive Medicinal Chemistry

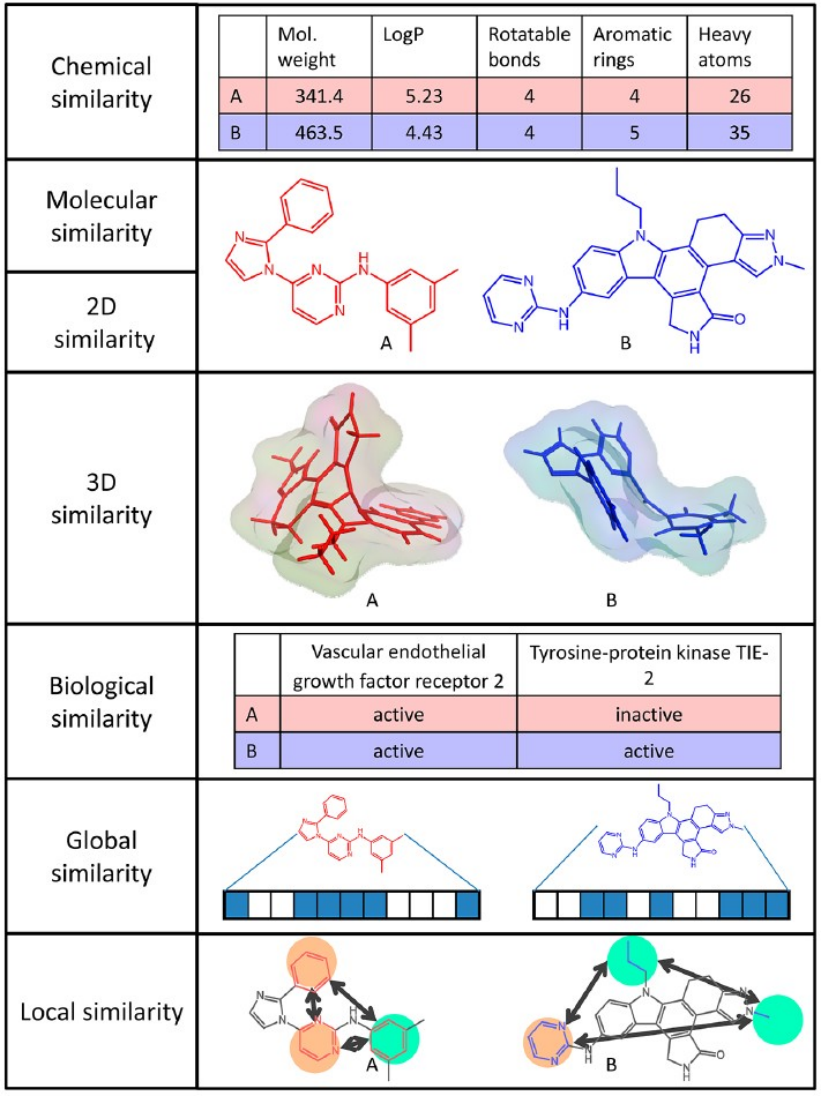
Виртуальный скрининг по структуре лигандов:Молекулярное подобие

Принцип подобия свойств:

Похожие структуры обладают

похожими свойствами

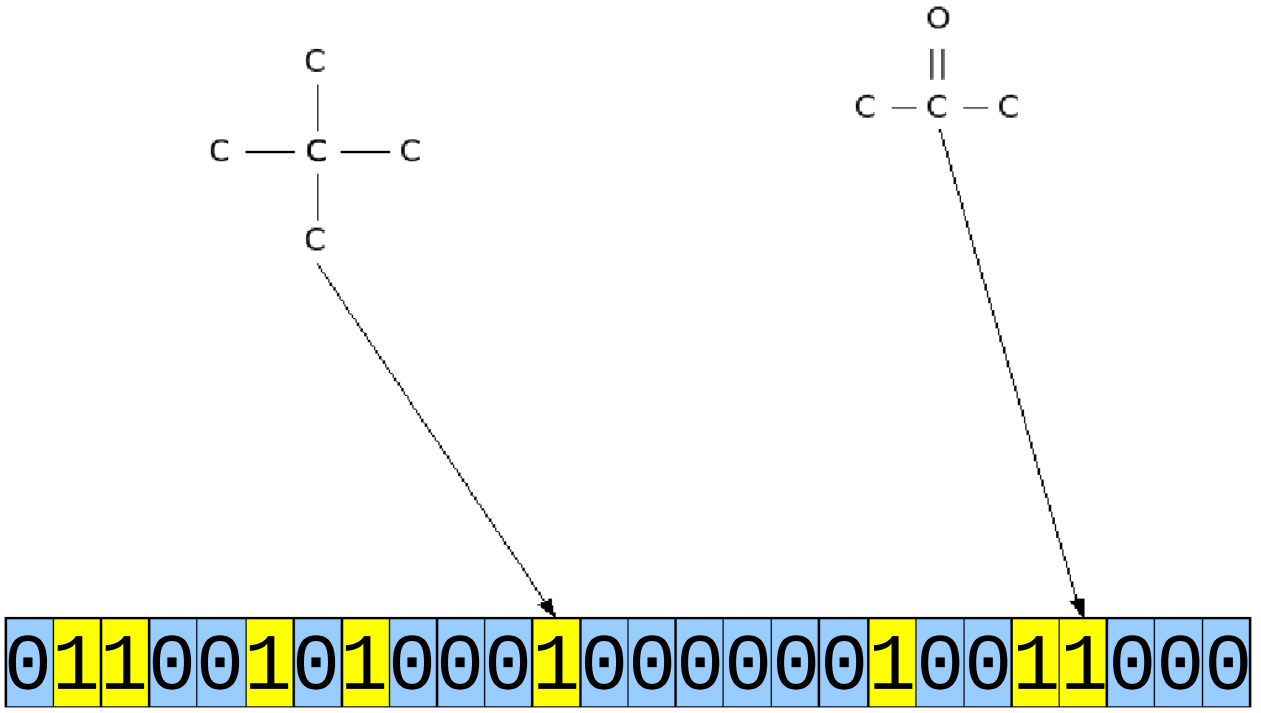
Но как понять,
что структуры
похожи?
И на что они
похожи?



	Mol. weight	LogP	Rotatable bonds	Aromatic rings	Heavy atoms
A	341.4	5.23	4	4	26
B	463.5	4.43	4	5	35

Maggiora *et al., JMC,* **2014,** *57,* 3186-3204

“Фингерпринты” (“Молекулярные отпечатки”)



Значение бита соответствует наличию или отсутствию определённого фрагмента, взаимодействия и т. д.

Индексы молекулярного подобия: Индекс Танимото

$$Tc(A,B)=\frac{A\cap B}{A\cup B} \qquad Tc(A,B)=\frac{c}{a+b-c}$$

a — число признаков в молекуле A
b — число признаков в молекуле B
c — число общих признаков

Tc = 0 — нет в молекулах ничего общего

Tc = 1 — молекулы идентичны

Утверждения верны в пределах изучаемого набора признаков!

Индексы молекулярного подобия: Индекс Тверски

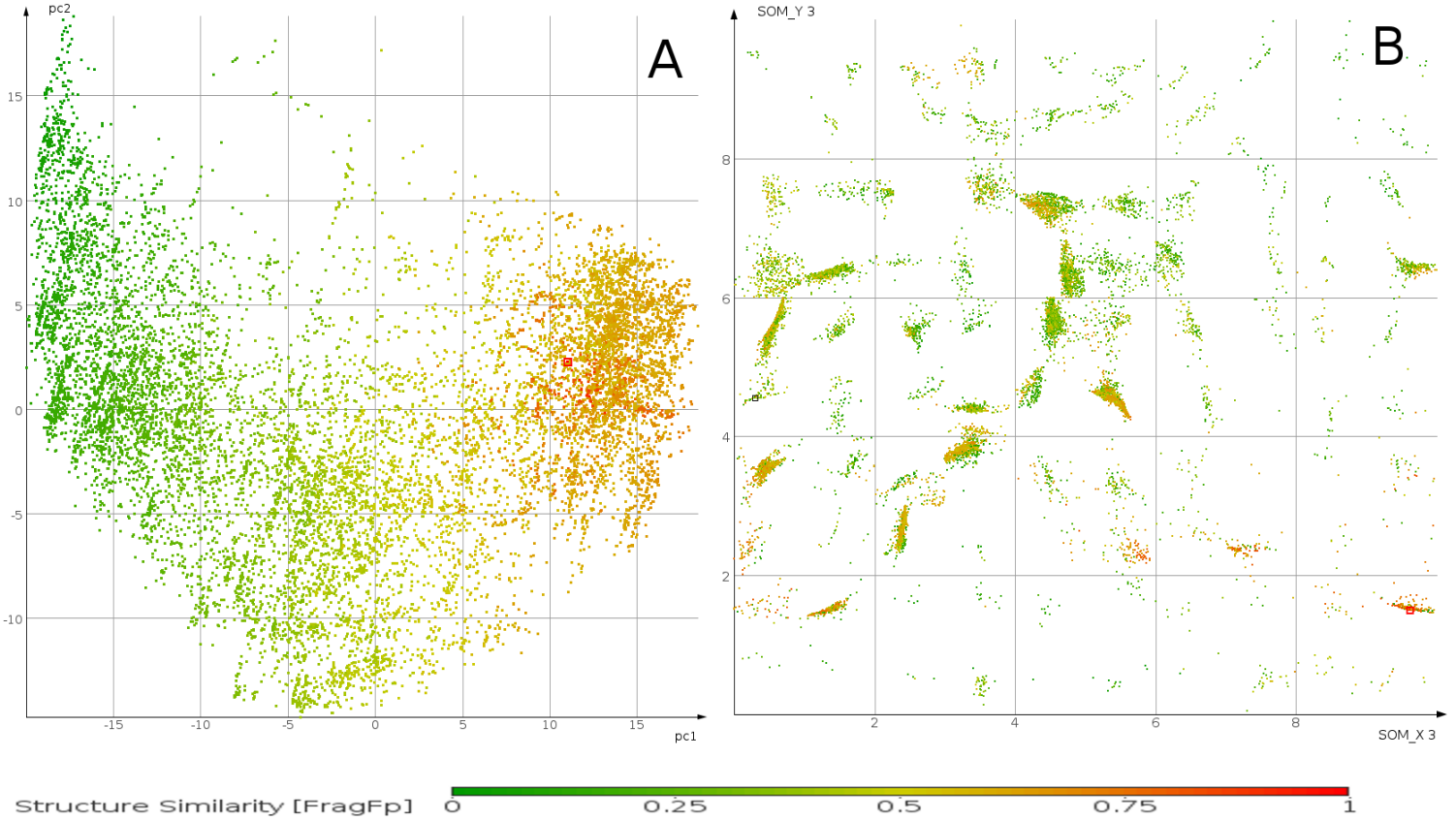
$$Tc(A,B)=\frac{c}{a+b-c}=\frac{c}{(a-c)+(b-c)+c}$$

$$Tv_{\alpha,\beta}(A,B)=\frac{c}{\alpha(a-c)+\beta(b-c)+c}$$

- Асимметричен
- Нормируется на единицу для удобства
- Позволяет оценивать и варьировать относительную важность общих и уникальных признаков

Amos Tversky, *Psychol. Rev.,* **1977,** *84,* 327-352

Визуализация подобия в химическом пространстве



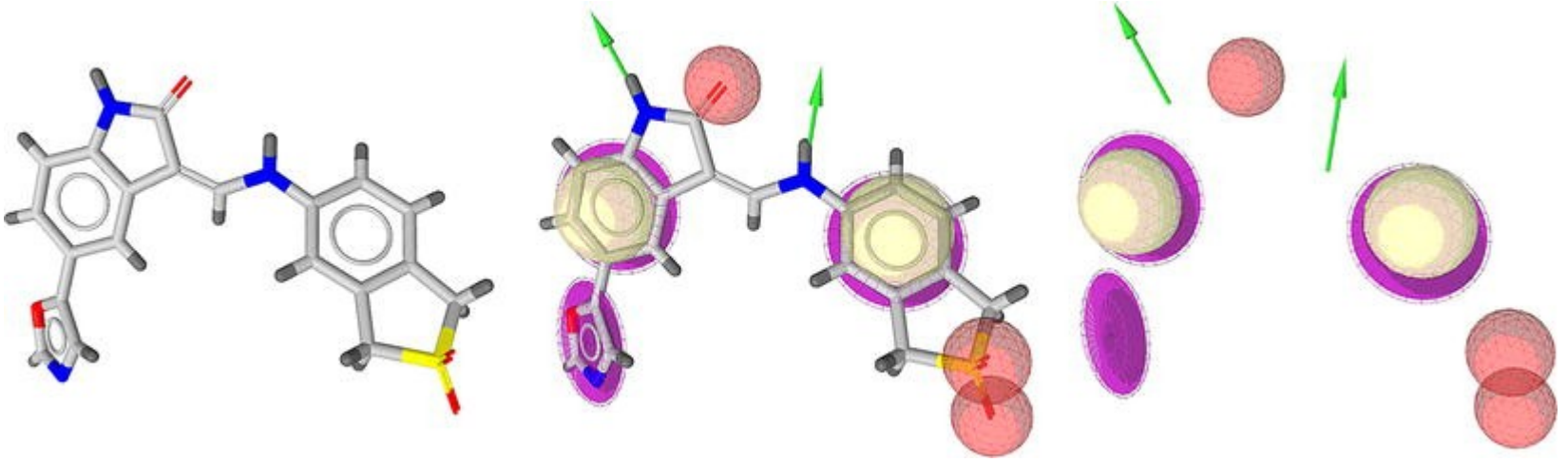
Главные компоненты

Карта Кохонена

Osolodkin *et al., Exp. Opin. Drug Discov.*, **2015,** *10,* 959-973

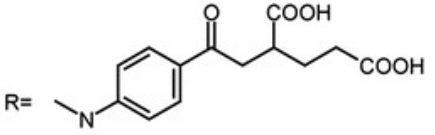
Фармакофор

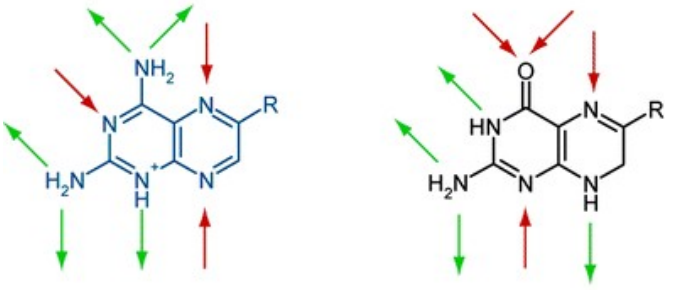
Фармакофор — это набор пространственных и электронных признаков молекулы, необходимых для обеспечения оптимальных супрамолекулярных взаимодействий со специфической биологической мишенью, которые могут вызывать (или блокировать) её биологический ответ.

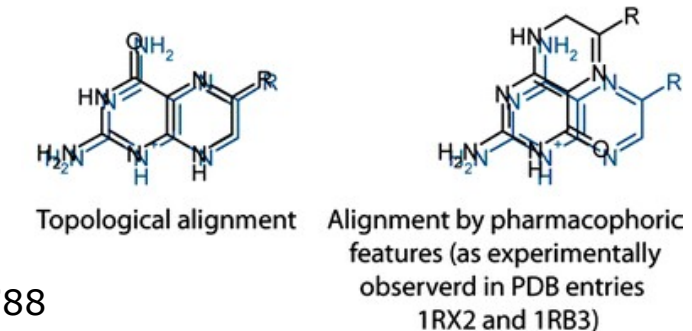


Wolber *et al., JCAMD,* **2006,** *20,* 773-788

Фармакофор и подобие

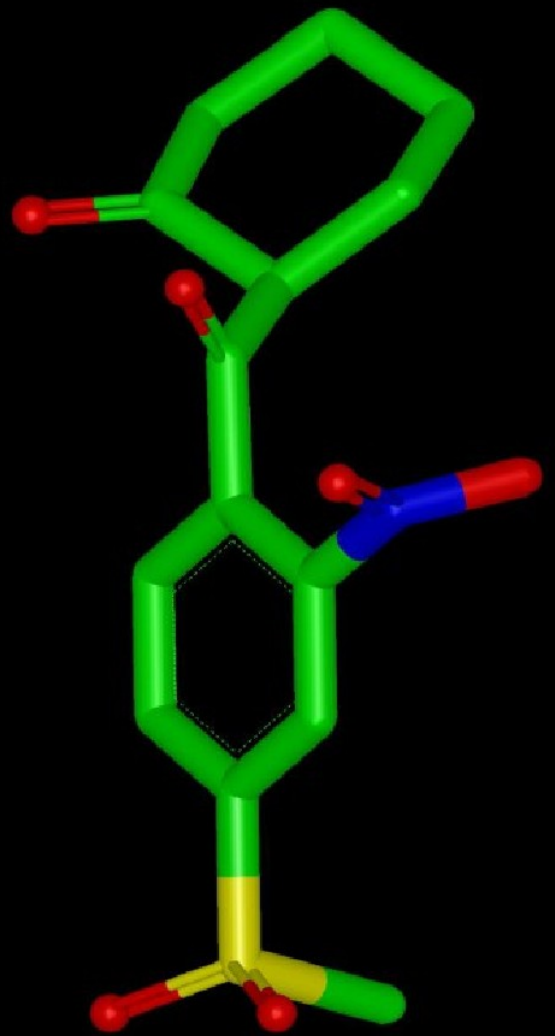






Wolber *et al.*, *JCAMD,* **2006,** *20,* 773-788

Генерация фармакофорных моделей

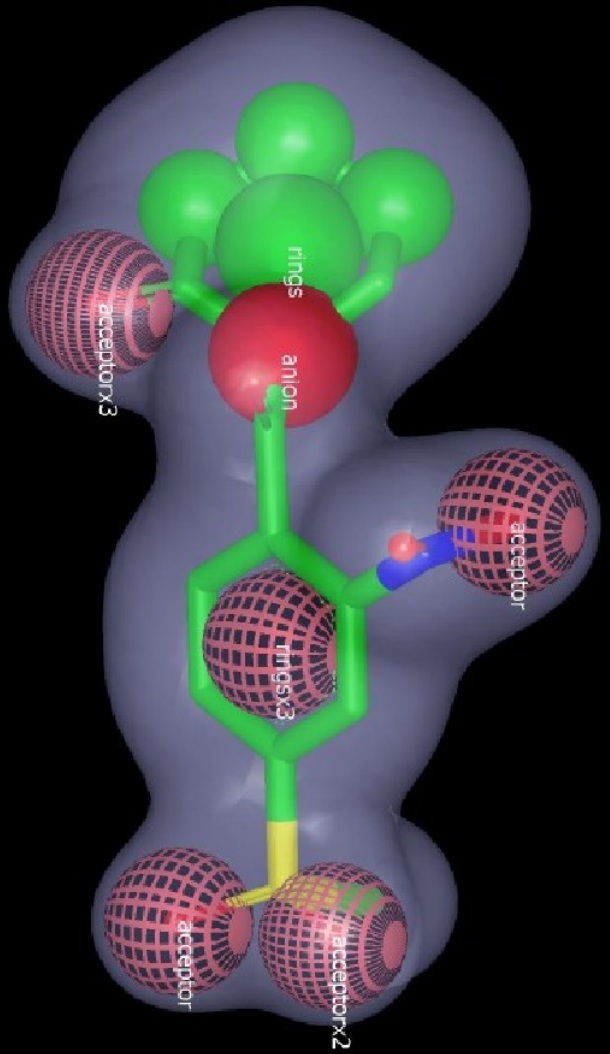


Generate queries

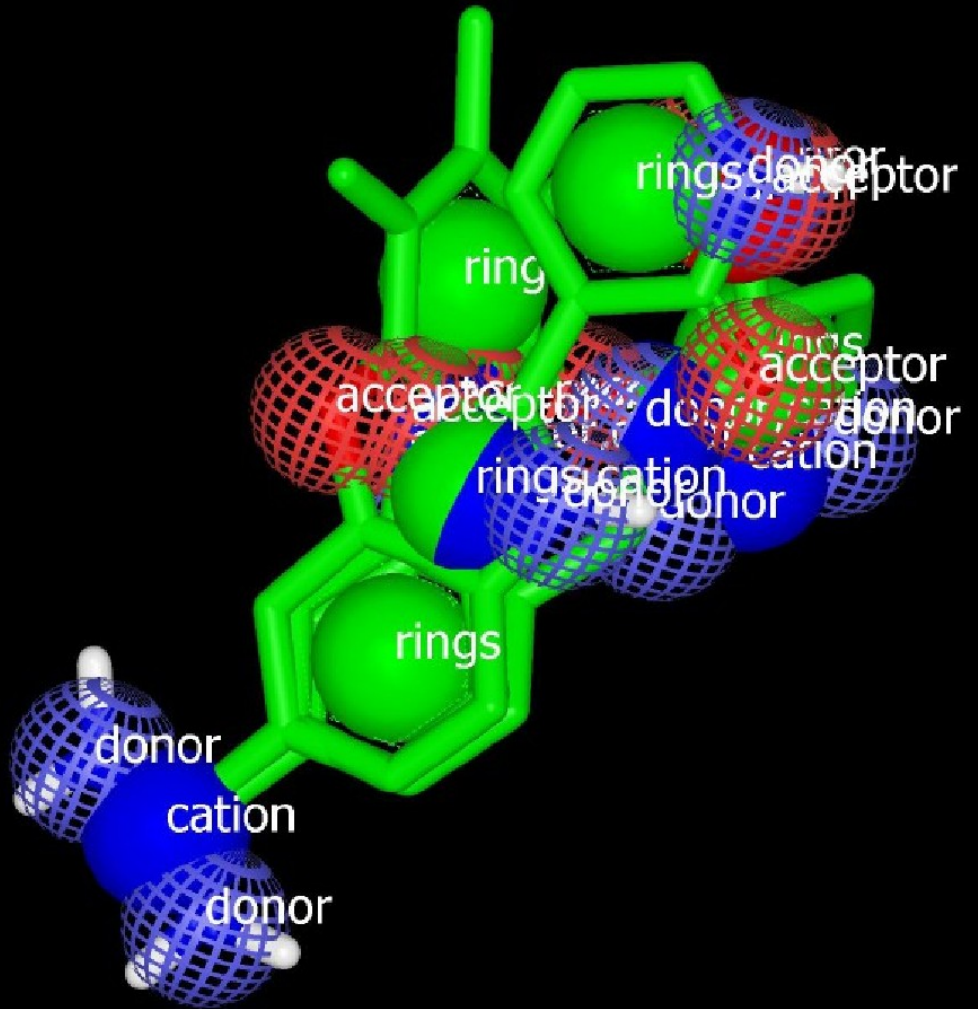
from molecules

Customised queries

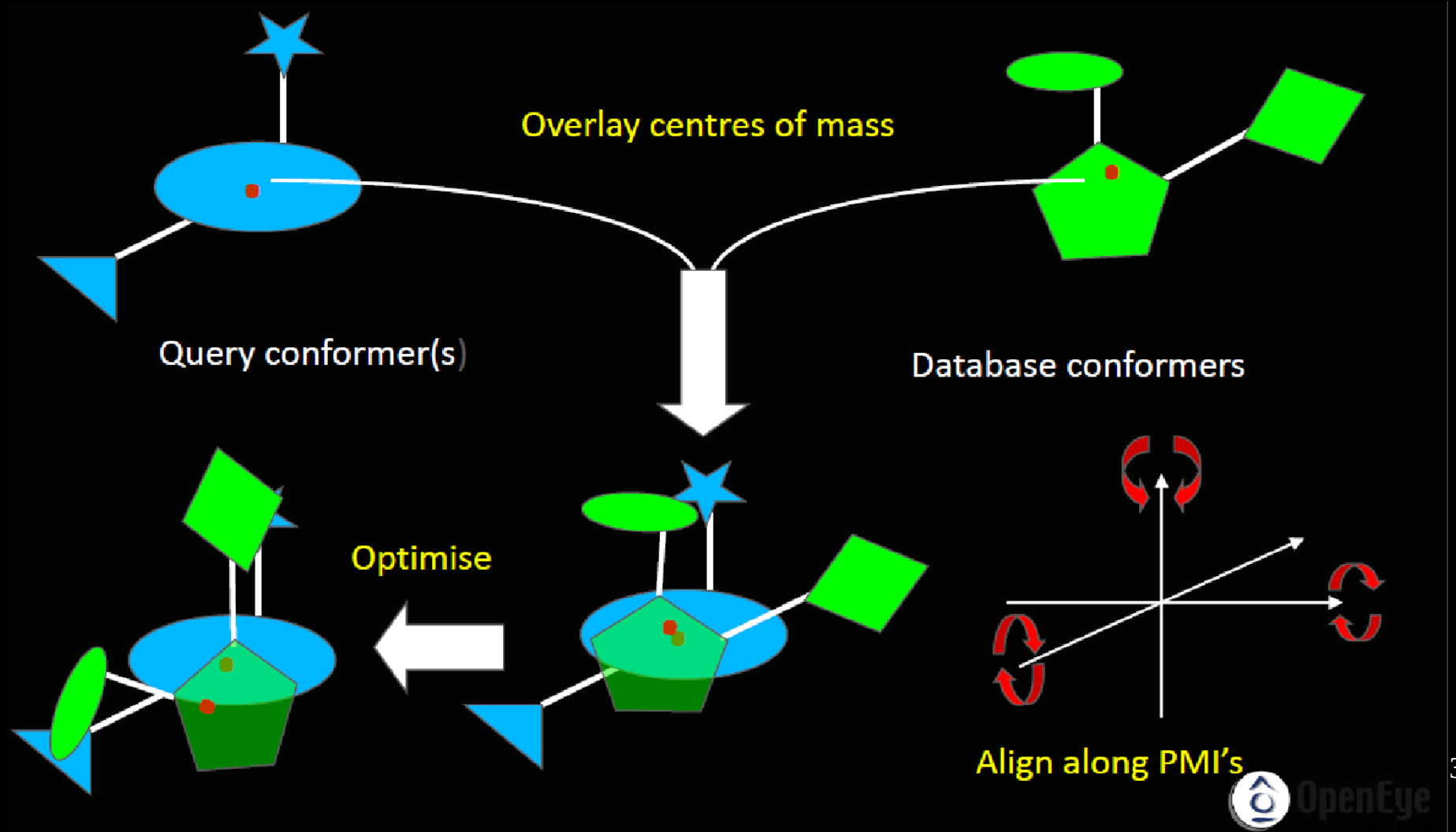
Multi-molecule queries



Фармакофорное наложение



Алгоритм фармакофорного скрининга



Оценка эффективности метода

- **Экспериментальная (“проспективная”)**
  - Оцениваем, сколько выбранных соединений действительно имеет предсказанную активность
  - Необходимо иметь критерий признания соединения активным

- **Статистическая (“ретроспективная”)**
  - Анализируем, насколько хорошо выбираются известные активные соединения
  - Необходимо иметь информацию об активных соединениях

Порог активности

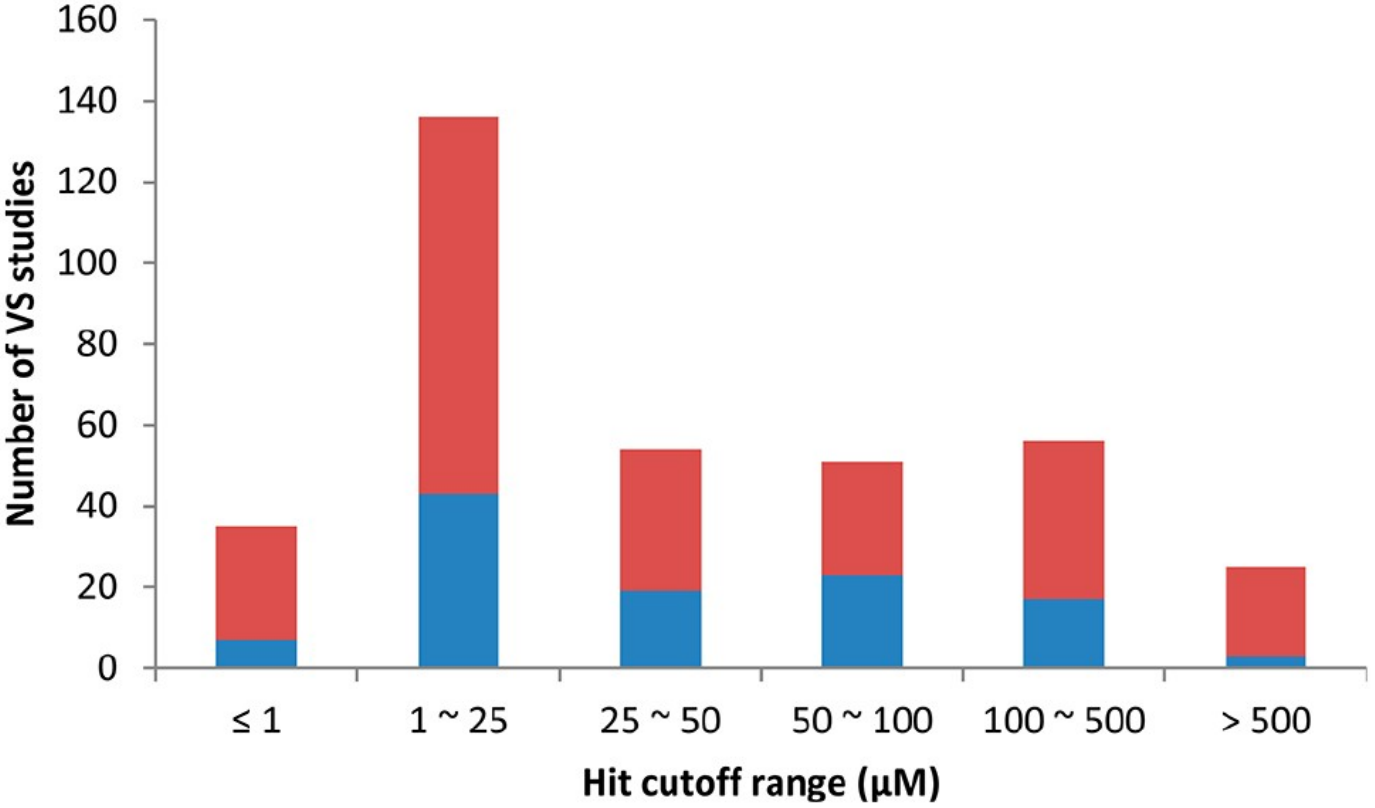


Figure 1. Hit cutoff ranges. Values in blue were obtained from studies with clearly defined hit cutoffs. Values in red are estimated from the lowest experimental activity reported for a hit.

Zhu *et al., JMC,* **2013,** *56,* 6560-6572

Если есть известные лиганды...

...можно оценить и сравнить качество моделей (классификаторов), используемых для виртуального скрининга. Это очень просто.

1. Генерируем тестовые выборки:
— истинно активных соединений (actives)
— ложных лигандов (decoys)

2. Ранжируем объединённую библиотеку истинных и ложных лигандов, и смотрим, наверху ли активные. Сравниваем результаты.

3. Выбираем метрику успешности скрининга.

4. На её основе выбираем самую лучшую модель / метод / программу.

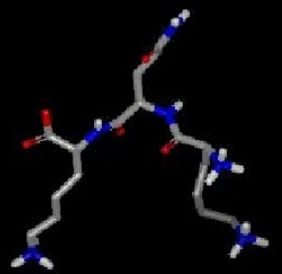
Это называется *«ретроспективное тестирование»*.

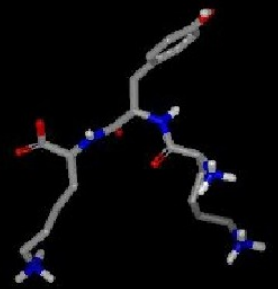
Выборка ложных лигандов

Задача различения между истиной и ложью
в ретроспективном тестировании должна быть **сложной**

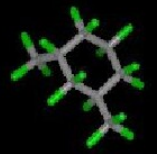
- If your dog can discriminate the actives from the decoys, you need a different set of decoys.

Actives





Decoys



Nicholls, *JCAMD,* **2008,** *22,* 239-255

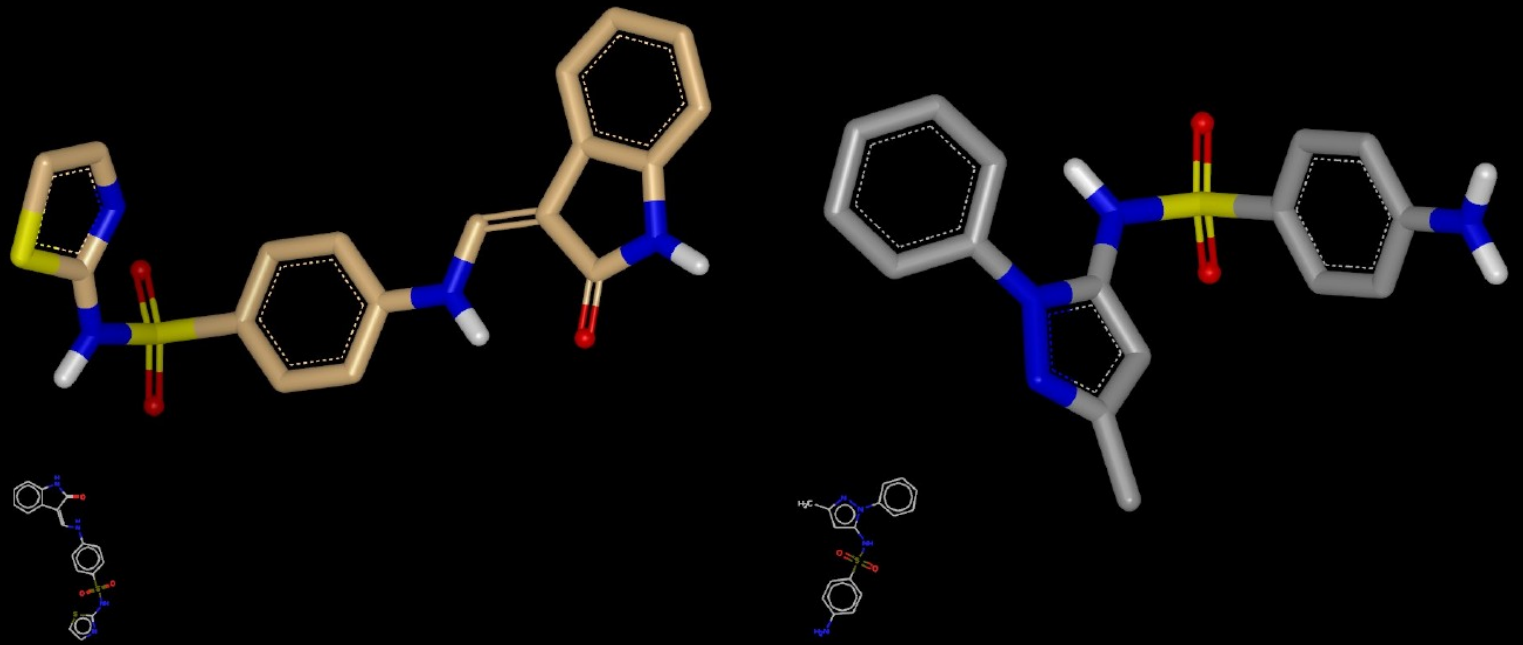
Идеальные результаты...

... подсказывают, что что-то идёт не так

Directory of Useful Decoys (DUD)

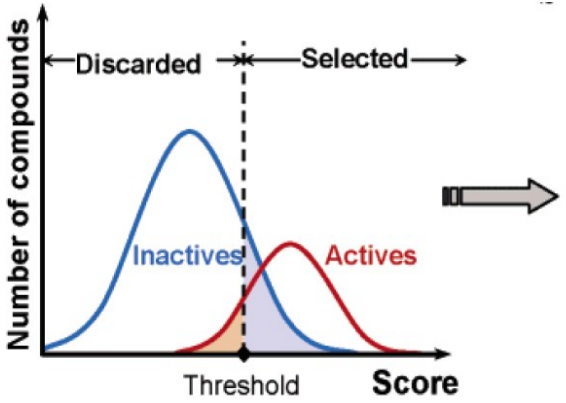
Ложные лиганды должны быть похожи на истинные.

Идея: для многих мишеней выбрать молекулы, похожие на истинные лиганды по физико-химическим свойствам, но отличающиеся топологически.



JMC, 2006, 49, 6789-6801

Ошибки прогнозирования



	in vitro	
in silico	Actives	Inactives
Selected	True Positives (*TP*)	False Positives (*FP*)
Discarded	False Negatives (*FN*)	True Negatives (*TN*)
	Se = *TP*/(*TP*+*FN*)	***Sp*** = *TN*/(*TN*+*FP*)

Что хуже — ложноположительные или ложноотрицательные результаты?

К чему стремиться? Чего бояться и избегать?

И как оценить степень обогащения?

Метрики обогащения

- Фактор обогащения (EF)
- Receiver Operating Characteristic (ROC)
- Boltzmann-Enhanced Discrimination of ROC (BEDROC)
- Robust Initial Enhancement (RIE)
- Сумма логарифмов рангов (SLR)
- pROC
- Cumulative Density Function (CDF)

  (a.k.a. Кривая накопления)

Фактор обогащения

$$EF = \frac{\textit{число извлечённых активных}}{\textit{число ожидаемых активных при равномерном распределении}}$$

$$EF = \frac{\sum_{i=1}^{n} \delta_i}{\chi n} \qquad \text{where } \delta_i = \begin{cases} 1, & r_i \le \chi N \\ 0, & r_i > \chi N \end{cases}$$

Рассчитывается для небольшой доли базы — наиболее высоко оцененных соединений: 0,5%, 1%, 1,5%, редко больше 5%.

Имеет кучу недостатков: нестрогий, зависит от выборки, не даёт возможность отбросить неактивные, штрафует ранжирование одного активного выше другого, не позволяет сравнивать различные эксперименты.

Зато простой и наглядный.

EF — свойство эксперимента, а не метода!

EF зависит от выборки

Какой максимальный фактор обогащения на 1%, если:

— 60 активных, 2940 неактивных (N = 3000)?

— 60 активных, 29 940 неактивных (N = 30 000)?

EF зависит от выборки

Какой максимальный фактор обогащения на 1%, если:

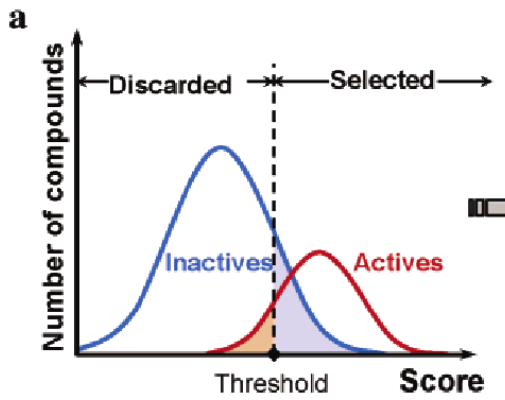
— 60 активных, 2940 неактивных (*N* = 3000)?

— 60 активных, 29 940 неактивных (*N* = 30 000)?

$$EF_1 = \frac{30}{0.01 \times 60} = 50$$

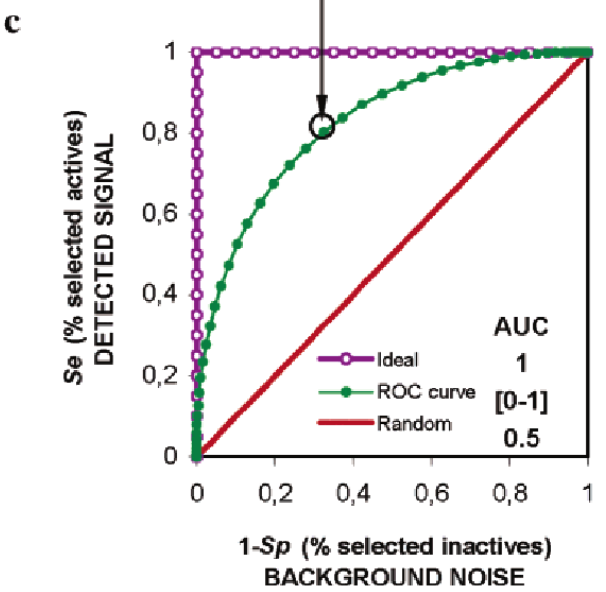
$$EF_2 = \frac{60}{0.01 \times 60} = 100$$

Receiver Operating Characteristic (ROC)



b

	in vitro	
in silico	Actives	Inactives
Selected	True Positives (*TP*)	False Positives (*FP*)
Discarded	False Negatives (*FN*)	True Negatives (*TN*)
	Se = *TP*/(*TP*+*FN*)	***Sp*** = *TN*/(*TN*+*FP*)



Основной числовой параметр — *площадь под кривой* (Area Under Curve, AUC) — имеет простой смысл: вероятность правильного ранжирования двух произвольных соединений.

JMC, **2005,** *48,* 2534-2547

Преимущества ROC

Самое главное преимущество — математическая и статистическая строгость. Следовательно, можно посчитать ошибку и доверительный интервал, что важно для использования в прогностических целях.

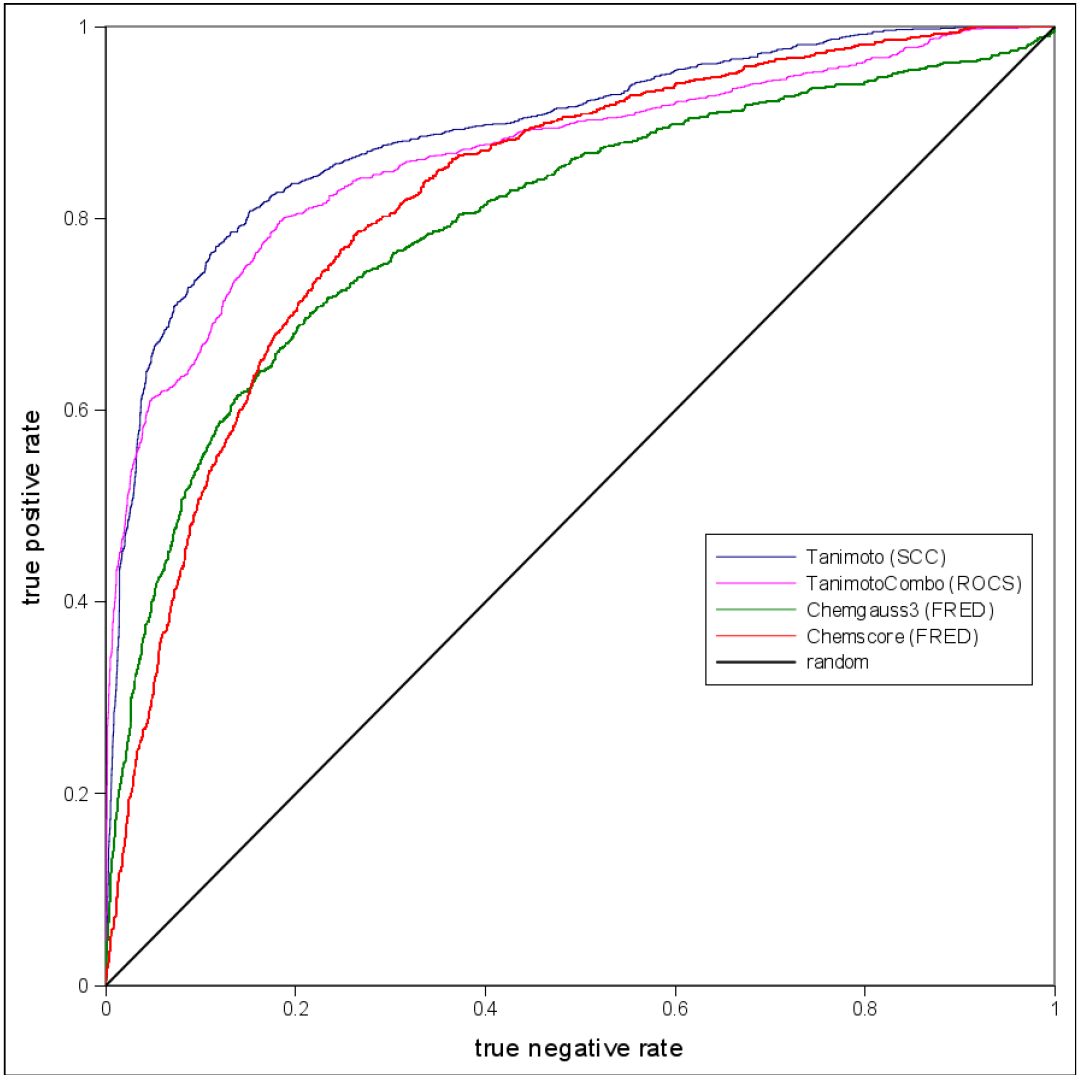
$$SE = \sqrt{\frac{A(1-A)+(n_a-1)(Q_1-A^2)+(n_n-1)(Q_2-A^2)}{n_a n_n}}$$

Кривая ROC (и площадь под ней) не зависит от выборки, что упрощает сравнение результатов анализа.

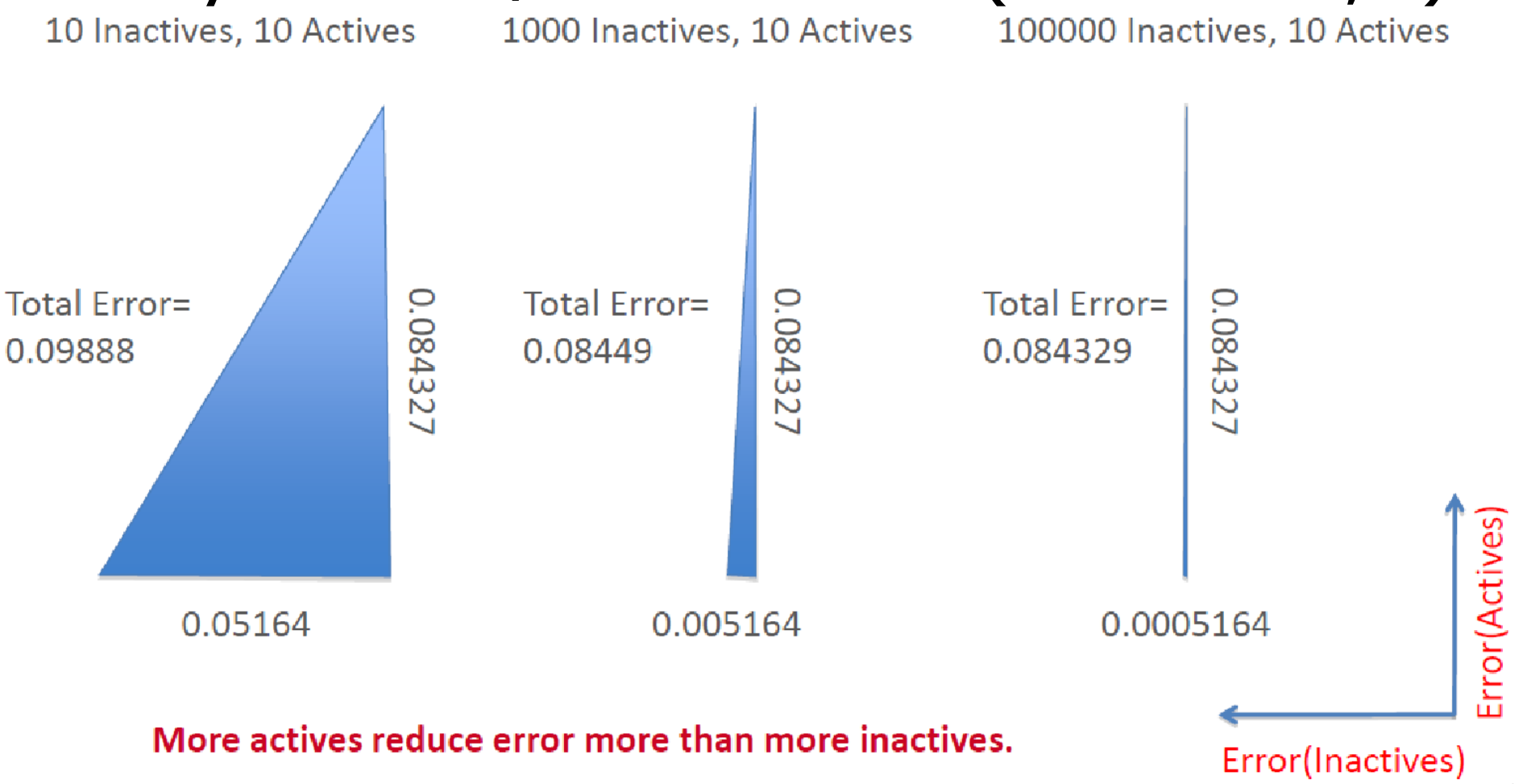
Основной недостаток ROC

Ничего не говорит о раннем обогащении

Раннее обогащение



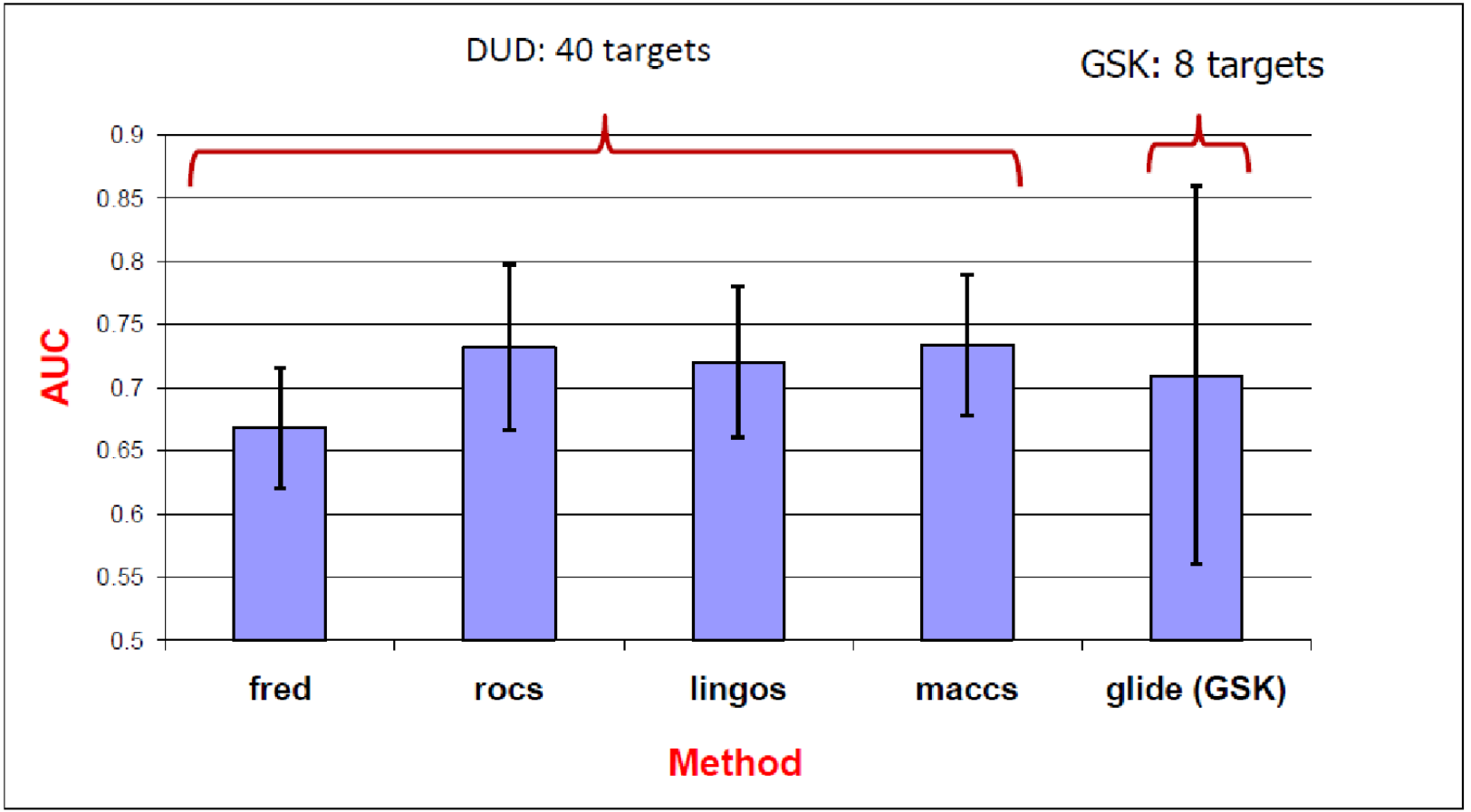
Визуализация ошибок (*AUC* = 0,8)



More actives reduce error more than more inactives.

Number of independent trials decreases error.

Анализ ошибок

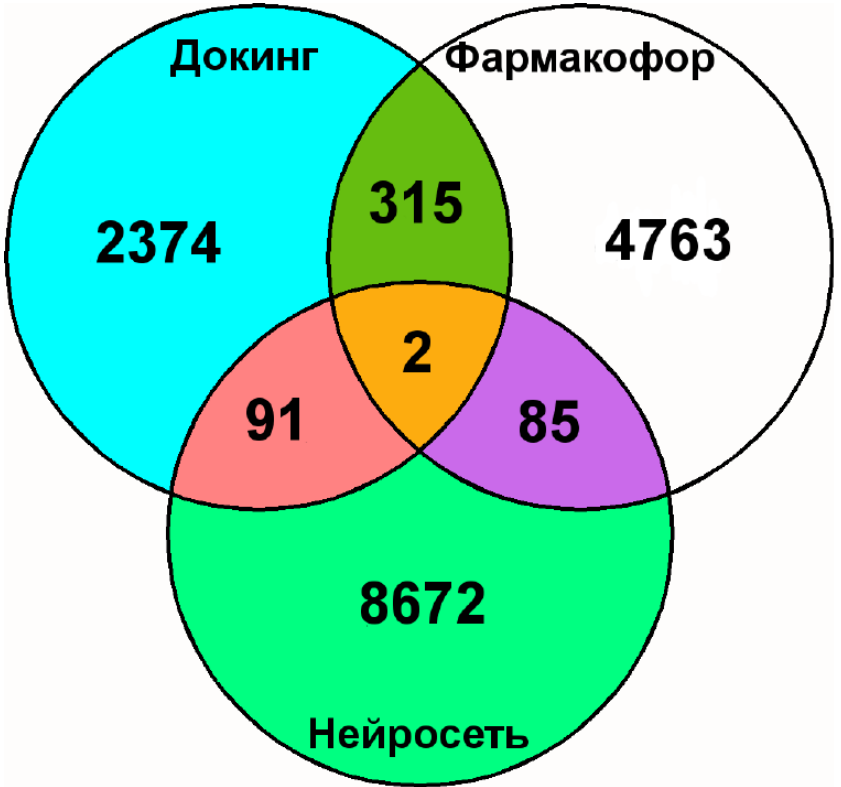


Разные методы статистически неразличимы!

Предупреждение

Два метода с одинаковой AUC не эквивалентны!

- Они могут приводить к извлечению различной информации
- Одни и те же активные могут иметь различный ранг



Ранговая корреляция: Сперман и Кендалл

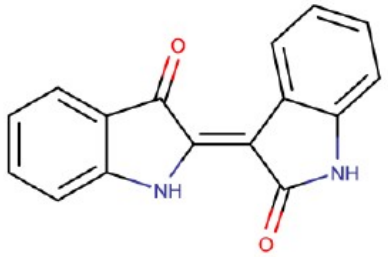
$$r_{Spearman} = 1 - \frac{\sum d_i^2}{n(n^2-1)}$$

$$\tau_{Kendall} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

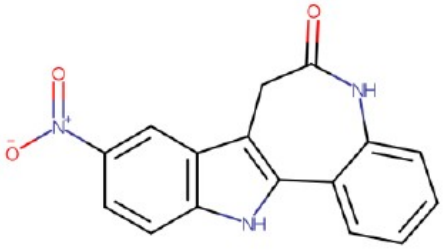
Пример

Target	LINGOS *AUC*	ROCS *AUC*	Spearman	Kendall
ACE	0,862	0,70	0,23	0,16
GR	**0,82**	**0,80**	**-0,18**	**-0,13**
	Одинаковая эффективность		Разные хиты	

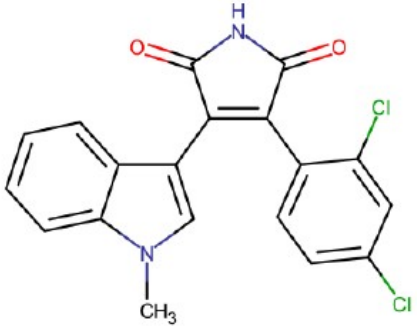
Мультипараметрическая оптимизация



?



?



?

MW

LogP

LogS

HBA

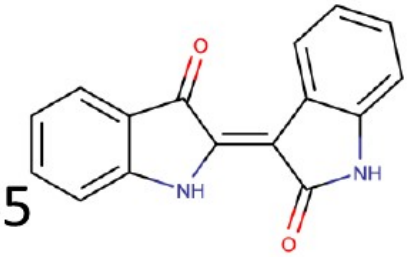
LogBB

Ic50

...

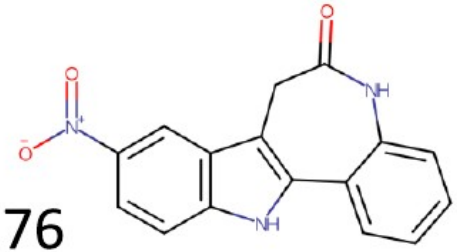
ОЦЕНКА

0.65



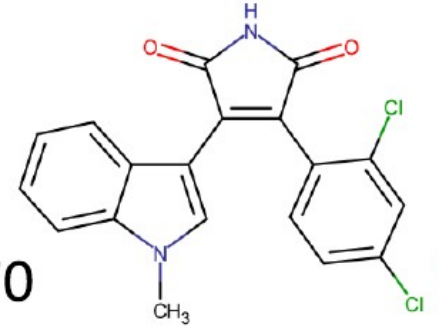
3

0.76



1

0.70



2

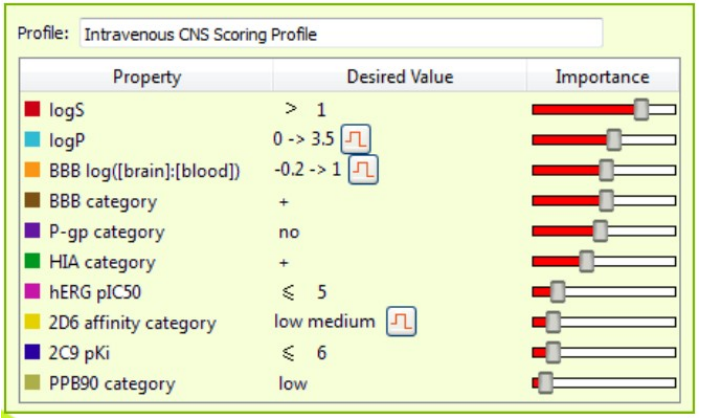
Функции желательности

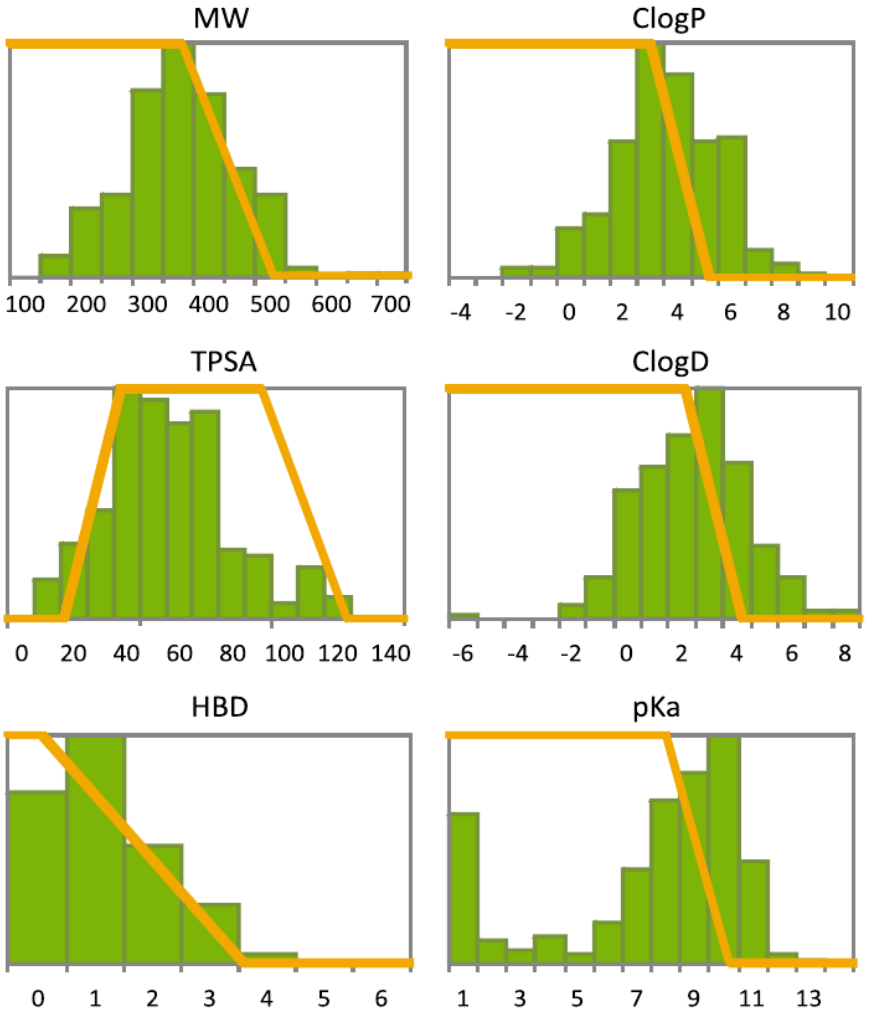
Аддитивный подход
среднее арифметическое
значений функций желательности

$$D = \frac{d_1(Y_1)+d_2(Y_2)+\ldots+d_n(Y_n)}{n}$$

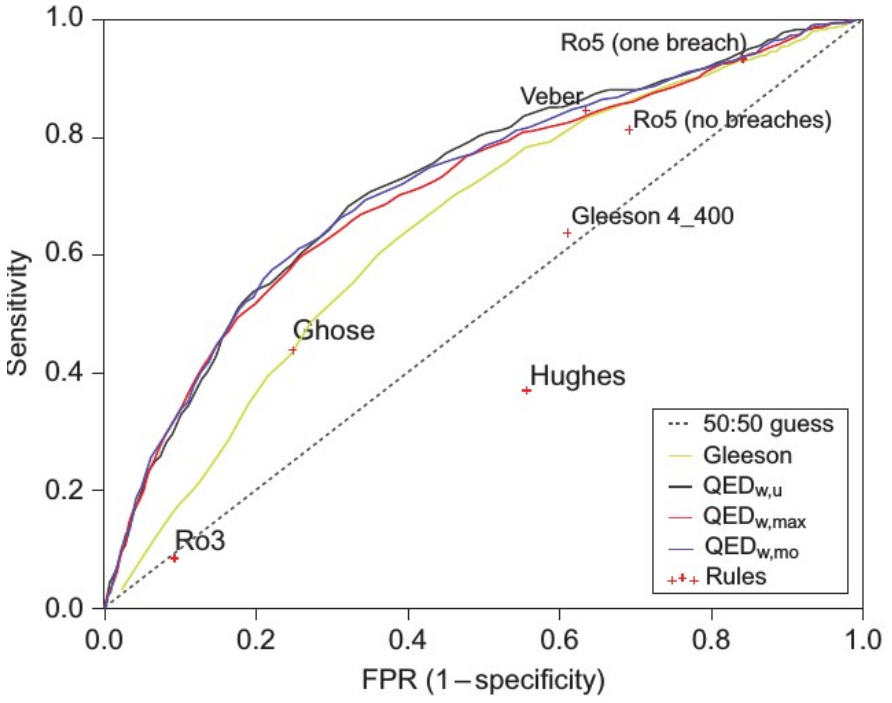
Мультипликативный подход
среднее геометрическое
значений функций желательности

$$D = \sqrt[n]{d_1(Y_1)\cdot d_2(Y_2)\cdot\ldots\cdot d_n(Y_n)}$$





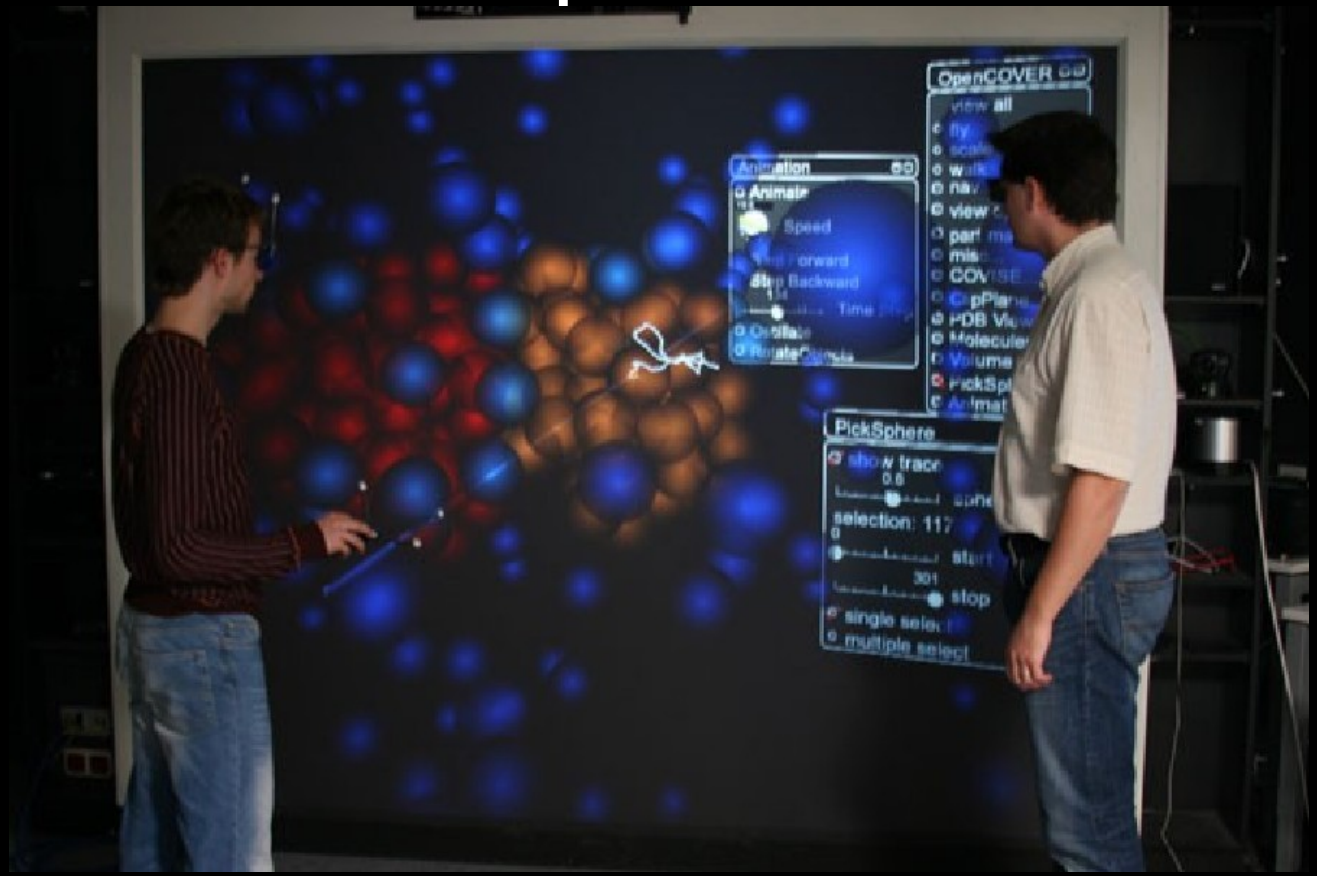
Результат: превосходит иные способы оценки



QED — Quantitative Estimate of Drug-likeness

Bickerton *et al., Nat. Chem.,* **2012,** *4,* 90-98

Постпроцессинг



Визуальный анализ результатов, желательно в компании химика-синтетика, который эти соединения будет синтезировать.

Скачать презентации можно здесь:

http://qsar.chem.msu.ru/ru/obrazov