

Лекция 2. Введение в дизайн лекарств

Курс: Методы машинного обучения в дизайне белков

by Головин А.В.¹ (¹МГУ им М.В. Ломоносова, Факультет Биоинженерии и Биоинформатики)

on Москва, 2023

» Содержание

Активные молекулы

Фарминдустрия

HTS

Хемоинформатика

QSAR

ML в хемоинформатике

Генеративные подходы

Frameworks



» Активные молекулы

- * В основном биологически активные молекулы взаимодействуют нековалентно с биополимерами
- * Агонисты связываются как нативные лиганды и дают тот же эффект
- * Антагонисты конкурируют или препятствуют связыванию нативного лиганда
- * Обратные агонисты связываются и оказывают эффект, обратный эффекту нативного лиганда
- * Хорошие молекулы показывают высокую комплементарность поверхности биополимера



» Свойства лекарства

- * Лекарством обычно являются не только те молекулы, которые хорошо связываются с биополимером.
- * Лекарство должно иметь приемлемую растворимость
- * Часто бывает, что лекарству надо проникнуть сквозь мембрану.
- * Хорошо когда лекарство в итоге метаболизируется, а не накапливается в тканях.



» Как искать активные молекулы?

- * Можно пытаться искать вещества в биоматериалах.
- * Можно проводить роботизированное сканирование библиотеки соединений на активность в разных тестах.
- * Недостаток сканирования: не все тесты можно адаптировать под робота.
- * Возможен высокий уровень шума из-за не специфических взаимодействий
- * Можно применить фильтрацию по подобию соединений, для этого нужны ИТ.



» Особенности деятельности фарм-производителей

Дженерик - лекарство без патентной защиты (срок вышел)

- * Рынок высоко конкурентен.
- * Разработка нового лекарства занимает от 10 до 20 лет.
- * Новые лекарства приносят основную прибыль
- * 4 основные фазы: открытие, разработка, испытания, продажи



» R&D



» Новые технологии

- * Чипы: экспрессия генов.
- * Структуры: роботизированный поиск комплексов с кристаллом белка.
- * Высоко-производительный поиск ингибиторов.
- * Виртуальный поиск.
- * Комбинаторная химия.

Все это в основном относится к стадии поиска ингибитора

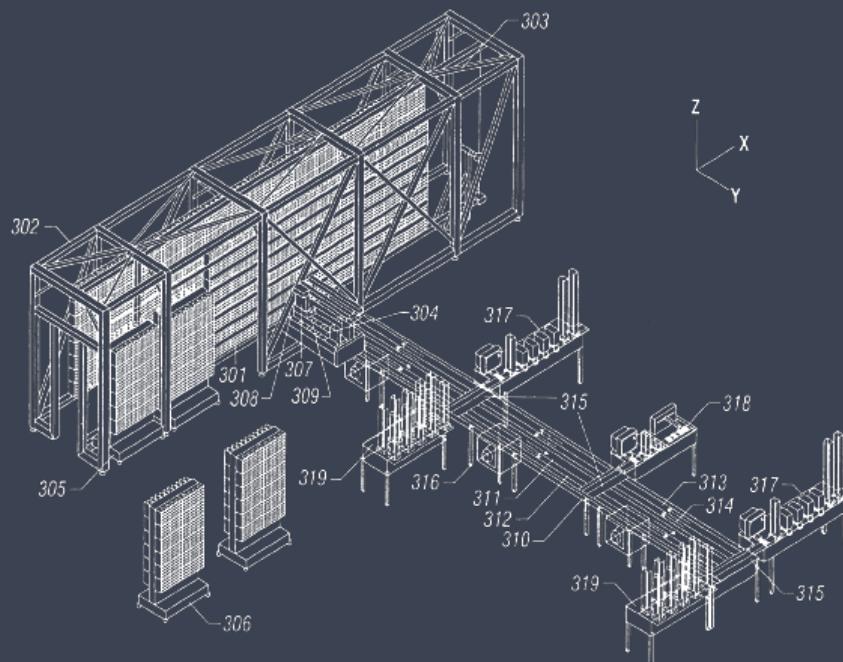


» Как хемоинформатика может помочь?

- * Разработка методов и управление информацией о лигандах.
- * Оценка данных *in silico* для минимизации рисков.
 - * Разработка библиотеки.
 - * Виртуальный поиск.
 - * Оценка стоимости и выгоды.
- * Организация доступа к информации.
- * Интеграция процессов.



» Пример: HTS, Высоко-производительный поиск ингибиторов



до 100000 соединений в день



» HTS и поток данных

- * Исполнить HTS.
- * Решить какие соединения активны а какие нет.
- * Кластеризация активных соединений в классы.
- * Визуализация.
- * Идентификация "основы" для каждого класса.
- * Поиск причин, элементов структуры, которые приводят к "не активности".
- * Использование структурной информации для объяснения активности.



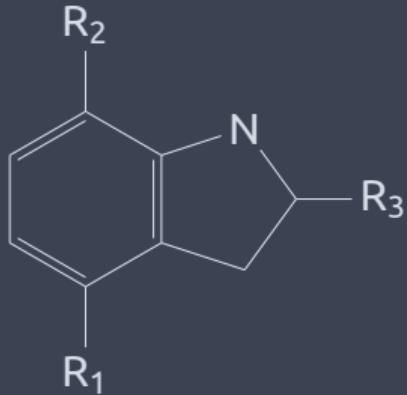
» Пример, комбинаторная химия

- * Исследователи используют "строительные блоки" для быстрого создания большого количества разных соединений.
- * Обычно используется некоторая "основа" и "строительные блоки" присоединяются к разным местам основы.



» Комбинаторная химия

"Основа"



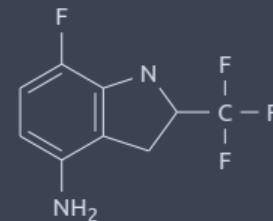
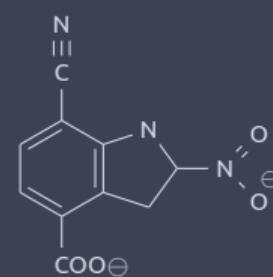
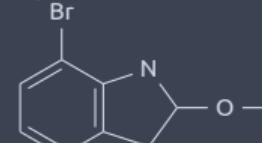
"Блоки"

$R_1 = OH, OCH_3, NH_2, Cl, COOH$

$R_2 = Phe, OH, NH_2, Br, F, CN$

$R_3 = CF_3, NO_2, OCH_3, OH, PheO$

Примеры



» Хемоинформатика и библиотеки

- * Какие блоки выбрать?
- * Какие библиотеки строить?
 - * Дополнение известных наборов
 - * Модификация под конкретный белок
 - * Полное "насыщение"библиотеки
- * Компьютерное профилирование библиотеки
 - * Виртуальными библиотеками удобно манипулировать на компьютере



» Компьютерное представление молекул

- * Хранение в компьютере молекулы как изображения имеет малую ценность
- * Большинство современных баз данных представляет молекулу как граф, с узлами и рёбрами
- * Графы представляются как таблицы связей.

Marvin 04200617372D

```
4 3 0 0 0 0          999 V2000
 0.0000   0.0000   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.7145  -0.4125   0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 -0.7145  -0.4125   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0.0000   0.8250   0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 4 2 0 0 0 0
2 1 1 0 0 0 0
3 1 1 0 0 0 0
M END
```



» Линейное представление молекул, SMILES

Молекула представляется в виде диаграммы и
каждый атом проходится только один раз

<chem>CC</chem>	ethane	<chem>[OH3+]</chem>	hydronium ion
<chem>O=C=O</chem>	carbon dioxide	<chem>[[2H]]O[[2H]]</chem>	deuterium oxide
<chem>C#N</chem>	hydrogen cyanide	<chem>[[235U]]</chem>	uranium-235
<chem>NN(CC)CC</chem>	triethylamine	<chem>F/C=C/F</chem>	E-difluoroethene
<chem>CC(=O)O</chem>	acetic acid	<chem>F/C=C/F</chem>	Z-difluoroethene
<chem>C1CCCCC1</chem>	cyclohexane	<chem>N[[C@@H]](C)C(=O)O</chem>	L-alanine
<chem>c1ccccc1</chem>	benzene	<chem>N[[C@H]](C)C(=O)O</chem>	D-alanine

Реакции в виде SMILES



» Стандартизация SMILES

- * Очевидно, что одну молекулу можно описать разными способами.
- * Морган в 1965 году предложил рассматривать каждый атом по свойству его окружения.
- * Стандартные SMILES называют Unique.

Input SMILES	Unique SMILES
OCC	CCO
[CH3][CH2][OH]	CCO
C-C-O	CCO
C(O)C	CCO
OC(=O)C(Br)(Cl)N	NC(Cl)(Br)C(=O)O
ClC(Br)(N)C(=O)O	NC(Cl)(Br)C(=O)O
O=C(O)C(N)(Br)Cl	NC(Cl)(Br)C(=O)O



» Описание SMILES: атомы

- * Одно буквенные атомы, а именно : B, C, N, O, P, S, F, Cl, Br, I записываются как есть, как один символ.
- * Все остальные атомы записываются в квадратных скобках [Pt]
- * Так как атомы водорода обычно не указываются, то “валентность” атомов определяется как наименьшая из ближайших Т.е. B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6).
- * “Валентности”, отличные от “нормальных”, указывают в скобках [S], [H+], [Fe+2], [OH-], [Fe++], [OH3+], [NH4+]



» Описание SMILES: связи

CC	этан
C=C	этилен
O=C=O	CO2
C#N	HCN
CCO	этанол
[H][H]	водород

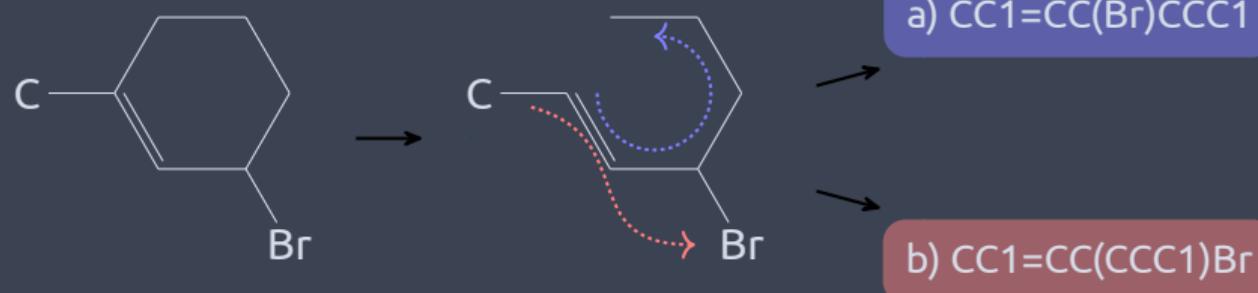
Ветвление цепи отображается в скобках ()

Пример: CCC(CC)COO



» Описание SMILES: циклы

* C1CCCCC1 циклогексан

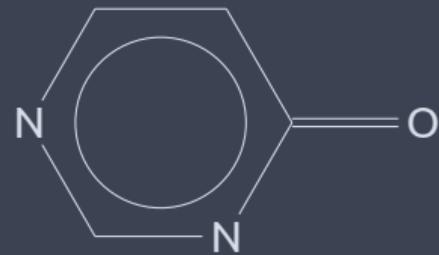


* Или более сложный пример:

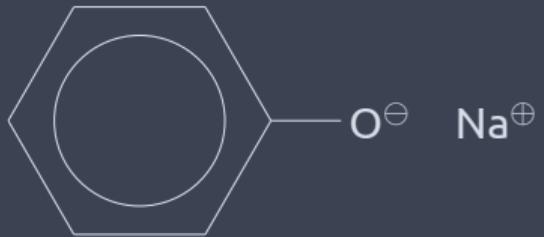


» Описание SMILES: ароматика

- * SMILES для определения ароматичности использует расширенный алгоритм Хюкеля.
- * c1ccccc1 eq C1=CC=CC=C1 тут все атомы находятся в sp₂-гибридизации
- * c1cccc1 eq C1=CC=CC1, последний атом в гибридизации sp₃.
- * Ароматичными могут быть атомы: C, N, O, P, S, As, Se, и *.
- * Пример: c1cnc[nH]c(=O)1



» Структуры где есть нековалентные связи



В SMILES нотации это:

$[\text{Na}^+].[O^-]c1ccccc1$

или

$c1cc([O^-].[Na^+])ccc1$



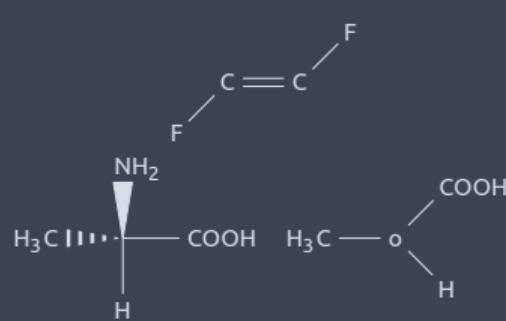
» Изомеры

Изотопы

—

[12C],[13C]

Цис-Транс



F/C=F/C или F\C=F\C

N[C@](C)(C(=O)O)H

N[C@@](C)(H)C(=O)O

Хиральность



» SMARTS: паттерны для SMILES

В принципе, SMARTS это SMILES + операторы логики и варианты в позициях.

Пример для атомов:

C	алифатический углерод
c	ароматический углерод
a	любой ароматический атом
[#6]	любой атом углерода
[++]	атом с зарядом +2
[R]	атом в кольце
[D3]	атом с тремя связями (не с атомами водорода)
[X3]	атом с тремя связями, включая атомы водорода
[v3]	атом с валентностью 3.



» SMARTS: логические операторы и примеры

Логика:

$!e1$	not $e1$
$e1 \& e2$	$a1 \text{ and } e2$
$e1, e2$	$e1 \text{ or } e2$
$e1; e2$	$a1 \text{ and } e2$

Пример:

$[\text{!C}; \text{R}]$	не алифатический С в кольце
$[\text{n}; \text{H1}], [\text{n} \& \text{H1}], [\text{nH1}]$	H в пирроле
$[\text{c}, \text{n} \& \text{H1}]$	C или H в пирроле
$[\text{X3} \& \text{H0}]$	Атом с тремя связями не с H
$[\text{c}, \text{n}; \text{H1}]$	N или C в связи с одним H1



» Линейное представление молекул, InChI

InChI = IUPAC International Chemical Identifier

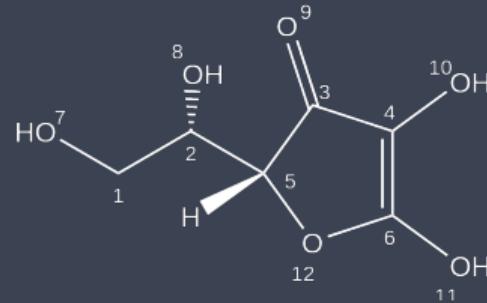
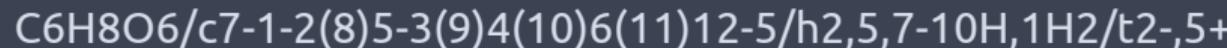
Структура молекул описывается слоями:

- * Основной слой содержит описание брутто формулы, связанности (c) и связей с водородами (h)



- * Слой с описанием заряда (p) кратности связей

- * Слой с описанием стереохимии и связей



» Дискрипторы, правило Лепински

- * Водородные связи
- * Гибкость молекулы
- * Гидрофобность

Правило пяти Лепински

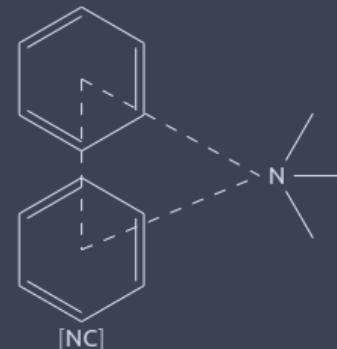
- * No more than 5 hydrogen bond donors
- * No more than 10 hydrogen bond acceptors
- * A molecular mass less than 500 daltons
- * An octanol-water partition coefficient $\log P$ not greater than 5

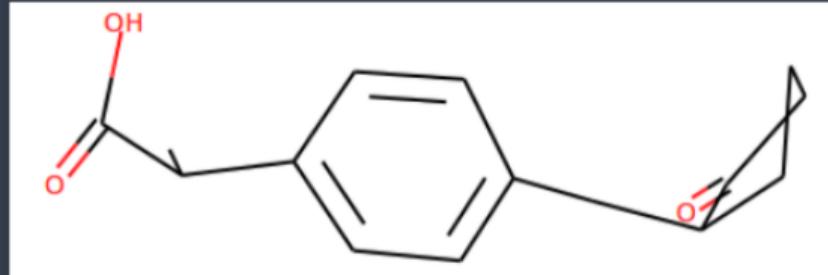


» Поиск по 3D-базам данным

- * Поиск в 2D-пространстве хорош для поиска подобных молекул, но биологически активные молекулы действуют благодаря специфической 3D-структуре.
- * Взаимодействие с биополимером может происходить благодаря нужному расположению в пространстве некоторых групп. При этом различие в 2D-структуре может быть весьма существенным.
- * Фармакофор — это набор свойств, которые являются общими для некоторой группы активных молекул.
- * Пример: Антигистаминный 3D-фармакофор

Антигистаминный 3D-фармакофор





```
fp.ToBitString()
```

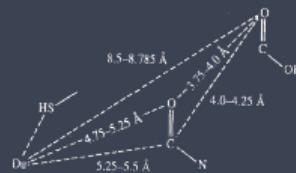
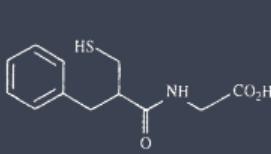
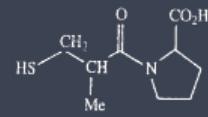
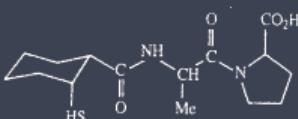
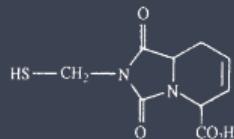
» Проблемы с фармакофорами

- * Если молекулы более или менее подвижны, то это накладывает дополнительные требования на учёт конформационных превращений.
- * Для определения фармакофора надо определить, какой набор групп располагается в биополимере идентично.
- * Надо быть уверенным, что выбранный набор молекул связывается с белком в одном и том же месте. Однозначное указание на это можно получить только экспериментально.



» Систематический поиск

- * Есть проблема:



- * Выбирают точки, которые по мнению исследователей определяют активность. Делают конформационный поиск для всех молекул. Если находят пересечения по геометрии, то на основе этих точек и геометрии пересечения формулируют фармакофор.



» Базы данных:

- * PubChem
- * Cambridge database
- * Inorganic structural database

The image shows two side-by-side screenshots of the PubChem website.

Left Screenshot (Search Interface):

- Header:** PubChem Structure Search, PubChem Compound, Search, Help.
- Search By:** Name (Text), Identity/Similarity, Substructure/Superstructure, Molecular Formula, 3D Conformer, Saved Search.
- Input Fields:** Draw a Structure, CID, SMILES, InChI, Structure File, Enter single structure identifier, Edit, Search, Preview.
- Options:** Options (dropdown: Identical Structures), Filters (dropdown).
- Buttons:** Clear, Save Query.

Right Screenshot (Compound Summary for Ibuprofen):

- Title:** Ibuprofen - Compound Summary (CID 3K7)
- Aliases:** Endol, Motrin, Doloft, Lopan, Hepten, Arthogen, Duramed, Butylone, Ibumetin.
- Molecular Formula:** $C_{12}H_{18}O_2$, **Molecular Weight:** 206.2800, **InChIKey:** HEPVWQDQXAWTRWUHPPAUA.N
- Description:** A nonsteroidal anti-inflammatory agent with analgesic properties used in the therapy of rheumatism and arthritis. From: MeSH
- Table of Contents:** User and Manufacturing, Pharmacokinetics, Biomedicinal Effects and Toxicity, Stability and Stability Indicators, Bioassay and Exposure Potential, Monitoring and Analysis Methods, Literature, Patent, Biomarker Interactions and Pathways, Biological Test Results, Dosage, Chemical and Physical Properties.
- Chemical Structure:** A 2D chemical structure of Ibuprofen is shown, featuring a benzene ring substituted with a propyl group at position 1 and a carboxylic acid group at position 2.
- Identification:** Depester-Applied Synonyms: Ibuprofen, Brufen, Motrin, Doloft, Lopan, Hepten, Arthogen, Duramed, Butylone, Ibumetin.
- Count:** see all 468



```
In [1]: import pubchempy as pcp
```

```
In [21]: name='Aspirin'
list=pcp.get_compounds(name, 'name')
asp=list[0]
hb=asp.h_bond_acceptor_count
form=asp.molecular_formula
xlogp=asp.xlogp
print name, ":", form," hb acceptors: " , hb , " Hydrophobicity: ", asp.xlogp
```

```
Aspirin: C9H8O4 hb acceptors: 4 Hydrophobicity: 1.2
```

```
In [30]: list=[]
for i in range(1,10):
    try:
        a=pcp.get_compounds('C1=N-C=C-N1', 'smiles', searchtype='similarity', listkey_count=50, listkey_start=i)
    except:
        print "ended on: ",i*50
        break
    list.extend(a)

print "Downloaded: ", 50*len(list), "compounds"
```

```
Downloaded: 22500 compounds
```

```
In [31]: list
```

```
Out[31]: [Compound(12749),
Compound(82140),
Compound(484),
Compound(96125),
Compound(283401),
Compound(559542),
Compound(2773261),
Compound(2773328),
Compound(125120)]
```

» ML в хемоинформатике

- * Улучшение анализа HTS данных, в основном регрессии
- * Улучшение предсказания афности, токсичности, фармокинетике для заданных соединений. Регресии и не только.
- * Генерация новых соединений под указанную задачу



» Дескрипторы

- * 0D Формула: молекулярный вес, количество атомов и связей
- * 1D Химические графы: фрагменты, функциональные группы
- * 2D Топология структуры: индексы Weiner, Balaban, Randic, BCUTS
- * 3D Геометрия молекулы: WHIM, autocorrelation, 3D-MORSE, GETAWAY
- * 4D Химическая информация: Volsurf, GRID, Raptor



» 1D

- * Одномерные дескрипторы - это скаляры: количество атомов, количество связей, молекулярный вес, суммы атомных свойств или количество фрагментов
- * Просты в вычислении, страдают от проблем вырожденности, когда различные соединения сопоставляются с идентичными значениями дескриптора
- * Одномерные дескрипторы обычно используются вместе с многомерными дескрипторами или выражаются как вектор из нескольких одномерных дескрипторов.



» 2D

- * Двумерные химические дескрипторы являются наиболее частым типом дескрипторов
- * Включают топологические индексы, молекулярные профили и двухмерные дескрипторы автокорреляции
- * Важной особенностью 2D-дескрипторов является инвариантность графа, когда на значения дескрипторов не влияет перенумерация узлов (вершин) графа.
- * Система *Mold²* быстро генерирует до 200 типов 2D-дескрипторов для больших составных наборов данных.
- * Коммерческие программные пакеты включают систему DRAGON, до 5000 дескрипторов.



» 3D

- * 3D дескрипторы извлекают химические особенности из трехмерных геометрий и наиболее чувствительны структурным изменениям
- * Могут включать дескрипторы автокорреляции, данные об заместителях , дескрипторы поверхности, объема и квантово-химические дескрипторы
- * Трехмерные химические дескрипторы полезны для идентификации «каркасов» - отдельных химических каркасов со сходной связывающей активностью
- * Ключевым ограничением является вычислительная сложность генерации конформеров и выравнивания структур



» 3D

- * Предсказанные конформации могут не соответствовать соответствующим биоактивным конформациям.
- * Химические дескрипторы 4D являются расширением дескрипторов 3D, которые одновременно рассматривают несколько структурных конформаций
- * Эш и Фурчес применили MD киназы ERK2 для вычисления трехмерных дескрипторов по сеткой на основе траектории 20 нс и показали, что такие четырехмерные химические дескрипторы могут эффективно отличать наиболее активные ингибиторы ERK2 от неактивных с более высокой степенью обогащения.

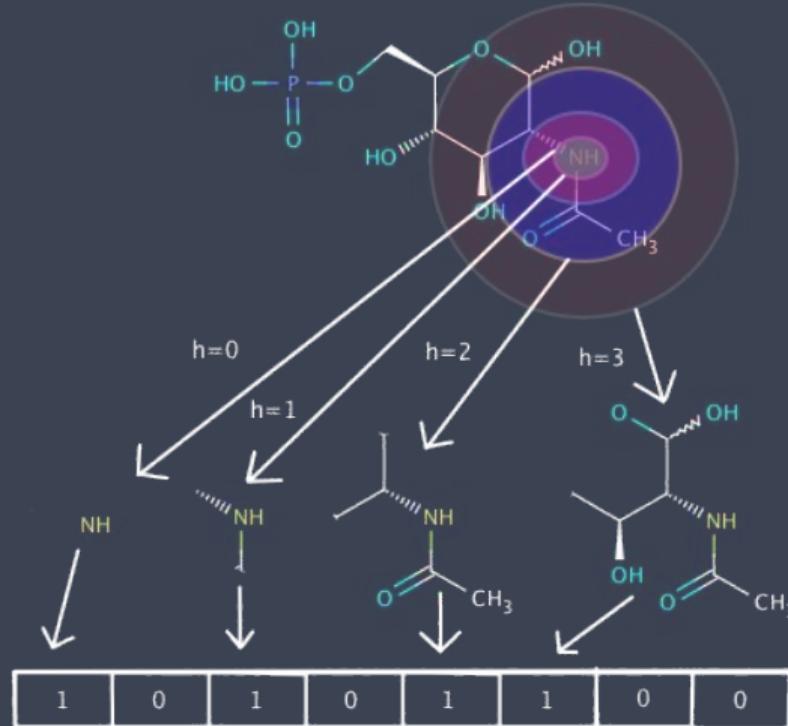


» Fingerprints

- * FP это многомерные векторы, элементами которых являются значения химических дескрипторов
- * MACCS представляют собой двумерные двоичные FP (0 и 1), каждый из которых 166 бит указывает на наличие или отсутствие определенных ключей подструктуры
- * Daylight FP и ECFP позволяют извлекать паттерны до определенной длины или диаметра из графа структуры, и могут динамически индексировать представления с использованием хэш-функций, что часто обеспечивают более высокую специфичность.



» ECFP



10.1021/ci100050t

» Fingerprints

- * Последние разработки - это continuous kernel и встроенные нейронные FP. Это внутренние представления, полученные с помощью SVM и нейронных сетей.
- * Duvenaud et al. распространил концепцию свертки на молекулы, представленные в виде двумерных молекулярных графов.



» 3D fingerprints

- * 3D FP включают химические характеристики, основанные на фармакофорных паттернах, свойствах поверхности, молекулярных объемах или взаимодействия молекул
- * MIF, реализованное в GRID. FP на основе MIF помещает лиганд в сетку с фиксированным интервалом и вычисляет электронный, стерический и гидрофобный вклад независимо в каждой точке сетки.

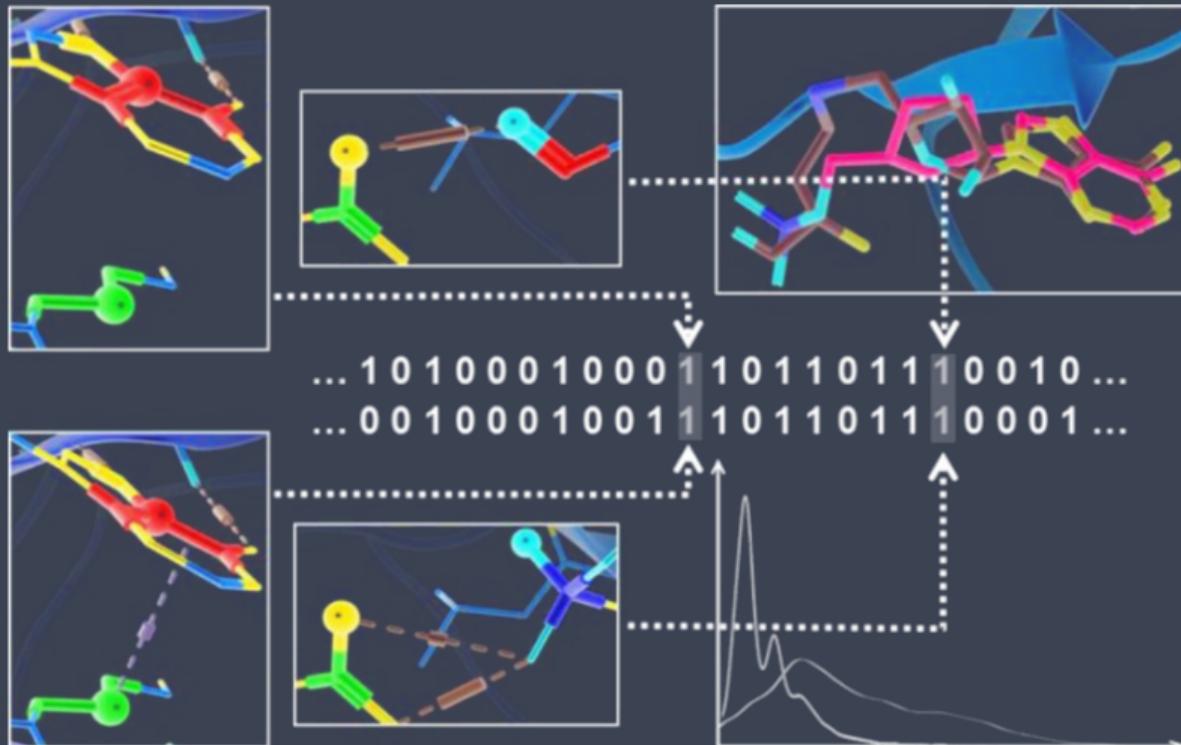


» 3D fingerprints

- * FP на основе MIF можно затем использовать в сравнительном анализе молекулярного поля (CoMFA) путем установления взаимосвязей между точками трехмерной сетки и активностями соединения.
- * Зависимость от относительной ориентации молекул внутри сетки является основным ограничением.
- * Баскин и Жохова недавно представили подход непрерывного молекулярного поля (CMF), который заменяет сетку непрерывной функцией



» SPLIF



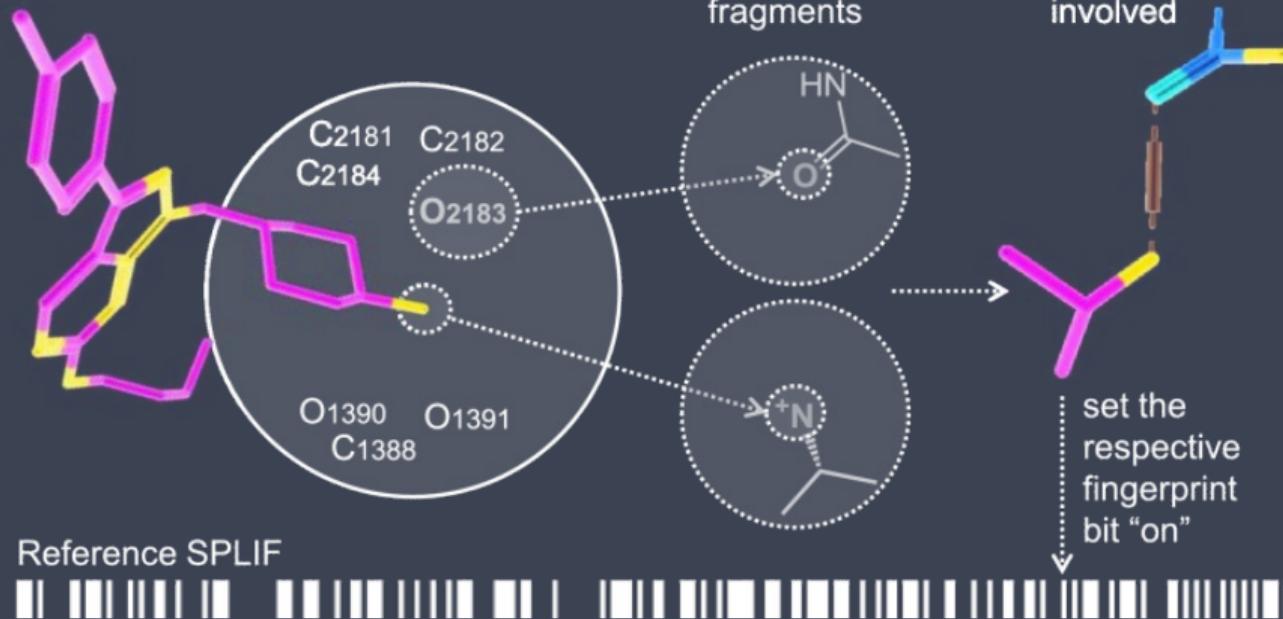
» SPLIF

read 3D coordinates
of the ligand-protein
complex

for each ligand atom,
identify close protein
atoms

expand ligand-
protein atom pairs
to 2D circular
fragments

retrieve 3D
coordinates
for the atoms
involved

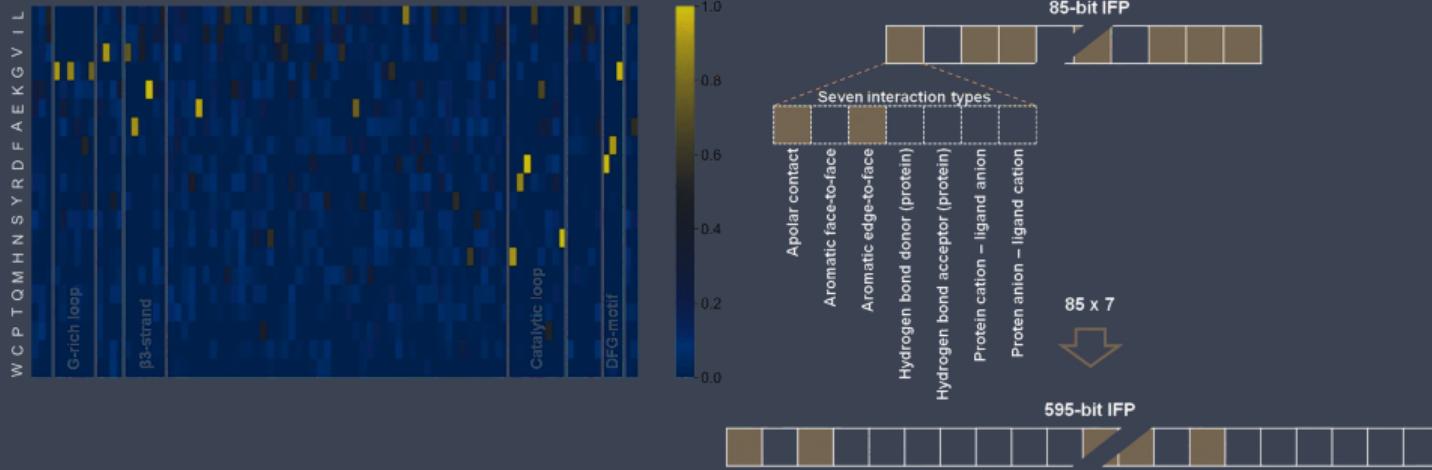


» Ингибиторы киназ

- * Киназы ключевой класс ферментов в передаче сигналов
- * Известно много ингибиторов киназ
- * Сайт связывания очень похож, это связывание ATP



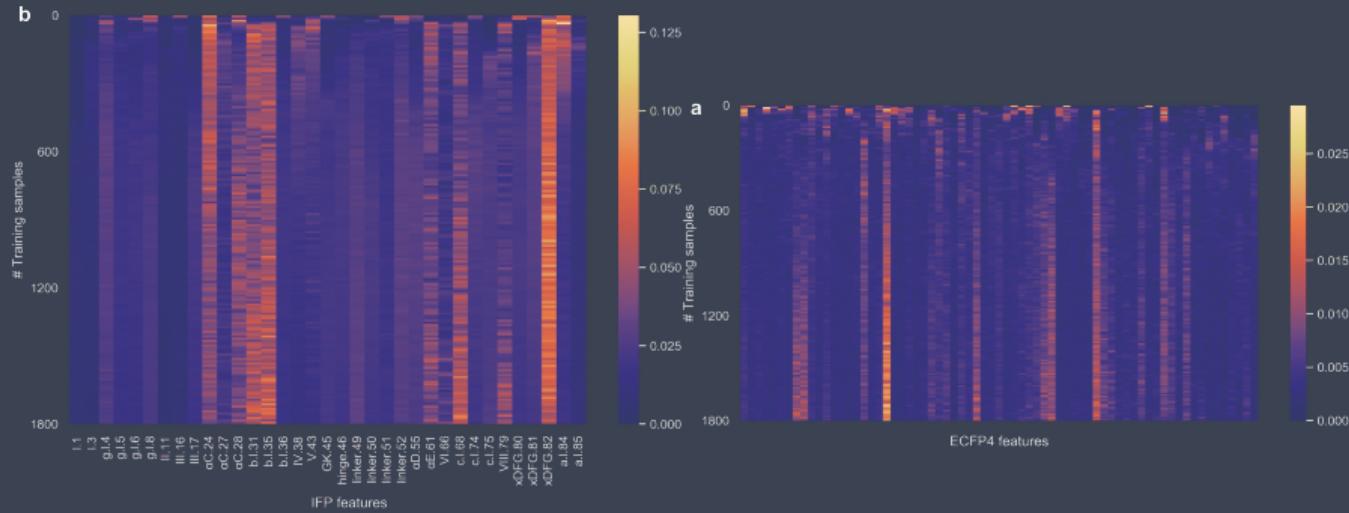
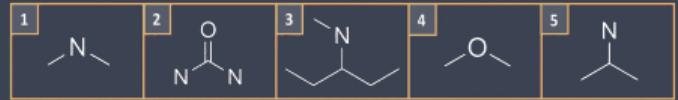
» Interaction fingerprints



Включение мишень-специфичной информации через IPF улучшило предсказание связывания примерно на 10% по сравнению с использованием традиционных FP.
doi.org/10.1186/s13321-020-00434-7



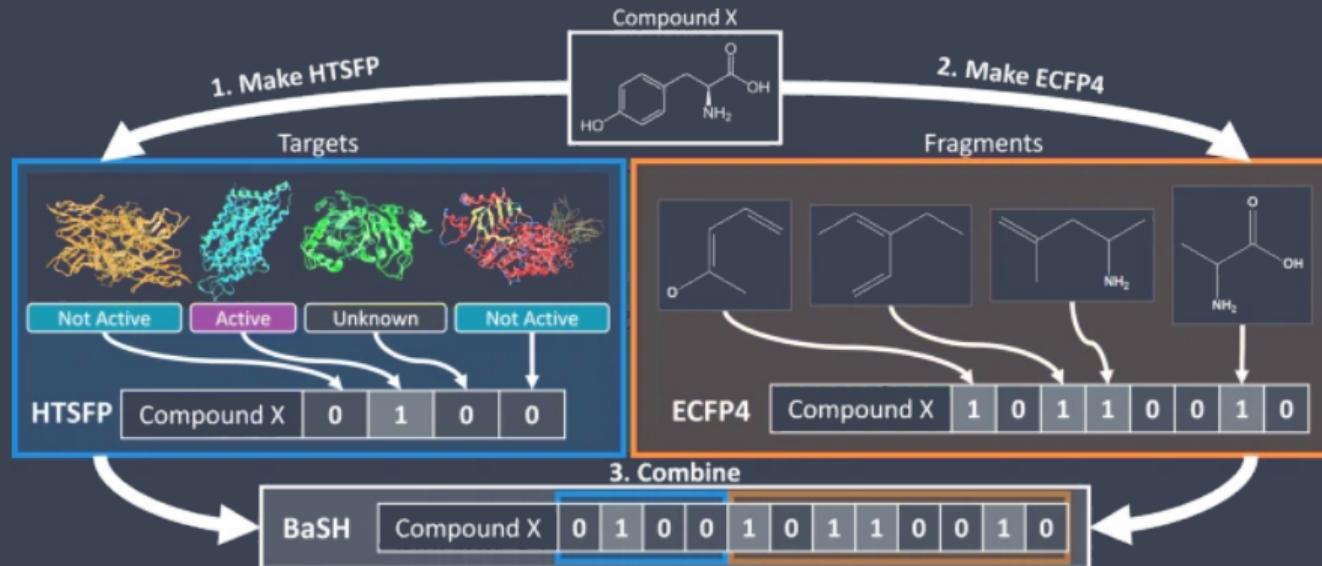
» Features



doi.org/10.1186/s13321-020-00434-7



» Включение bioassay

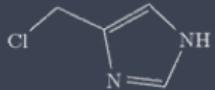


Комбинация структурных и данных об активности позволяет улучшить производительность и показывает эффективный переход между scaffolds
[10.1186/s13321-019-0376-1](https://doi.org/10.1186/s13321-019-0376-1)



» Агент

Graph:

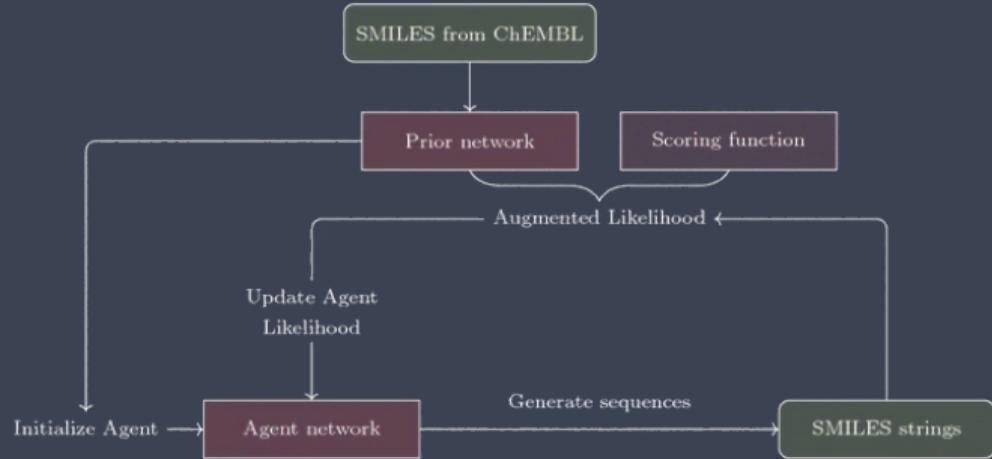


SMILES:

ClCc1c[nH]cn1

One-hot
encoding:

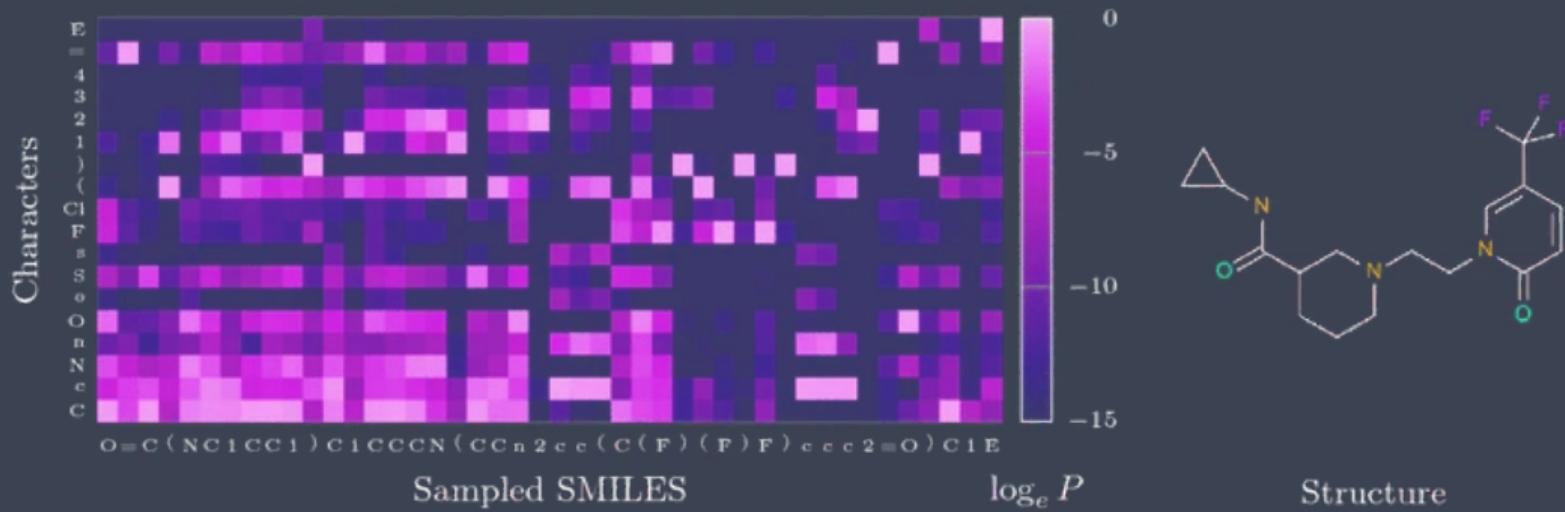
	Cl	C	c	1	c	nH	c	n	1
C	0	1	0	0	0	0	0	0	0
c	0	0	1	0	1	0	1	0	0
n	0	0	0	0	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
nH	0	0	0	0	0	1	0	0	0
Cl	1	0	0	0	0	0	0	0	0



doi.org/10.1186/s13321-017-0235-x



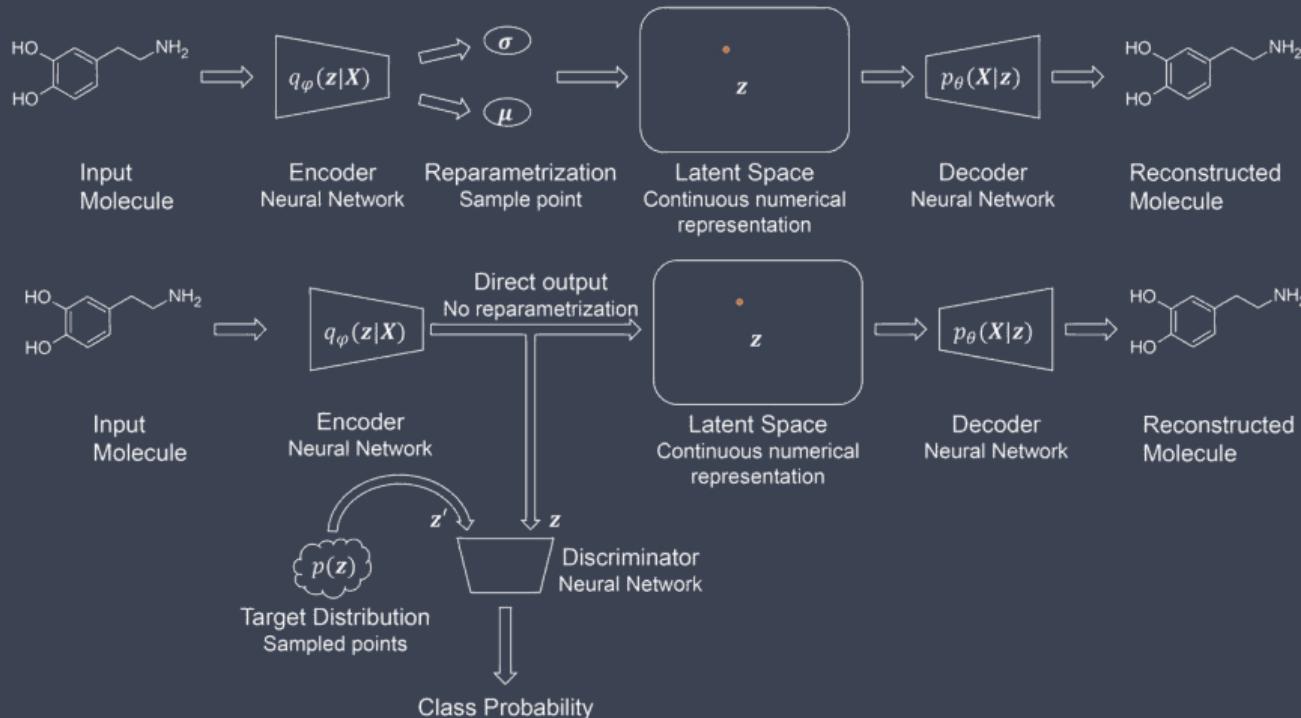
» Генерация по подобию



doi.org/10.1186/s13321-017-0235-x



» Generative Autoencoder



разработан для использования с RNN

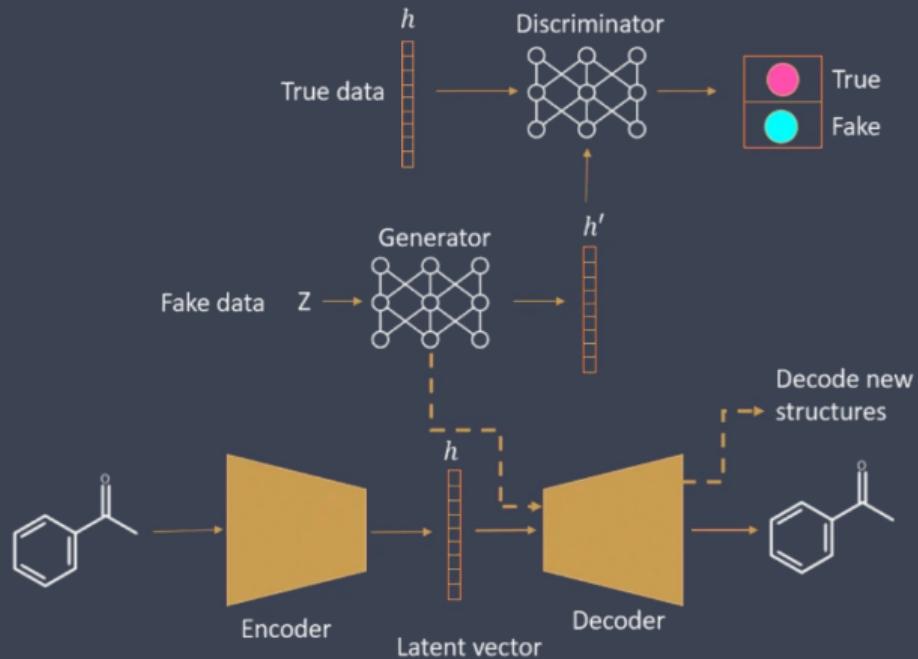


» AE и GAN

- * Предварительно обученный автокодер использовался для сопоставления молекулярной структуры с латентным вектором
- * GAN был обучен с использованием латентных векторов в качестве входных и выходных данных
- * После обучения GAN выбранные скрытые векторы были отображены обратно в структуры



» АЕ и GAN



» DeepChem

- * Имеет множество модулей для Featurization молекул
- * Упрощенное использование TensorFlow, Pytorch и тд
- * Коллекция рецептов



» OpenChem

Model	Module	Task
<p>Smiles2Label</p> <ul style="list-style-type: none"> Prediction of properties from sequential input <p>Graph2Label</p> <ul style="list-style-type: none"> Prediction of property from molecular graphs <p>GraphRNNModel</p> <ul style="list-style-type: none"> Generation of molecular graphs <p>GenerativeRNN</p> <ul style="list-style-type: none"> Generation of SMILES strings 	<p>Embedding</p> <ul style="list-style-type: none"> Token embeddings Positional embeddings <p>Encoder</p> <ul style="list-style-type: none"> Recurrent encoder (RNN, GRU, LSTM) Convolutional encoder Graph convolutional encoder <p>MLP</p> <ul style="list-style-type: none"> Multi-layer perceptron with custom activation functions 	<ul style="list-style-type: none"> Classification Regression Multi-task Generation



» Направления для работы

- * Методы основанные на подобии рассматривают подобные вещества и белки, равномерное распределение отсутствует.
- * Описание *features* сделать количественным.
- * Методы основаны на *datasets*. Нужна адаптация под успешные предсказания.



» Направления для работы

- * Объединение баз данных. Комбинирование максимально доступного количества данных для пары белок-ингибитор.
- * Правильно включение структурно-функциональных данных для лигандов и белков.

