

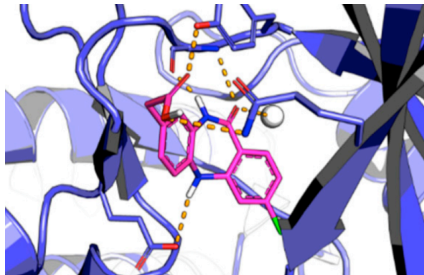
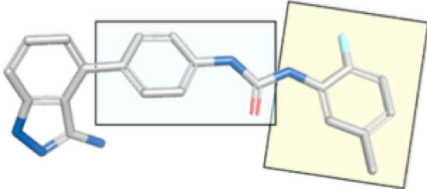
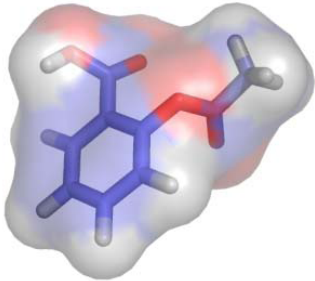
# ХИМИЧЕСКАЯ ИНФОРМАТИКА: ДОПОЛНИТЕЛЬНЫЕ ГЛАВЫ

---

**Лекции 6-7**

Молекулярный дизайн *De Novo*

# COMPUTER-AIDED DRUG DESIGN (CADD)

	Лиганды известны	Лиганды неизвестны
Структура белка известна	<p><b>Structure-based drug design (SBDD)</b></p>  <p>A 3D molecular model showing a protein structure in blue and grey, with a ligand molecule in pink and orange. Dotted lines indicate interactions between the ligand and the protein's binding site.</p>	<p><b><i>De novo</i> design</b></p>  <p>A diagram illustrating the de novo design process. It shows a target molecule (a benzene ring with a blue substituent) being broken down into fragments (a benzene ring, a methylene group, and a carbonyl group) which are then reassembled into the target molecule.</p>
Структура белка неизвестна	<p><b>Ligand-based drug design (LBDD)</b></p>  <p>A 3D molecular model showing a ligand molecule (a benzene ring with a red substituent) docked into a protein's binding site. The protein is represented by a grey surface, and the ligand is shown in blue and red.</p>	<p><b>Experimental data required</b></p>

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН *DE NOVO* IN DRUG DESIGN

Methods' categories

---

Fragment-Based

---

Reaction-Driven

---

Natural product-based

---

Bioisosteric replacement

---

Evolutionary Algorithms

---

Multiobjective Design

---

Pharmacophore-based

---

Machine Learning

---

МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO,  
ОСНОВАННЫЙ НА ФРАГМЕНТАХ

---

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН, ОСНОВАННЫЙ НА ФРАГМЕНТАХ: КАКИЕ СОЕДИНЕНИЯ ОТНОСЯТ К ФРАГМЕНТАМ

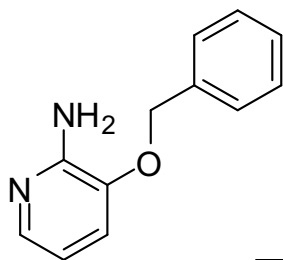
Molecular Weight  $M_r$  ~300 Da

H-bond donors (HBD) <3

H-bond acceptors (HBA) <3

Clog P <3 (a measure of lipophilicity of a compound)

Polar Surface Area (PSA) <60 (a measure of permeability through the cell membrane)



Fragment

$IC_{50} = 1.3\text{mM}$

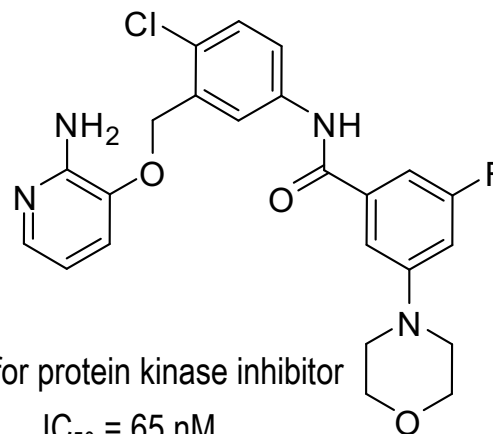
$M_r = 200$

HBD= 2

HBA=3

Clog P=1.92

PSA=48.14




Lead for protein kinase inhibitor

$IC_{50} = 65\text{ nM}$

$M_r = 456$

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН, ОСНОВАННЫЙ НА ФРАГМЕНТАХ: ХАРАКТЕРНЫЕ ОСОБЕННОСТИ



Возможность использовать молекулярные фрагменты как синтоны\* для органического синтеза

Методы определяют не только соединения-кандидаты, но и обоснованную схему синтеза

Если индивидуальные фрагменты являются часто встречающимися в соединениях, подобных лекарству, то их сочетание предположительно является drug-like соединением, химически устойчивым и синтетически доступным

\* реальная или идеализированная структурная единица молекулы, которая может быть введена в химический синтез известными приемами

# FRAGMENT LIBRARIES: ADDITIONAL REQUIREMENTS

Sample relevant chemical space by including pharmacophores that can be responsible for fragment binding

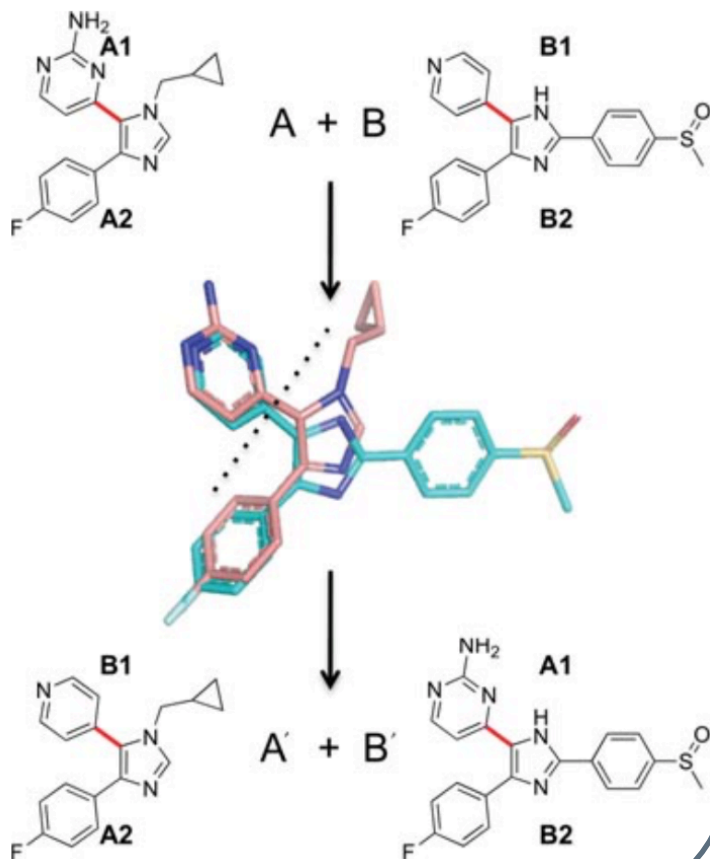
Contain an appropriate size distribution and a balance of differently shaped fragments of appropriate complexity

Contain a diversity of synthetically accessible growth vectors so that fragment hits can be effectively optimized into lead compounds

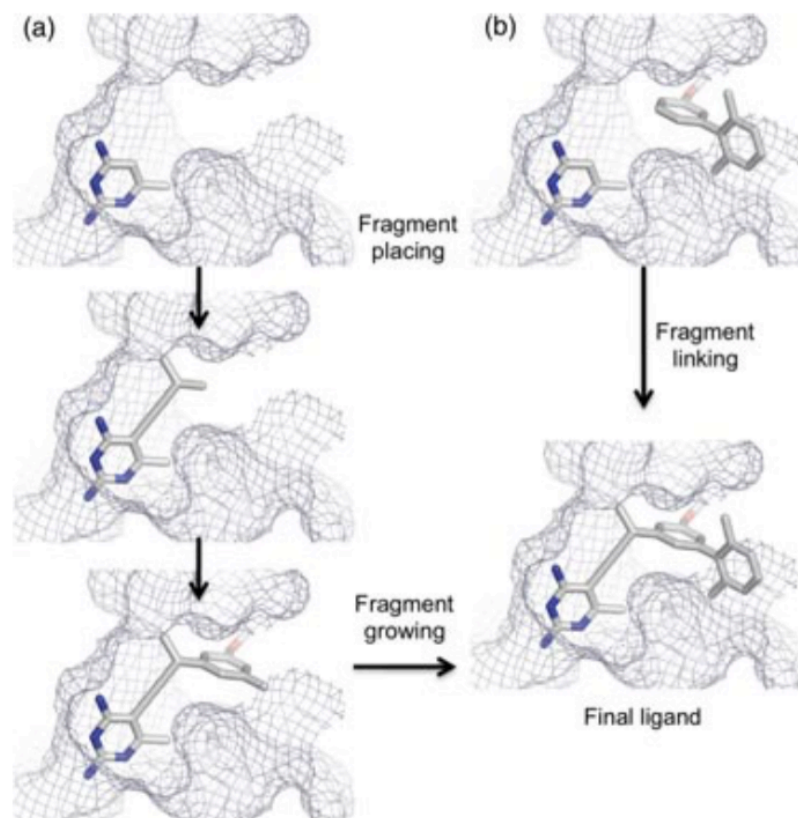
Avoid groups known to be associated with high reactivity, aggregation in solution, or persistent false positive data

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН *DE NOVO*, ОСНОВАННЫЙ НА ФРАГМЕНТАХ

Alignment-based assembly technique



Receptor-based assembly technique



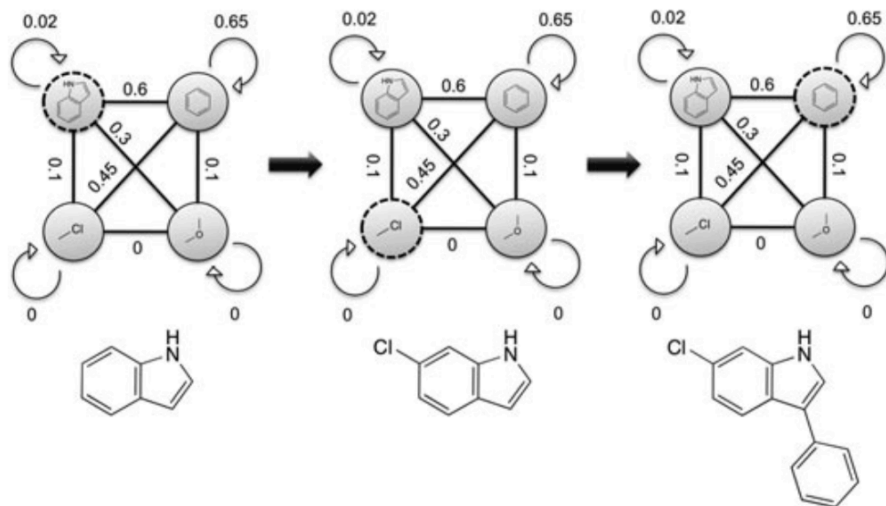
Enabling future drug discovery by *de novo* design.

M. Hartenfeller et al (2011) *Comput Mol Sci*, 1: 742–759. doi:10.1002/wcms.49



# МОЛЕКУЛЯРНЫЙ ДИЗАЙН *DE NOVO*, ОСНОВАННЫЙ НА ФРАГМЕНТАХ

## Ligand-based assembly technique

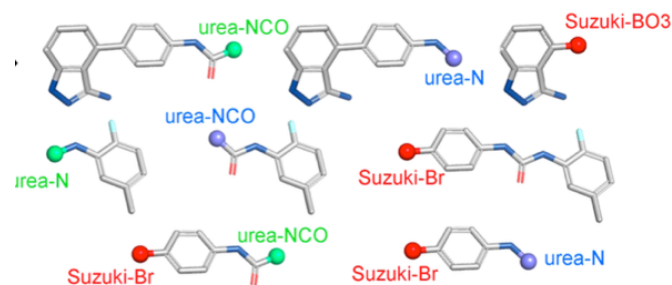


extracting connection frequencies of fragments from training set examples

converting frequencies to probabilities

Walk on a graph where each fragment is equivalent to node (edges between nodes are labelled with transition probabilities)

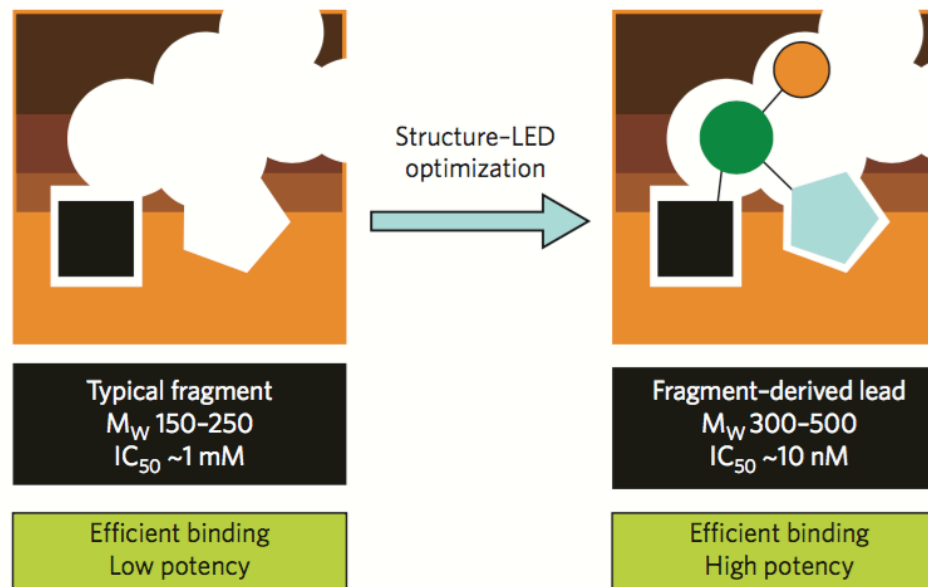
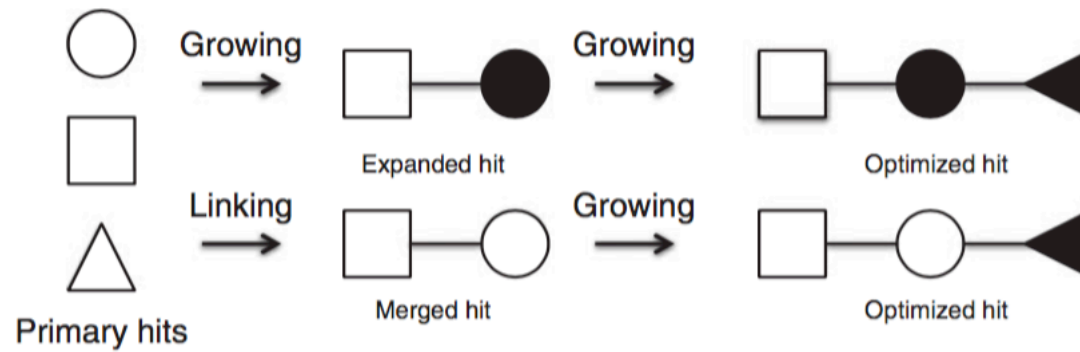
## Reaction-based assembly technique



Enabling future drug discovery by *de novo* design.

M. Hartenfeller et al (2011) *Comput Mol Sci*, 1: 742–759. doi:10.1002/wcms.49

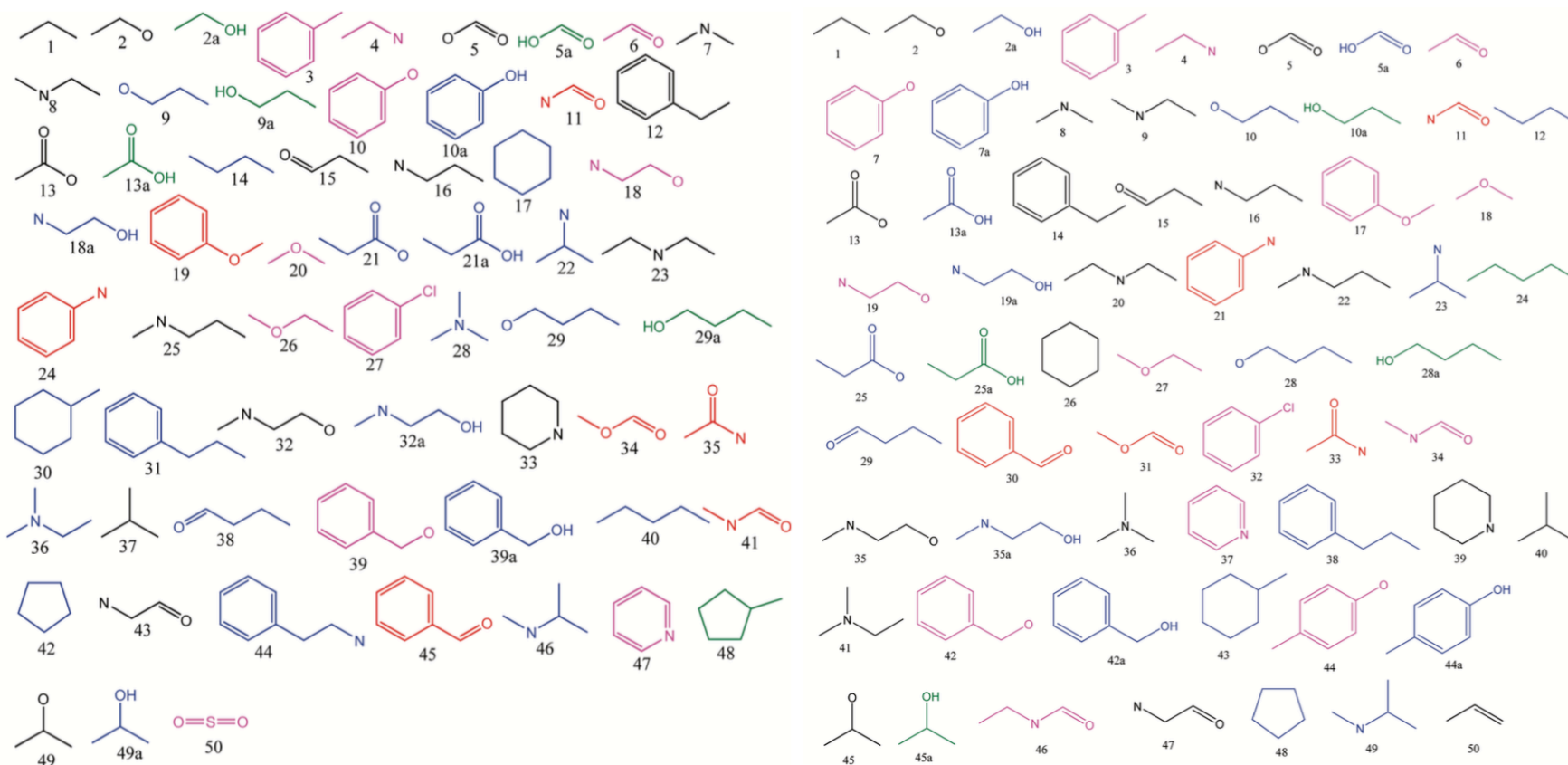
# МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO, ОСНОВАННЫЙ НА ФРАГМЕНТАХ



# FRAGMENT LIBRARIES

ACB Blocks	<a href="http://www.acbblocks.com">www.acbblocks.com</a>	<sup>19</sup> F NMR-oriented, RO3 compliant, predicted to be soluble, purity >96%	1280
Asinex	<a href="http://www.asinex.com">www.asinex.com</a>	fragment library	22524
Beactica	<a href="http://www.beactica.com">www.beactica.com</a>	SPRINT: validated for SPR. 2000 purchased fragments	1946
ChemBridge	<a href="http://www.chembridge.com">www.chembridge.com</a>	ChemBridge Fragment Library, RO3 compliant with predicted solubility; minimum purity 90% by <sup>1</sup> H NMR	>7000
ChemDiv	<a href="http://www.chemdiv.com">www.chemdiv.com</a>	3D designed fragment library	4283
Enamine	<a href="http://www.enamine.com">www.enamine.com</a>	RO3 compliant golden	18108 1794
		fragment library (diverse subset of full library), "simple" fragment library: RO3 compliant ≤20 heavy atoms from screening collection	126597
AnCoreX	<a href="http://www.ancorex.com">www.ancorex.com</a>	MetaKel (metal chelating. MW < 300)	>500
		TCI-Frag (targeted covalent inhibitor fragment screening; mildly reactive functionalities, RO3 compliant)	>100
Key Organics	<a href="http://www.keyorganics.net">www.keyorganics.net</a>	fragment library	26000
		2nd generation with assured aqueous solubility, RO3 compliant	1166
		fragments from nature: RO3 compliant, assured solubility and high Fsp <sup>3</sup> content	183
		CNS fragment library: more stringent filters (e.g., mw <240)	700
Life Chemicals	<a href="http://www.lifechemicals.com">www.lifechemicals.com</a>	general	31000
		RO3 compliant (and subsets of predicted soluble, fluorinated, brominated, and Fsp <sup>3</sup> enriched, covalent and PPI focused)	14000
Maybridge	<a href="http://www.maybridge.com">www.maybridge.com</a>	RO3 compliant diversity fragment library with assured solubility in DMSO and PBS buffer; 1000 fragment subset available	2500
		fragment collection, filtered by purity, mw <350 and substructures	>30000
Otava	<a href="http://www.otavachemicals.com">www.otavachemicals.com</a>	general RO3 compliant, predicted to be soluble	12486
		assured solubility in DMSO and PBS	1000
		fluorine	1217
		metal chelator	1023
		halogen-enriched with bromine for X-ray studies	618
Prestwick Chemical	<a href="http://www.prestwickchemical.com">www.prestwickchemical.com</a>	Prestwick Fragment Library mainly derived from drug fragments, RO3 compliant	910

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН *DE NOVO*, ОСНОВАННЫЙ НА ФРАГМЕНТАХ: НАИБОЛЕЕ ЧАСТО ВСТРЕЧАЮЩИЕСЯ DRUG-LIKE МАЛЫЕ ФРАГМЕНТЫ



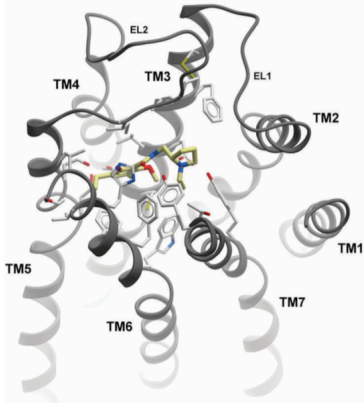
\* цвет соответствует относительной частоте представленности в двух рассматриваемых базах данных

## Drug and Drug Candidate Building Block Analysis

J. Wang et al *Journal of Chemical Information and Modeling* 2010 50 (1), 55-67 DOI: 10.1021/ci900398f

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO, ОСНОВАННЫЙ НА ФРАГМЕНТАХ: ОГРАНИЧЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ МЕТОДОВ

Сложности применения методологии к GPCR рецепторам и ионным каналам (при отсутствии информации о структуре)



**Рецепторы, сопряжённые с G-белком** выполняют функцию активаторов внутриклеточных путей передачи сигнала, связанными с физиологическими и патофизиологическими реакциями, влияющими на иммунитет, сердечно-сосудистую и эндокринную системы

Только около 30 GPCR структур расшифрованы (> 800 присутствуют в человеческом геноме)

Expanding the horizons of G protein-coupled receptor structure-based ligand discovery and optimization using homology models

*Cavasotto et al Chem. Commun.*, 2015, 51, 13576--13594

Необходимость адаптации высокопроизводительного скрининга для идентификации связывания фрагментов (обладают меньшей афинностью по отношению к мишени=>требуют высоких концентраций и количества материала)

В последние годы активно для решения этих задач используются следующие методы:

- ❖ functional/high-concentration screening
- ❖ fluorescence-based thermal shift assays (TSA)
- ❖ Поверхностный плазмонный резонанс (SPR)
- ❖ Масс-спектрометрия (MS)
- ❖ Ядерный магнитный резонанс (NMR)
- ❖ Рентгеноструктурный анализ

МОЛЕКУЛЯРНЫЙ ДИЗАЙН:  
СТОХАСТИЧЕСКИЕ АЛГОРИТМЫ

---

# МНОГОЗАДАЧНЫЙ МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO ПОСРЕДСТВОМ АДАПТИВНОЙ ПРИОРИТЕЗАЦИИ ФРАГМЕНТОВ

Angew. Chem. Int. Ed. 2014, 53, 4244–4248

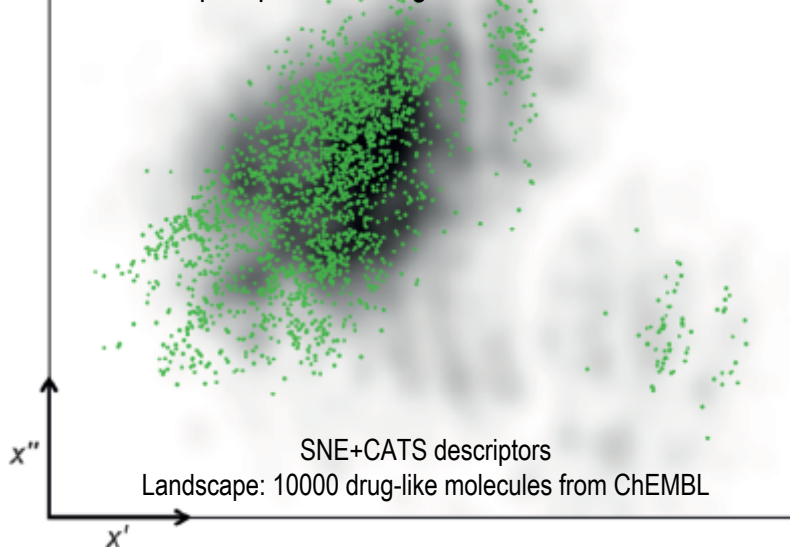
## Составляющие метода молекулярного дизайна:

- ❖ Схема синтеза
- ❖ Метод прогнозирования величины афинности полученных продуктов
- ❖ Метод оптимизации «строительных» блоков

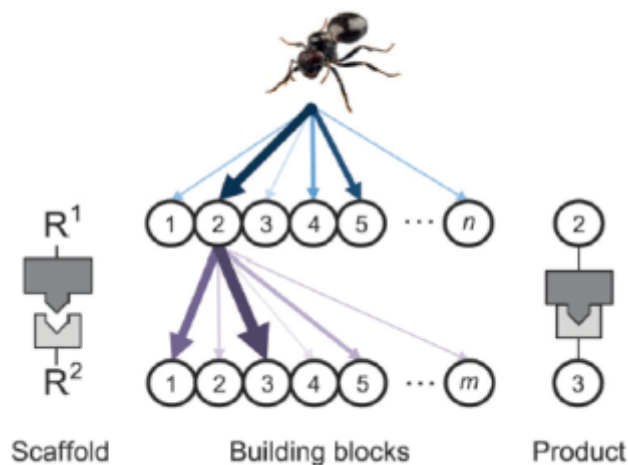
## В этой работе:

Реакция восстановительного аминирования (альдегиды/кетоны и амины в качестве строительных блоков)

Распределение 5000 виртуальных продуктов реакции восстановительного аминирования в химическом пространстве drug-like соединений



## Метод муравьиных колоний



Affinity prediction: Gaussian process (GP) regression for 640 human targets annotated in ChEMBL (v14), [10] based on ~280000 compounds with ~570000 measured bioactivities.

Selected candidates to match different criteria (“positive design”):

- 1) Potent and selective ( $\sigma_1$ ) or multitarget-modulating (dopamine D4) ligands
- 2) Target-subtype-selective ligands
- 3) Exploratory molecules, lying outside the training domain as expressed by Morgan fingerprint Tanimoto similarities  $< 0.20$
- 4) Inactive compounds that are nearest neighbors to known high-affinity ligands in ChEMBL bioactivity space.

МОЛЕКУЛЯРНЫЙ ДИЗАЙН:  
БИОИЗОСТЕРНОЕ ЗАМЕЩЕНИЕ

---



# МОЛЕКУЛЯРНЫЙ ДИЗАЙН: БИОИЗОСТЕРНОЕ ЗАМЕЩЕНИЕ



*Irving Langmuir*  
1881-1957

*Изостеры – это молекулы или ионы, содержащие одинаковое число атомов, а также имеющие одинаковое количество и расположение электронов.*

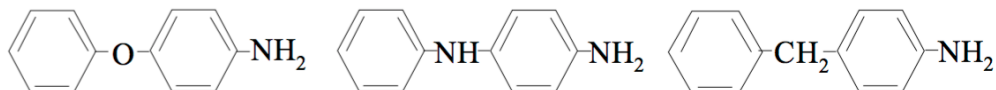
Закон гидридного сдвига Гримма

*“Атомы в любом месте периодической системы в пределах четырех клеток до инертного газа при присоединении к ним от одного до четырех атомов водорода изменяют свои свойства таким образом, что образующиеся комбинации ведут себя как псевдоатомы, аналогичные элементам в группах, находящимся справа от них от одной до четырех клеток соответственно”*

Группа				
IV	VI	V	VII	0
C	N	O	F	Ne
	CH	NH	OH	FH
		CH <sub>2</sub>	NH <sub>2</sub>	OH <sub>2</sub>
			CH <sub>3</sub>	NH <sub>3</sub>
				CH <sub>4</sub>

Ганс Эрленмайер

*“атомы, ионы или молекулы, в которых наружные электронные оболочки могут считаться идентичными” -> биологическая активность*

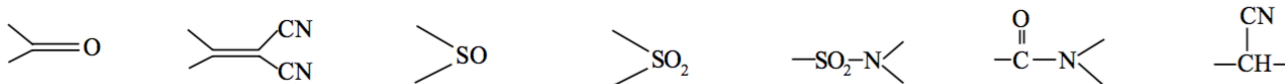


# МОЛЕКУЛЯРНЫЙ ДИЗАЙН: ПРИМЕРЫ НЕКЛАССИЧЕСКИХ БИОИЗОСТЕРОВ

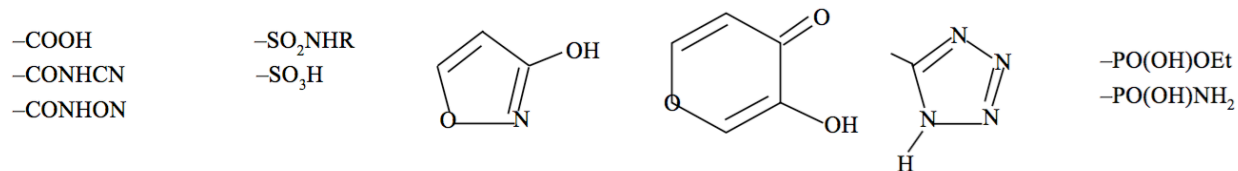
Корвин Ганч

Биоизостеры - соединения, обладающие сходными гидрофобными, электронными и стерическими признаками, образующих их подструктур

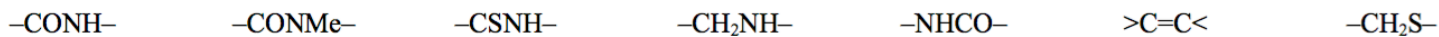
Карбонильная группа



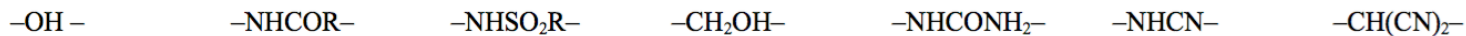
Карбоксильная группа



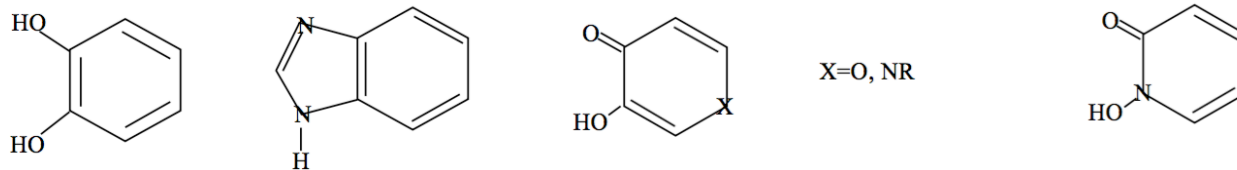
Амидная группа



Гидроксильная группа



Пирокатехин



# МОЛЕКУЛЯРНЫЙ ДИЗАЙН: БИОИЗОСТЕРНОЕ ЗАМЕЩЕНИЕ

Г. Фридман (1951)

*«Биоизостеры – все соединения, которые удовлетворяют самому широкому определению изостеров и имеют тот же тип биологической активности»*

Торнбер (1957)

*«Биоизостеры – это группы или соединения, характеризующиеся химическим и физическим сходством и оказывающие сходные биологические эффекты»*

К. Ганч (1964)

*«Биоизостеры - соединения, вызывающие идентичный биохимический или фармакологический ответ в стандартной тест- системе»*

## **Возможности использования биоизостерного замещения:**

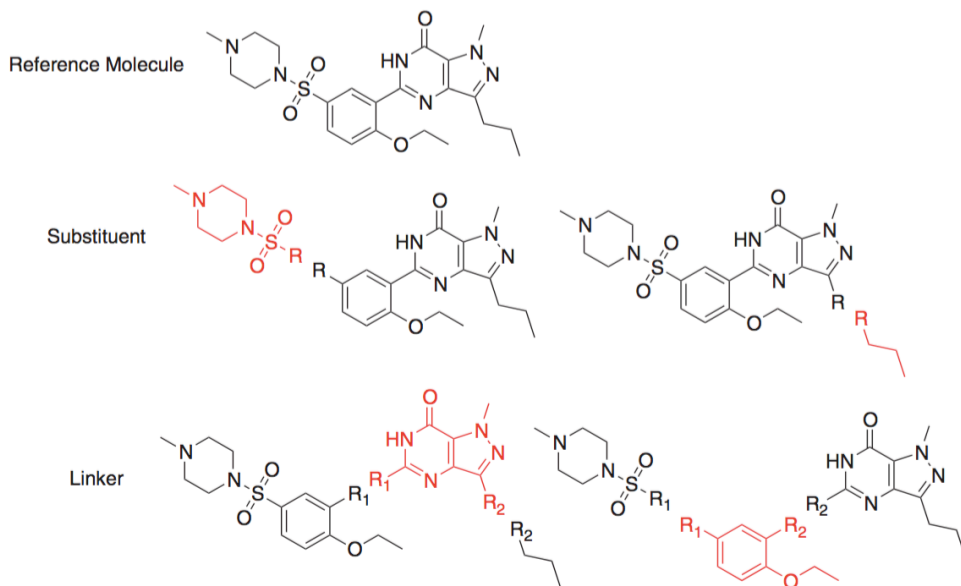
- ❖ Улучшение селективности
- ❖ Удаление нежелательных свойств (побочных эффектов, токсичности)
- ❖ Улучшение фармакокинетических характеристик (растворимость/гидрофобность)
- ❖ Оптимизация синтеза
- ❖ ...

# МОЛЕКУЛЯРНЫЙ ДИЗАЙН: БИОИЗОСТЕРНОЕ ЗАМЕЩЕНИЕ

Fragmentation

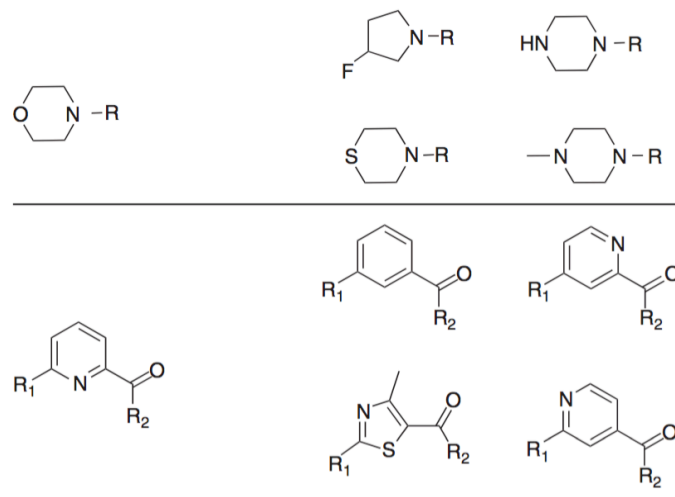


Replacement



Original fragment

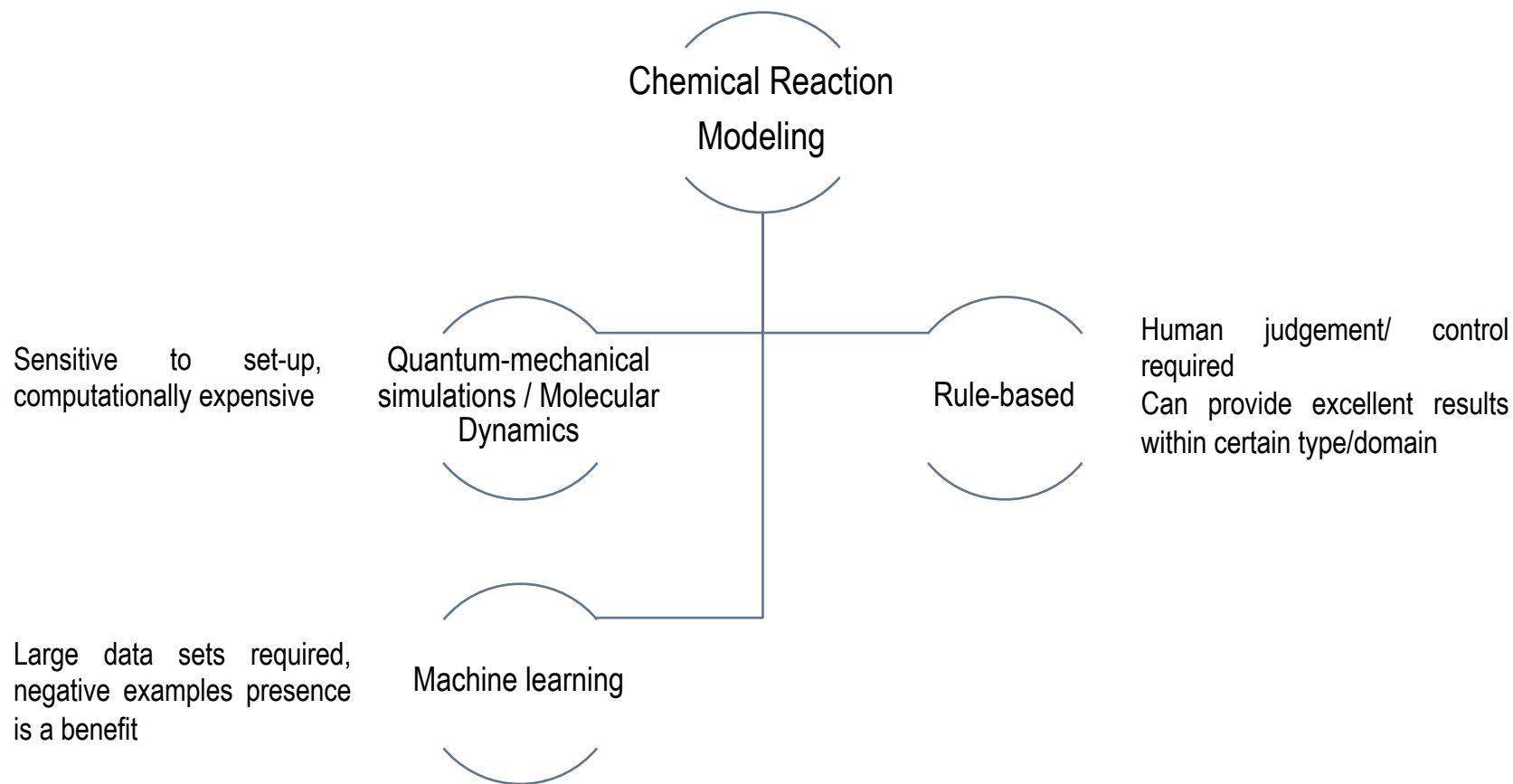
Replacement fragment



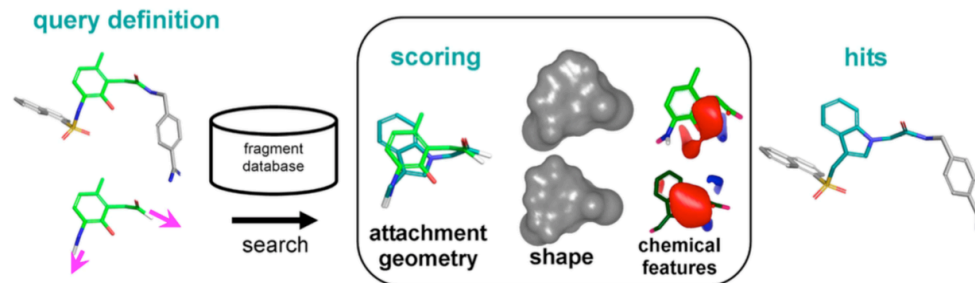
МОЛЕКУЛЯРНЫЙ ДИЗАЙН НА ОСНОВЕ МОДЕЛИРОВАНИЯ  
ХИМИЧЕСКИХ РЕАКЦИЙ

---

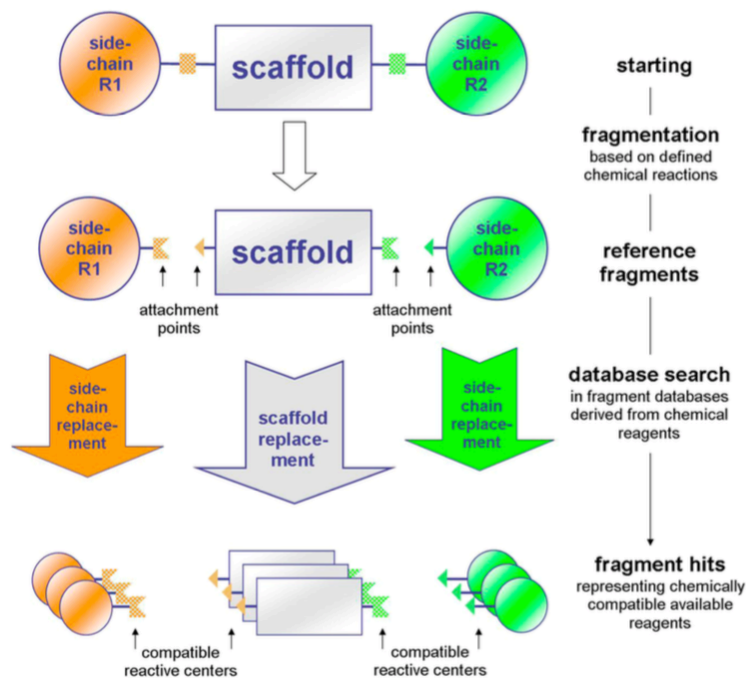
# МОЛЕКУЛЯРНЫЙ ДИЗАЙН НА ОСНОВЕ МОДЕЛИРОВАНИЯ ХИМИЧЕСКИХ РЕАКЦИЙ



# REACTION-DRIVEN RESCAFFOLDING AND SIDE-CHAIN OPTIMIZATION



Общая схема фрагментации исходных соединений и идентификация реагентов и реакционных центров для замены молекулярного скелета и заместителей

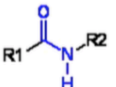
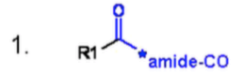
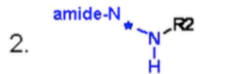
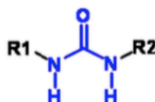
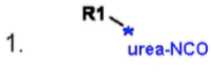

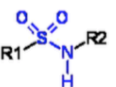
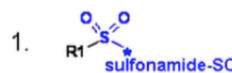


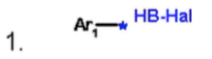

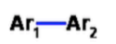
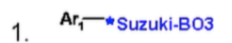




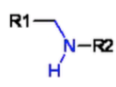
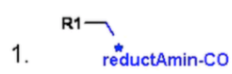


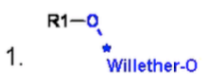



**CROSS: An Efficient Workflow for Reaction-Driven Rescaffolding and Side-Chain Optimization Using Robust Chemical Reactions and Available Reagents**

A. Evers et al *Journal of Medicinal Chemistry* 2013 56 (11), 4656-4670 DOI: 10.1021/jm400404v

# REACTION-DRIVEN RESCAFFOLDING AND SIDE-CHAIN OPTIMIZATION

Химические реакции, применяемые для ретросинтетической декомпозиции исходных соединений. С точкой разрыва ассоциирована информация о способах получения и реакционных группах.

cleavable bond	retrosynthetic decomposition	cleavable bond	retrosynthetic decomposition
 amide	1.  2. 	 urea	1.  2. 
 sulfonamide	1.  2. 	 Hartwig-Buchwald	1.  2. 
 Suzuki	1.  2. 	 Mitsunobu phenol	1.  2. 
 reductive amination	1.  2. 	 Williamson ether	1.  2. 

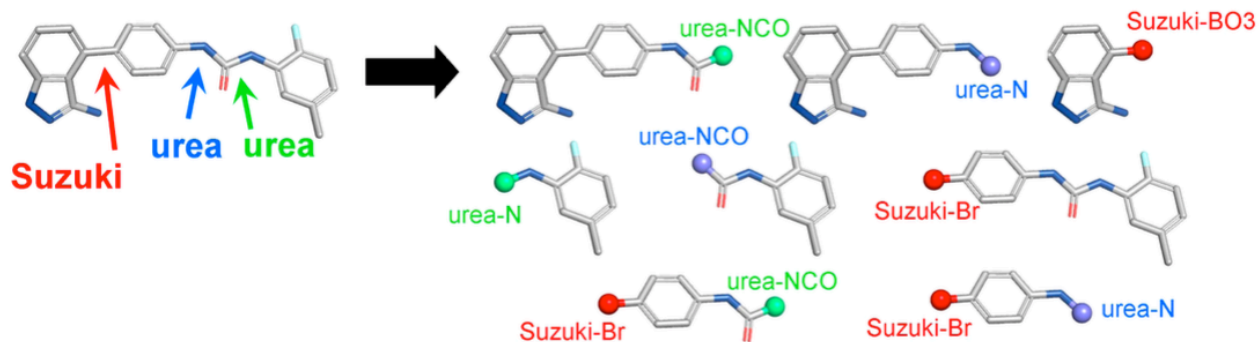
**CROSS: An Efficient Workflow for Reaction-Driven Rescaffolding and Side-Chain Optimization Using Robust Chemical Reactions and Available Reagents**

A. Evers et al *Journal of Medicinal Chemistry* 2013 56 (11), 4656-4670 DOI: 10.1021/jm400404v

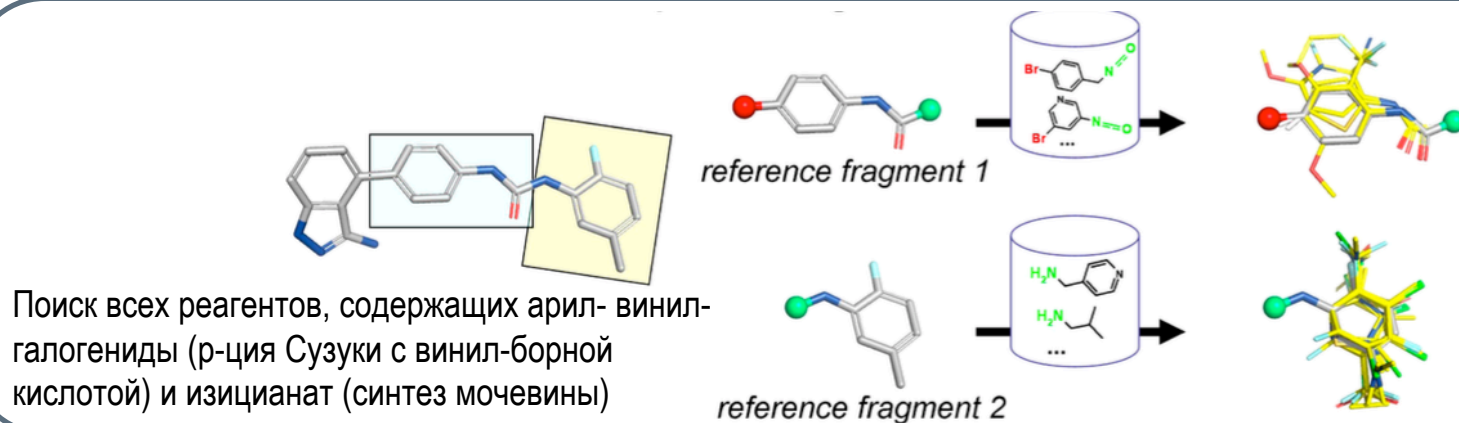


# REACTION-DRIVEN RESCAFFOLDING AND SIDE-CHAIN OPTIMIZATION

Ретросинтетический анализ исходных соединений



Определение фрагментов для замены, поиск реагентов



**CROSS: An Efficient Workflow for Reaction-Driven Rescaffolding and Side-Chain Optimization Using Robust Chemical Reactions and Available Reagents**

A. Evers et al *Journal of Medicinal Chemistry* **2013** 56 (11), 4656-4670 DOI: 10.1021/jm400404v

# SCUBIDOO (Screenable Chemical Universe Based on Intuitive Data OrganizatiOn ): DATABASE OF COMPUTATIONALLY GENERATED SYNTHETIC TRACTABLE COMPOUNDS

~18000  
«строительных  
блоков»

58 наиболее  
распространенных  
в медицинской  
химии реакций

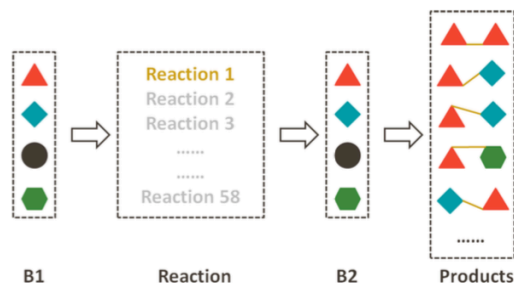
21 миллион новых  
соединений

Предварительная фильтрация:

- ❖ Удаление дубликатов
- ❖ Удаление противоионов
- ❖ Контроль «молекулярной сложности»:

сложности»:

- $MW \leq 250$  Da. (для продуктов с большей вероятностью  $MW \leq 500$  Da).
- Количество конформационных связей  $\leq 2$ . (для продуктов  $\leq 6$  => возможность применения методов молекулярного докинга)
- Количество хиральных центров  $\leq 1$  (для продуктов  $\leq 3$  => упрощает синтез)



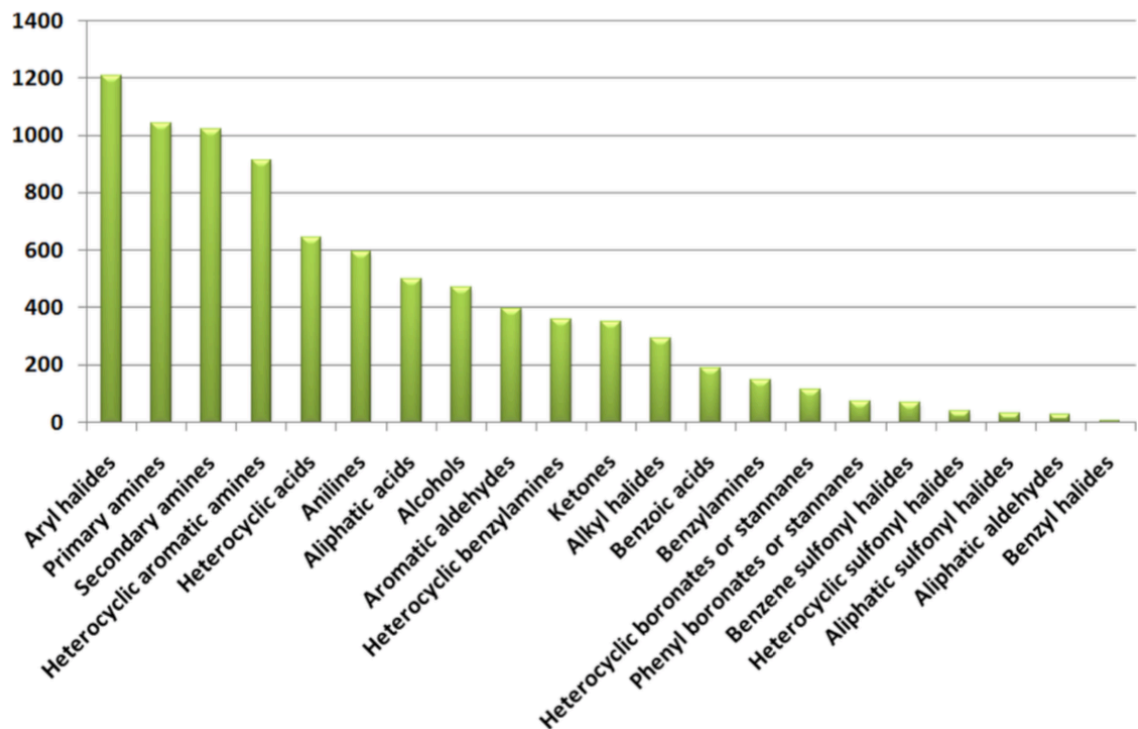
Для каждого соединения в базе данных представлена информация по методу синтеза, потенциальным побочным реакциям, и альтернативным способам получения

**SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability**

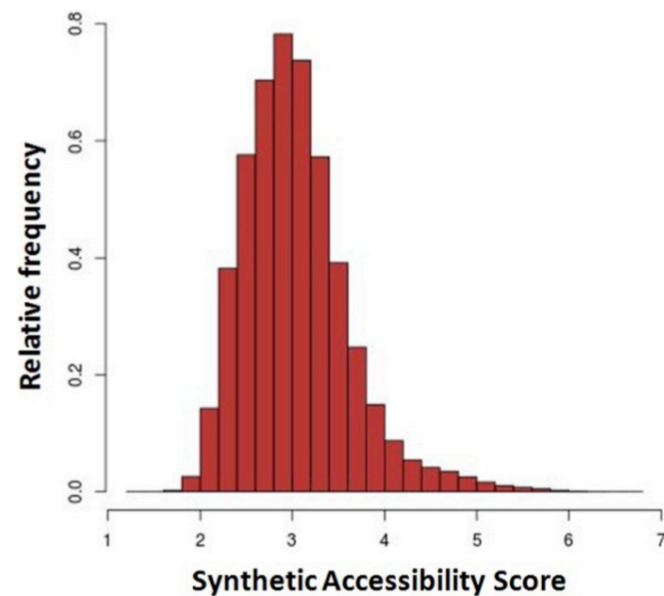
F. Chevillard et al *Journal of Chemical Information and Modeling* 2015 55 (9), 1824-1835 DOI: 10.1021/acs.jcim.5b00203

# SCUBIDOO (Screenable Chemical Universe Based on Intuitive Data OrganizatiOn ): DATABASE OF COMPUTATIONALLY GENERATED SYNTHETIC TRACTABLE COMPOUNDS

Классы реагентов («строительные блоки»)



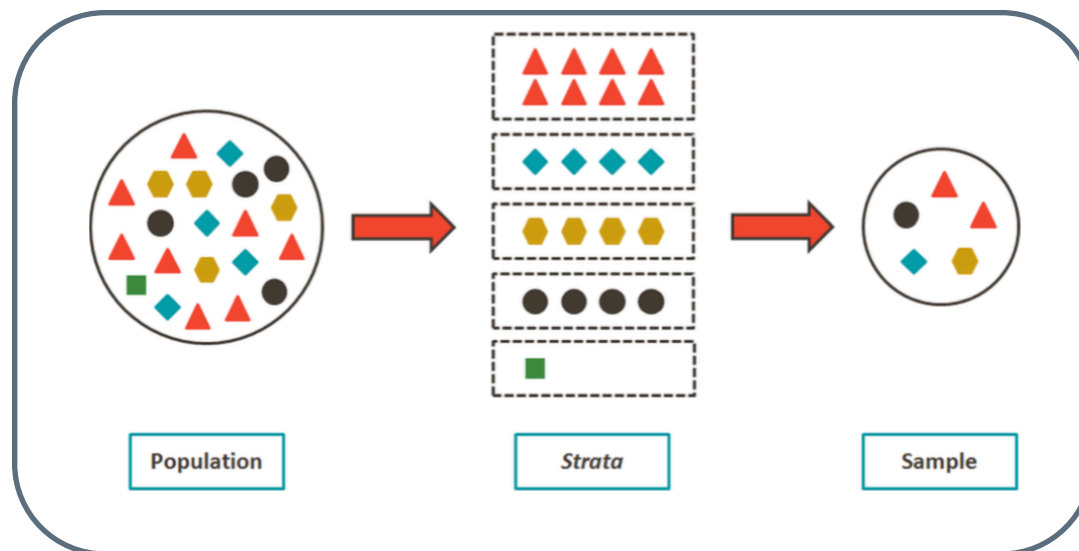
Синтетическая доступность



**SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability**

F. Chevillard et al *Journal of Chemical Information and Modeling* 2015 55 (9), 1824-1835 DOI: 10.1021/acs.jcim.5b00203

# SCUBIDOO (Screenable Chemical Universe Based on Intuitive Data OrganizatiOn ): DATABASE OF COMPUTATIONALLY GENERATED SYNTHETIC TRACTABLE COMPOUNDS



- ❖ Для получения максимально разнородной и репрезентативной набора данных были выделены подгруппы, называемые «стратами».
- ❖ В этой работе каждая страта определяется реакциями, каждый продукт относится только к одной страте.
- ❖ Для получения репрезентативного набора данных по продуктам химических реакций использовались простейшие молекулярные дескрипторы (molecular weight, logP, number of H-bond donors, number of H-bond acceptors, and topological polar surface area) и метод сбалансированного сэмплинга (Jean-Claude Deville, Yves Tillé; Efficient balanced sampling: The cube method. *Biometrika* 2004; 91 (4): 893-912. doi: 10.1093/biomet/91.4.893)
- ❖ Размер каждой страты определялся пропорционально представленности в исходной базе данных

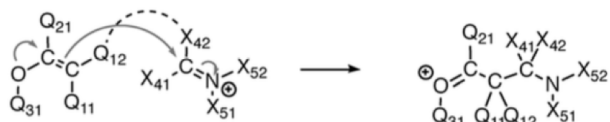
**SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability**

F. Chevillard et al *Journal of Chemical Information and Modeling* 2015 55 (9), 1824-1835 DOI: 10.1021/acs.jcim.5b00203

# DEEP LEARNING FOR CHEMICAL REACTION PREDICTION

Database: 11000 elementary reactions (SMILES/SMIRKS format)

## Combinatorial Reaction Generation



Q21 = [H],C,CC,C9=CC=CC=C9,OC,OCC,N(C)C,SC

Q11 = [H],C,CC,C7=CC=CC=C7,C(=O)C,C(=O)OC,C(=O)N(C)C  
Q12 = [H],C,CC

X41 = [H],C,C8=CC=CC=C8,OC,OCC,SC,C#N,C(=O)C  
X42 = [H],C,CC

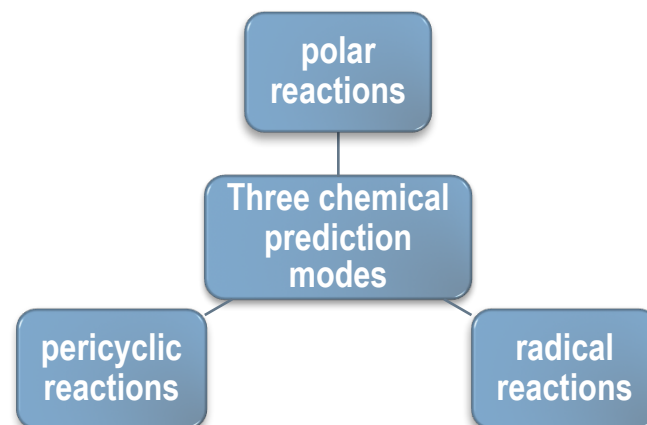
X51 = C,CC  
X52 = C,CC

## Pipeline:

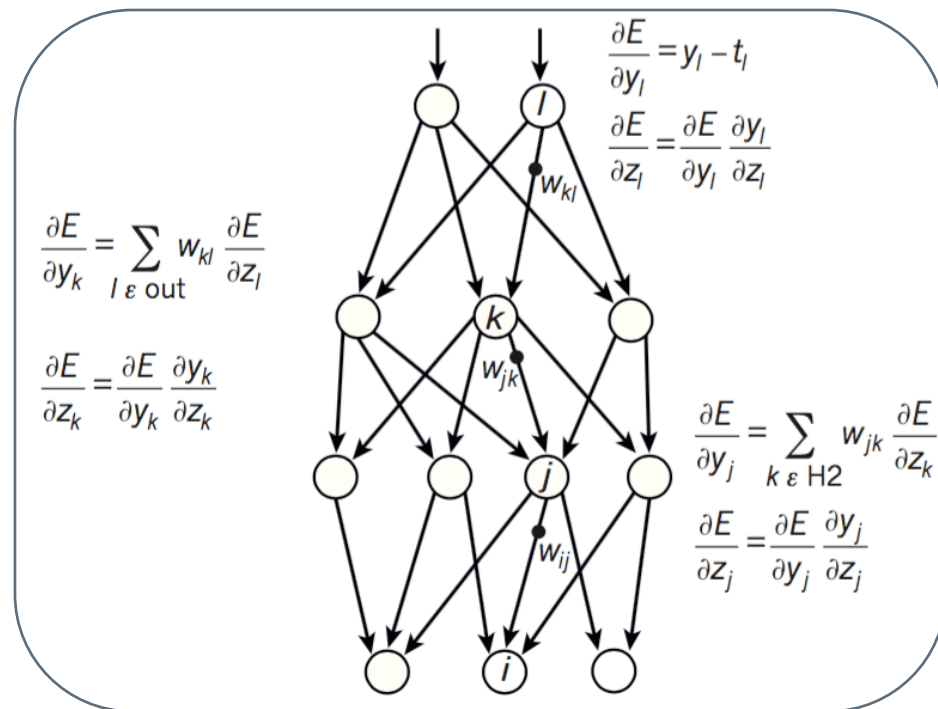
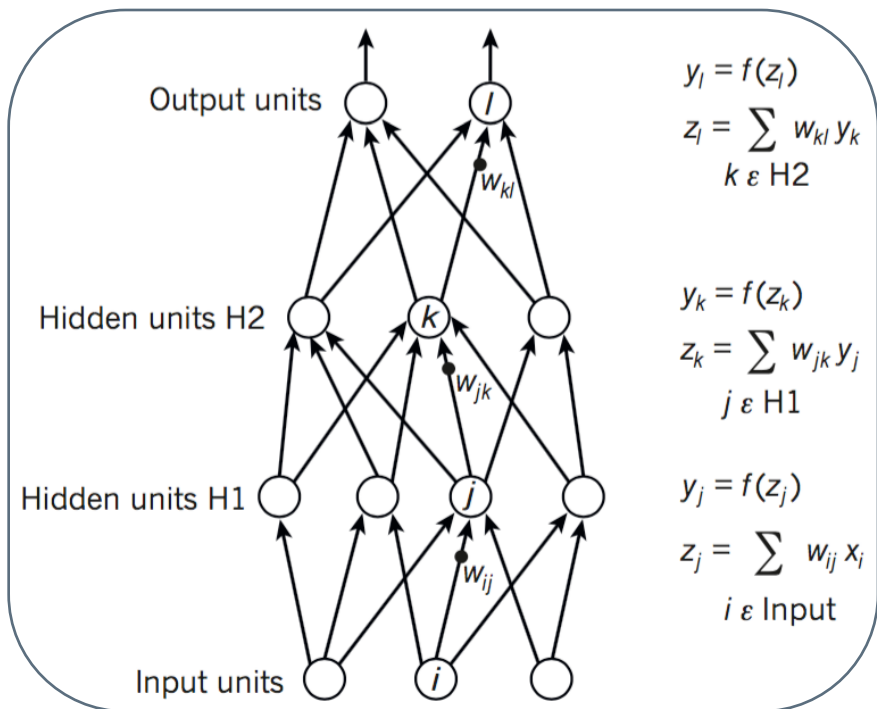
- Identify core molecular template and appropriate electron movement
- Varying substituents
- Random sampling of reaction subsets for each mechanism

## Deep Learning

- Enumerate all possible electron sources and electron sinks within the input reactant molecules
- Filter the list of candidate sources and sinks, predicting a smaller list containing only the most reactive sources and sinks.
- Propose reactions by enumerating all combinations of source-sink pairings.
- Rank the proposed reactions by favorability.
- Iterate the above process to identify global reactions, or search for unidentified products.

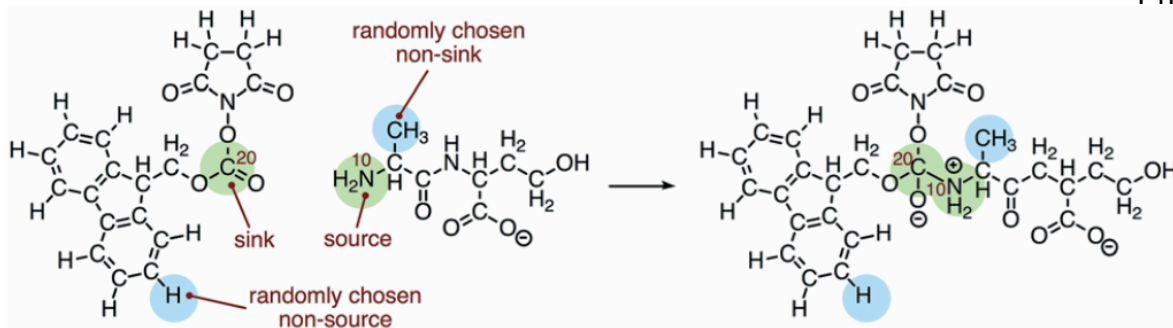


# DEEP LEARNING FOR CHEMICAL REACTION PREDICTION



# DEEP LEARNING FOR CHEMICAL REACTION PREDICTION

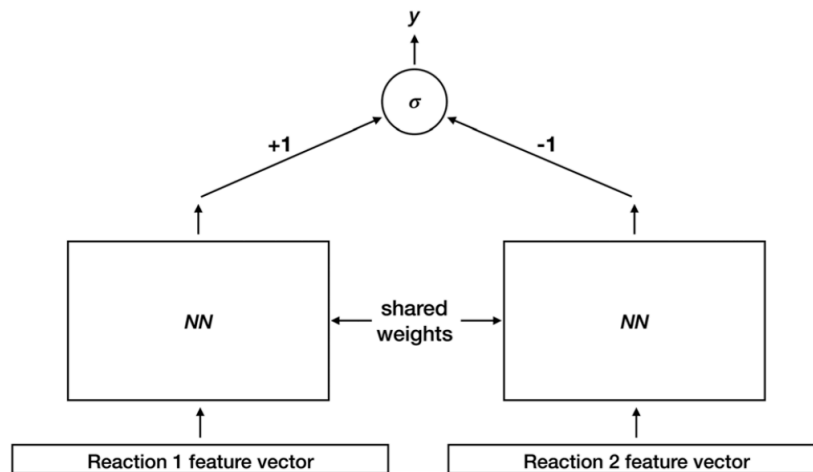
## Input Data Pre-Processing



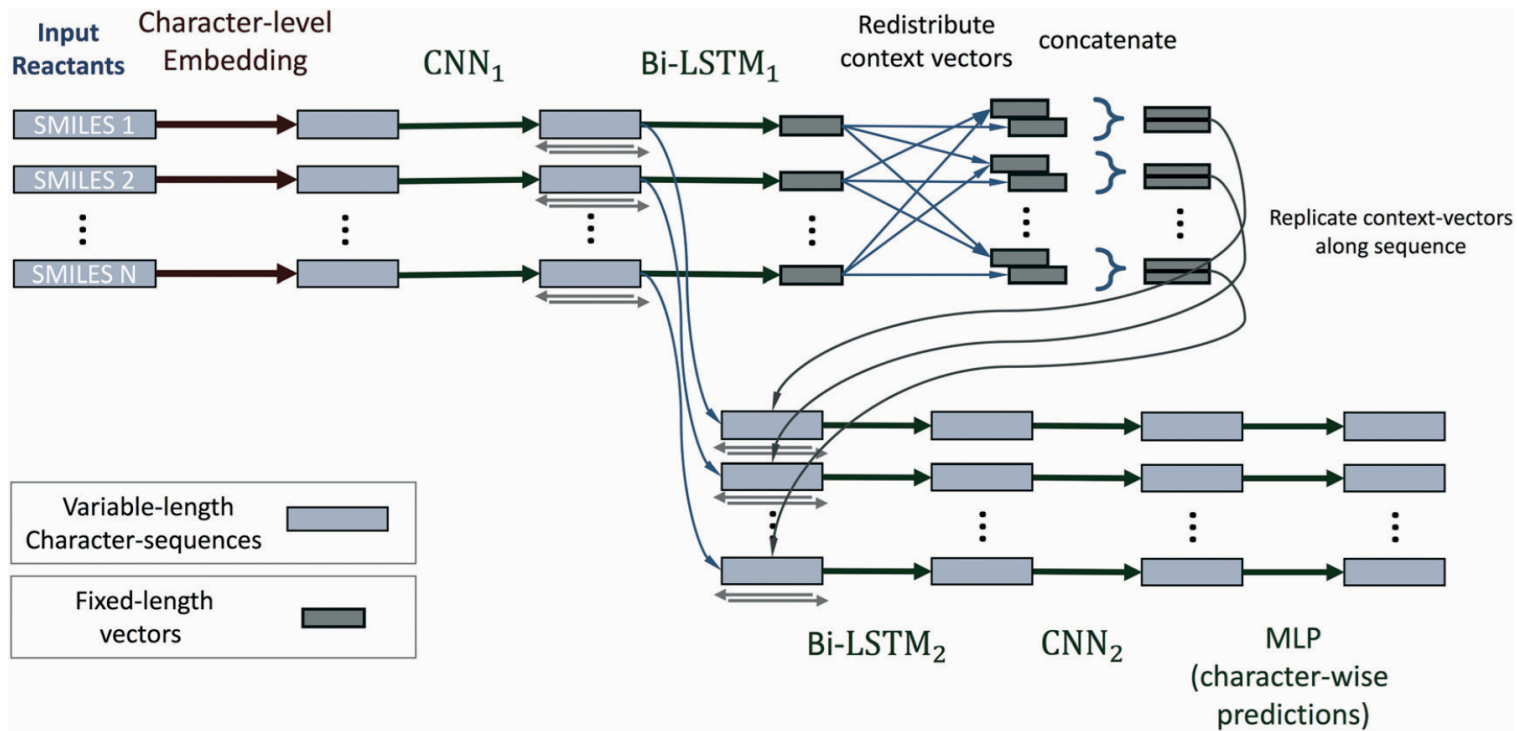
## Physicochemical and graph-topological descriptors

- Combination of concatenated source and sink atomic level features
- Type of orbital involved
- Net change features (dynamic bonds)

## Reaction Ranking Using Siamese Neural Network



# DEEP LEARNING FOR CHEMICAL REACTION PREDICTION

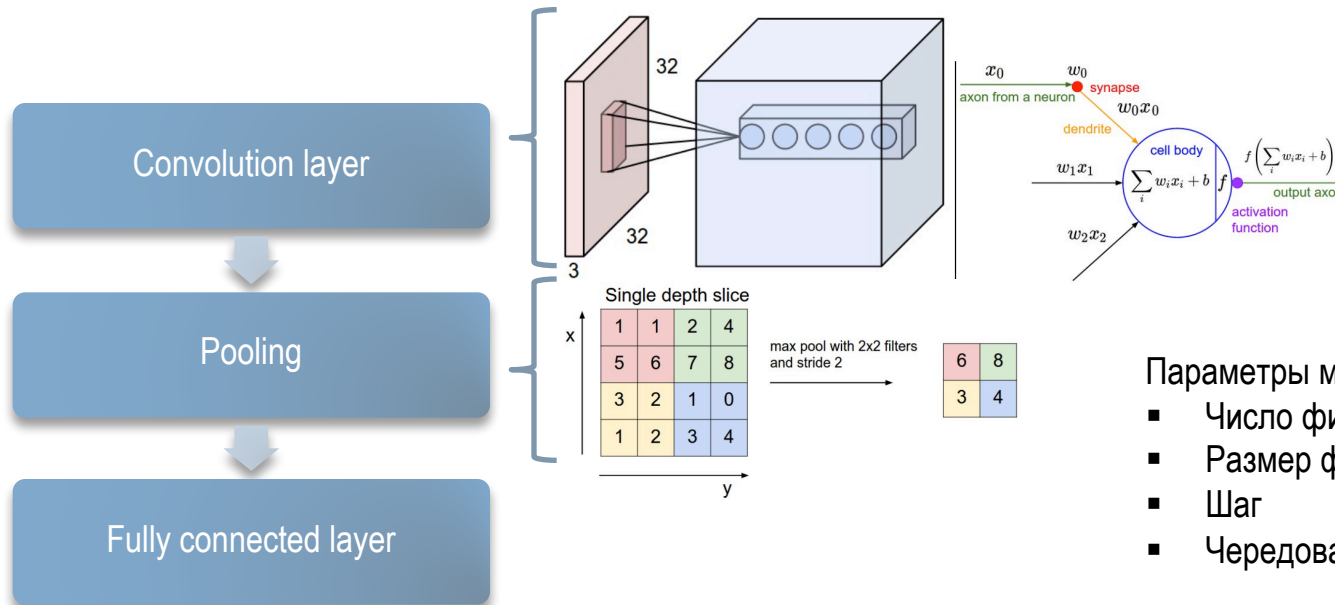
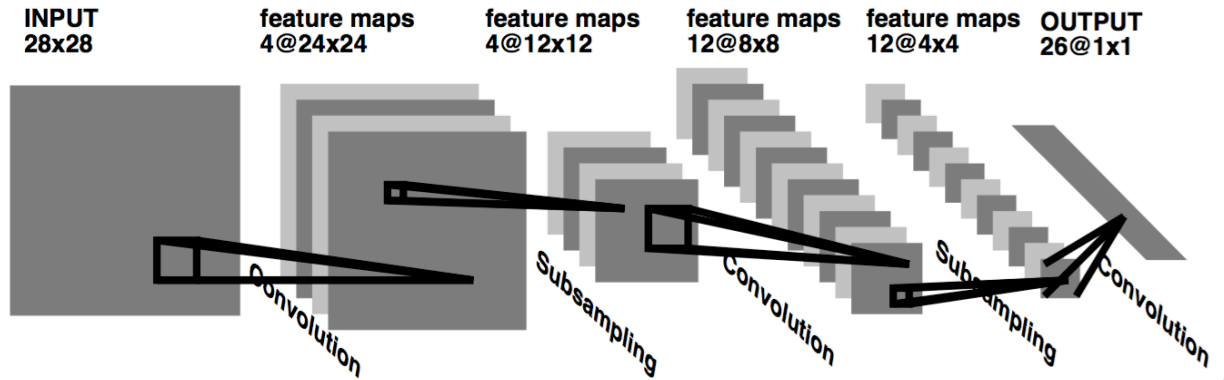




# CONVOLUTIONAL NEURAL NETWORKS

Specific characteristics:

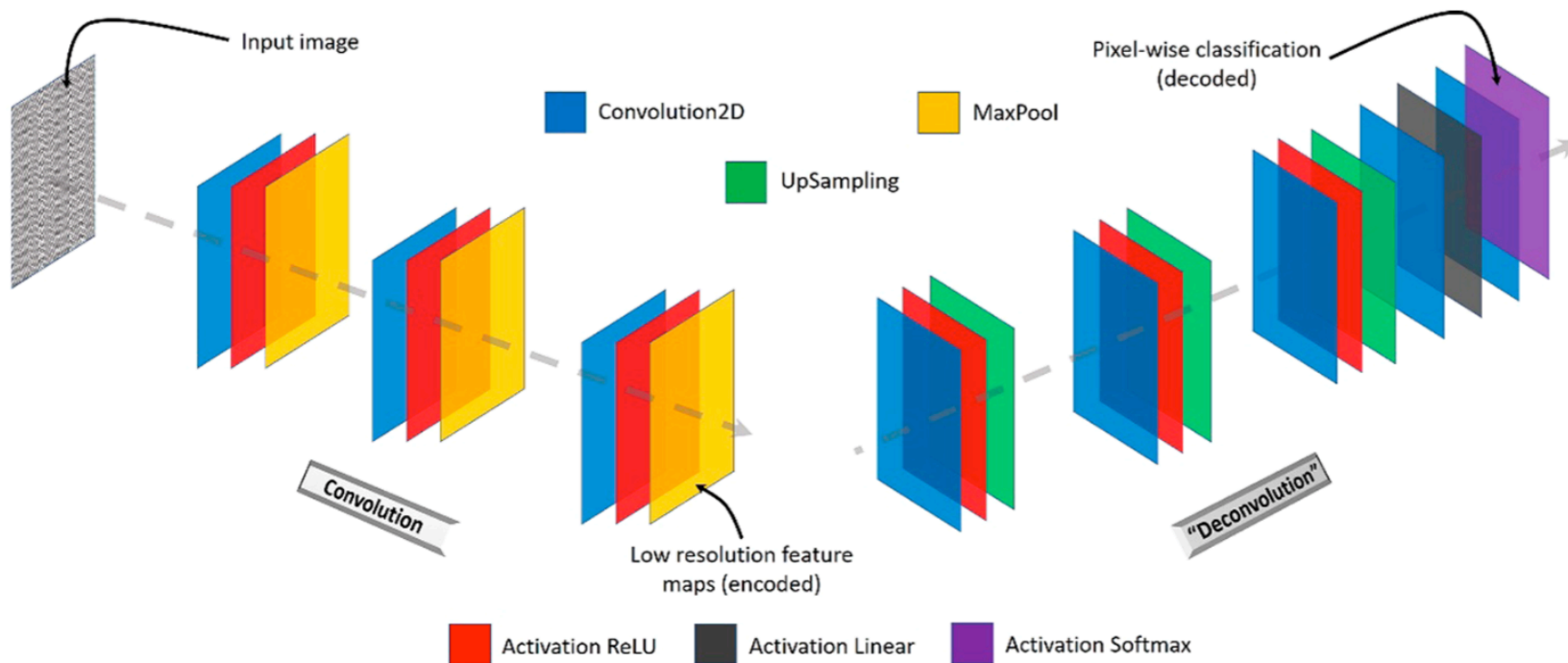
- local connections
- shared weights
- pooling
- using many layers



Параметры метода:

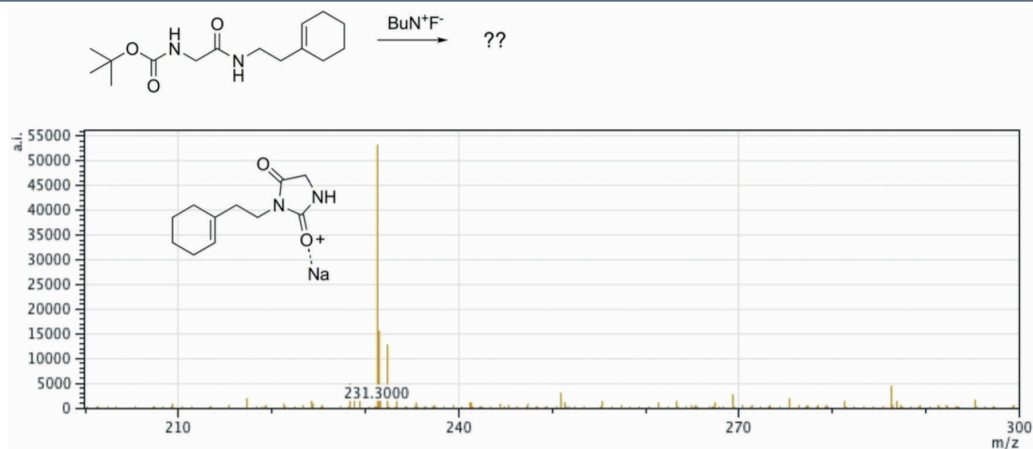
- Число фильтров
- Размер фильтров
- Шаг
- Чередование слоев

# АНАЛИЗ ЛОКАЛЬНОЙ СТРУКТУРЫ МАТЕРИАЛА: ИСПОЛЬЗОВАНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

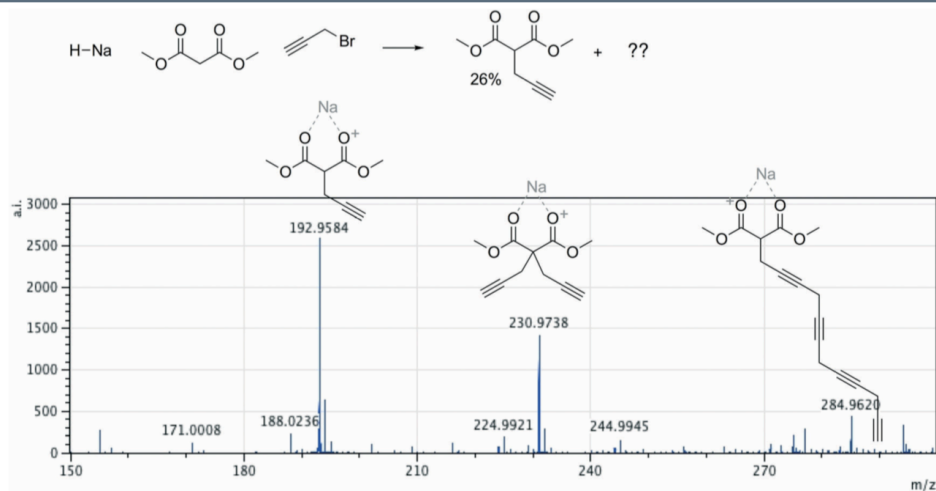


# DEEP LEARNING FOR CHEMICAL REACTION PREDICTION

## Identifying unexpected product



## Reaction optimization: identification of malonate overalkylation by propargyl bromide

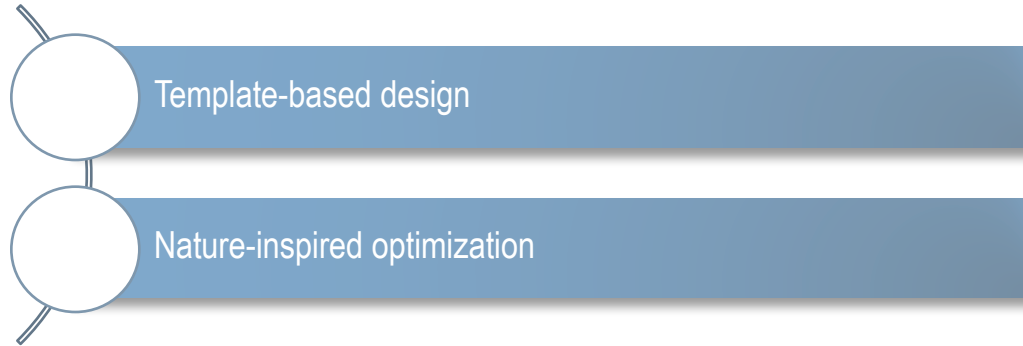


МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO:  
МАКРОМОЛЕКУЛЫ

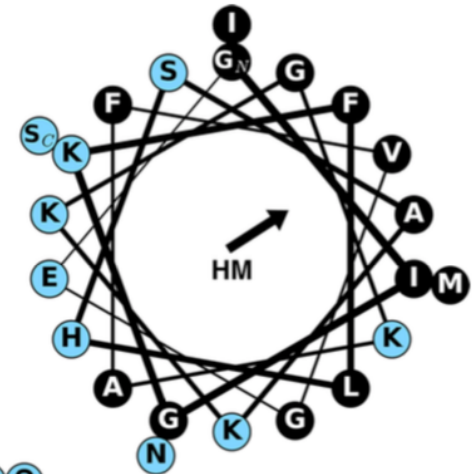
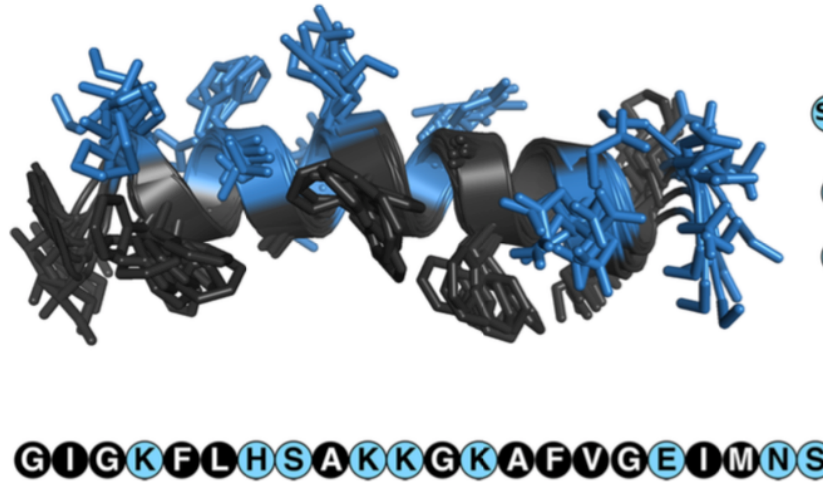
---

# COMPUTER-ASSISTED DESIGN OF MACROMOLECULES: PEPTIDES

Easily synthesized, often selective for macromolecular targets (e.g. membrane receptors, ion channels)



The hydrophobic amino acid side chains - black  
the polar residues - blue

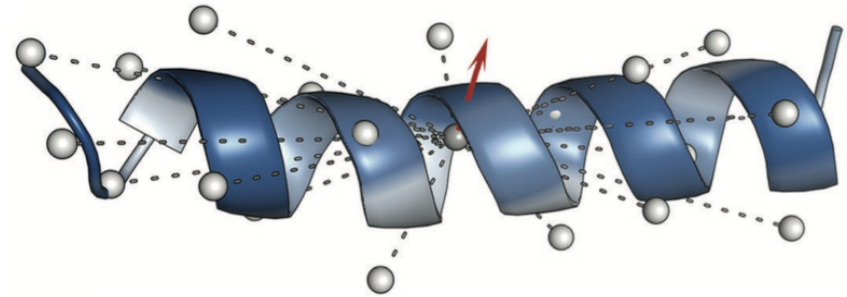


helical wheel plot of magainin 2

# RATIONAL DESIGN OF MEMBRANE-PORE-FORMING PEPTIDES

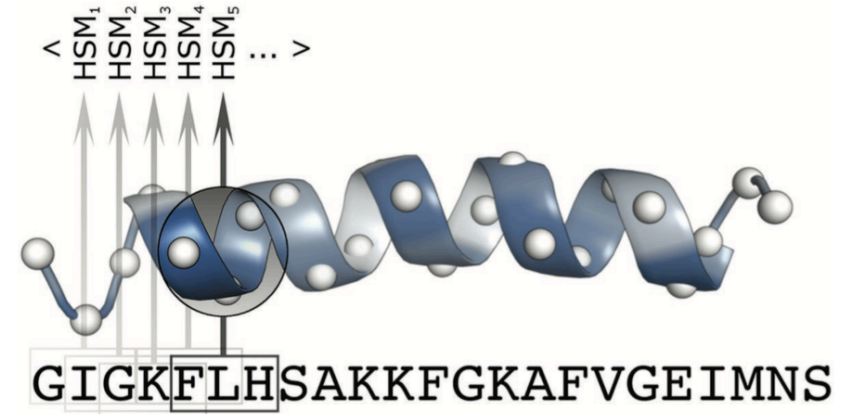
Eisenberg's mean hydrophobic moment

$$\langle \mu_H \rangle = \frac{1}{n} \sum_{i=1}^n \vec{H}_i$$



residue	Tanford	von Heijne-Blomberg	Janin	Chothia	Wolfenden	consensus
Ile	5.0	4.4	0.7	0.24	2.15	0.73
Phe	5.0	5.2	0.5	0.0	-0.76	0.61
Val	3.0	3.9	0.6	0.09	1.99	0.54
Leu	3.5	4.2	0.5	-0.12	2.28	0.53
Trp	6.5	3.9	0.3	-0.59	-5.88	0.37
Met	2.5	2.1	0.4	-0.24	-1.48	0.26
Ala	1.0	2.9	0.3	-0.29	1.94	0.25
Gly	0.0	1.9	0.3	-0.34	2.39	0.16
Cys	0.0	-0.08	0.9	0.0	-1.24	0.04
Tyr	4.5	3.6	-0.4	-1.02	-6.11	0.02
Pro	1.5	1.1	-0.3	-0.90	—	-0.07
Thr	0.5	1.2	-0.2	-0.71	-4.88	-0.18
Ser	-0.5	0.36	-0.1	-0.75	-5.06	-0.26
His	1.0	-1.5	-0.1	-0.94	-10.3	-0.40
Glu	—	-4.0	-0.7	-0.90	-10.2	-0.62
Asn	-1.5	-1.0	-0.5	-1.18	-9.68	-0.64
Gln	-1.0	-0.52	-0.7	-1.53	-9.38	-0.69
Asp	—	-5.6	-0.6	-1.02	-10.9	-0.72
Lys	—	-2.3	-1.8	-2.05	-9.52	-1.1
Arg	—	-9.4	-1.4	-2.71	-19.9	-1.8

Faraday Symp. Chem.Soc., 1982,17,109-120



$$\text{RMSD}_{V,W} = \sqrt{\frac{1}{n} \sum_{i=1}^n \lambda_i \left[ (v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2 \right]}$$

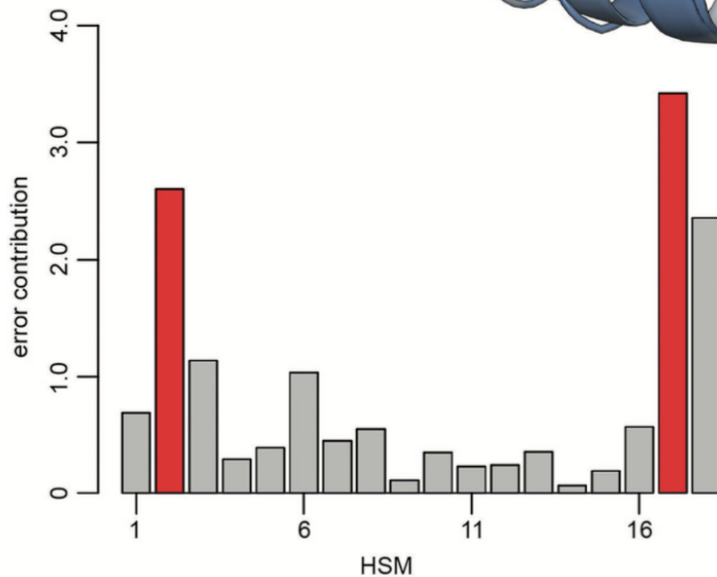
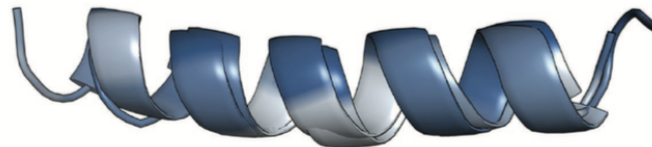
$$\lambda_i = 1 + \frac{1}{m} \sum_{j=1}^m (H_{j_v} - H_{j_w})^2$$

$v_i$  and  $w_i$  are the coordinates of the  $i$ -th HSM for two peptides  $V$  and  $W$

# RATIONAL DESIGN OF MEMBRANE-PORE-FORMING PEPTIDES

Pool of 50 000 sequences of 20 amino acid residues length  
 Amino acid distributions are based on the observed amino acid frequencies of all alpha-helical AMPs listed in the antimicrobial peptide database APD2

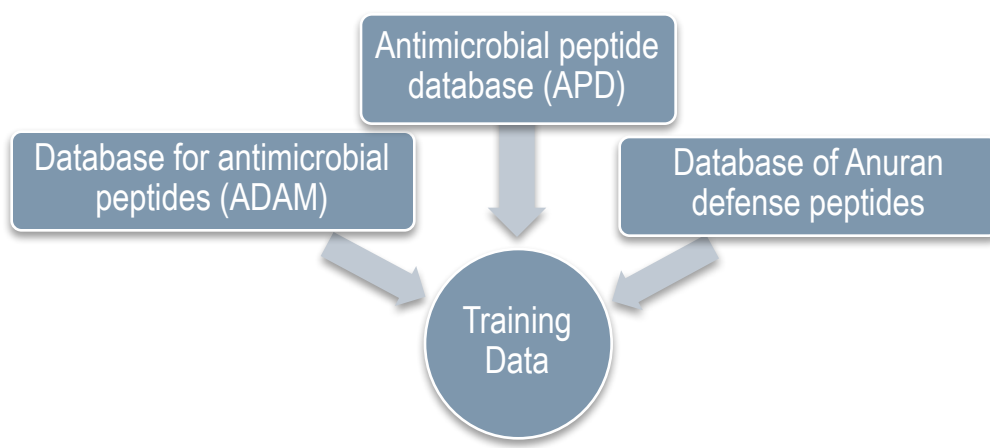
Rank	Score	ID	Sequence
1	1.027	Lavracin	WDPYFAGVKKLTKAILAVRA-NH <sub>2</sub>
2	1.030	P43745	LWLGDKKGAARLASALGLF-NH <sub>2</sub>
3	1.043	P14315	AELDGAKRGAEQKIGWLTVK-NH <sub>2</sub>
4	1.043	P15709	GFKHHQGSKKDDFPFLGKVD-NH <sub>2</sub>
5	1.047	P49224	IELKTGKLLDRVILFVALHA-NH <sub>2</sub>
6	1.048	P46981	ASKLGASLPLMKAKVGNVVGK-NH <sub>2</sub>
7	1.052	P12258	LFSTTDHVIGRKVLEIPAT-NH <sub>2</sub>
8	1.053	P27868	SFKSKQRKLSLPIMLLWGDA-NH <sub>2</sub>
9	1.059	P43971	EQIGQARIVAQNLAGHYLKS-NH <sub>2</sub>
10	1.068	P12245	IHWLESCRMHAHIIILIKTLG-NH <sub>2</sub>



Magainin-2 - Lavracin alignment

MAG **GIGKFL**HS AKKFGKAFV**GEIM**NS  
 ↓ ↓  
 LAV **-WDPYFAGVKKLTKAILAVRA--**

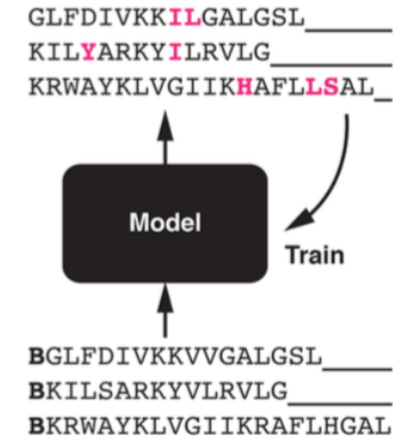
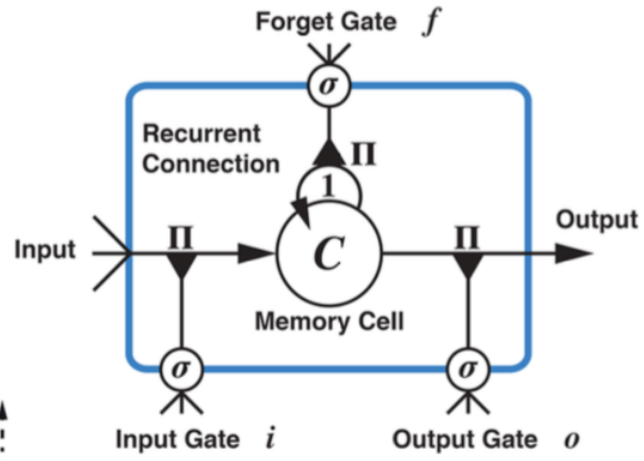
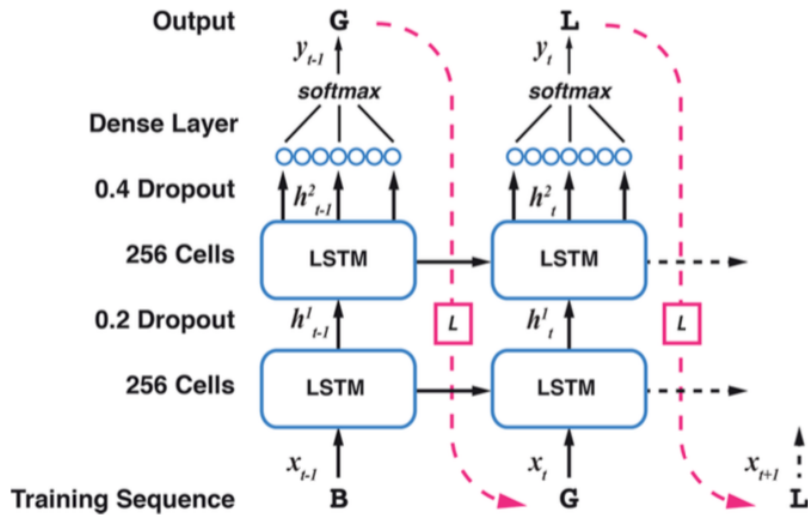
# RECURRENT NEURAL NETWORK MODEL FOR CONSTRUCTIVE PEPTIDE DESIGN



## Sequence Generation

The temperature-controlled probability  $P$  of picking amino acid  $y$  at sequence position  $i$

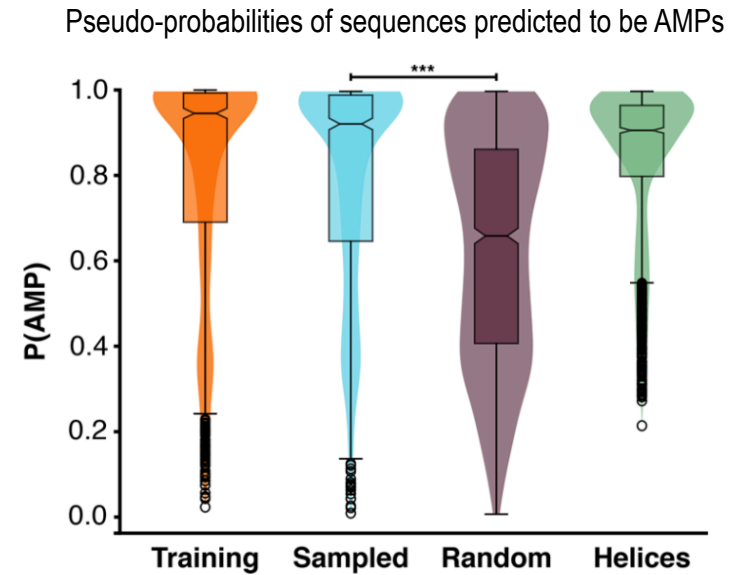
$$P(y_i) = \exp(y_i/T) / \left( \sum_{j=1}^n \exp(y_j/T) \right)$$





# RECURRENT NEURAL NETWORK MODEL FOR CONSTRUCTIVE PEPTIDE DESIGN

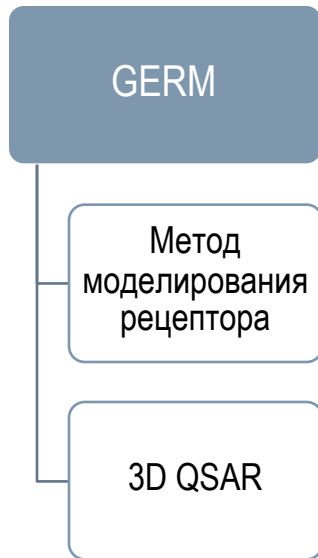
feature	training	generated
charge	$3.2 \pm 2.3$	$3.0 \pm 1.9$
length	$20.8 \pm 7.7$	$20.4 \pm 7.7$
molecular weight	$2226 \pm 809$	$2159 \pm 805$
charge density	$0.0014 \pm 0.0009$	$0.0014 \pm 0.0008$
Eisenberg hydrophobicity	$0.22 \pm 0.29$	$0.26 \pm 0.27$
Eisenberg hydrophobic moment	$0.36 \pm 0.14$	$0.38 \pm 0.13$
isoelectric point	$11.4 \pm 2.0$	$11.3 \pm 1.7$
aromaticity	$0.091 \pm 0.078$	$0.085 \pm 0.075$



МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO:  
модели псевдоресептора

---

# LIGAND-BASED MOLECULAR DESIGN USING PSEUDORECEPTORS: GENETICALLY EVOLVED RECEPTOR MODEL (GERM)



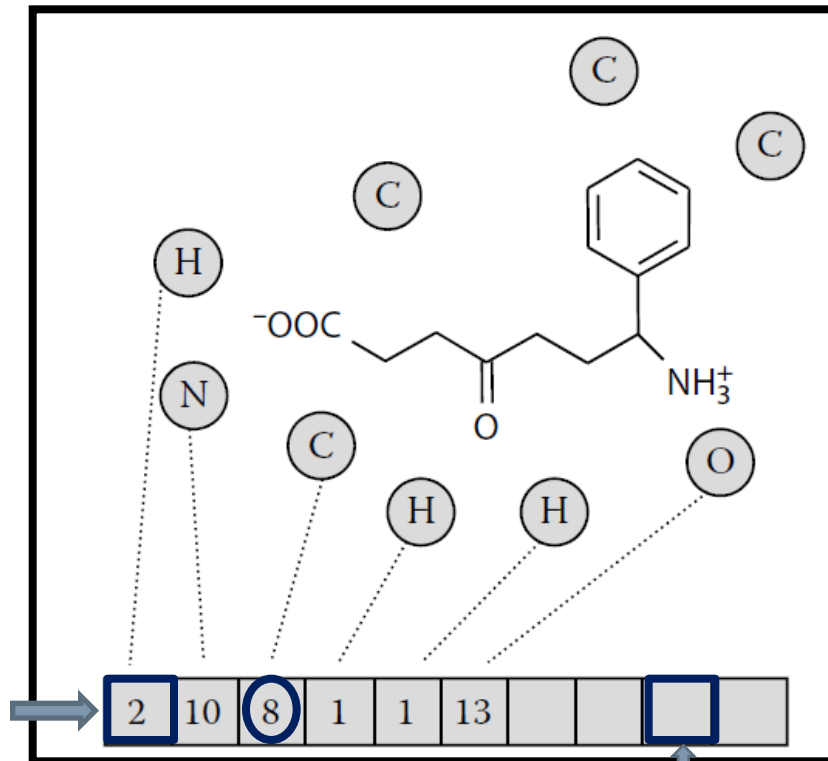
## Исходные требования:

SAR серия, для которой определены «приемлемые» выравнивания «подходящих» конформеров

## Предположения и ограничения:

- один лиганд – один конформер
- Значение биологической активности пропорционально энергии лиганд-рецепторного взаимодействия
- Влияние деформаций лиганда или гибкости рецептора не учитываются

# LIGAND-BASED MOLECULAR DESIGN USING PSEUDORECEPTORS: GENETICALLY EVOLVED RECEPTOR MODEL (GERM)



Атомы - «зонды» равномерно распределяются по поверхности сферы, окружающей лиганд, атом углерода помещается в каждую из точек, позиция корректируется для обеспечения наилучших ван-дер-Ваальсовых взаимодействий с лигандами

15 типов атомов

40 – 60 атомов- «зондов»

$$15 \cdot 60 \approx 4 \cdot 10^{70}$$



Генетический алгоритм

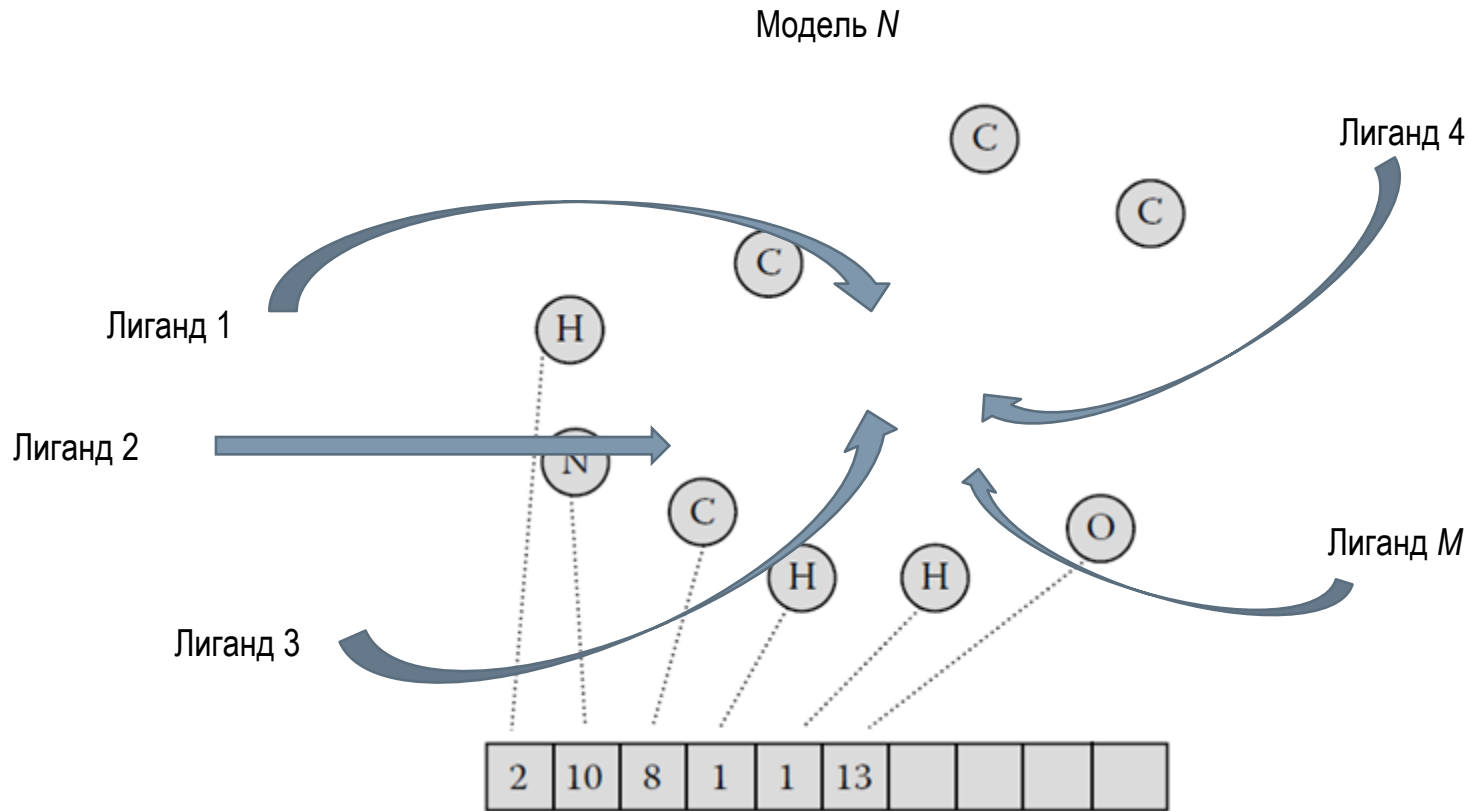
Положение в строке – положение в пространстве

Цифровое значение - тип атома-зонда

«Пустые» позиции ( возможность учета молекул растворителя)

# LIGAND-BASED MOLECULAR DESIGN USING PSEUDORECEPTORS: GENETICALLY EVOLVED RECEPTOR MODEL (GERM)

- ВДВ и электростатические энергии взаимодействия
- Корреляция между  $\log$  (значения свойства) и рассчитанной энергией взаимодействия

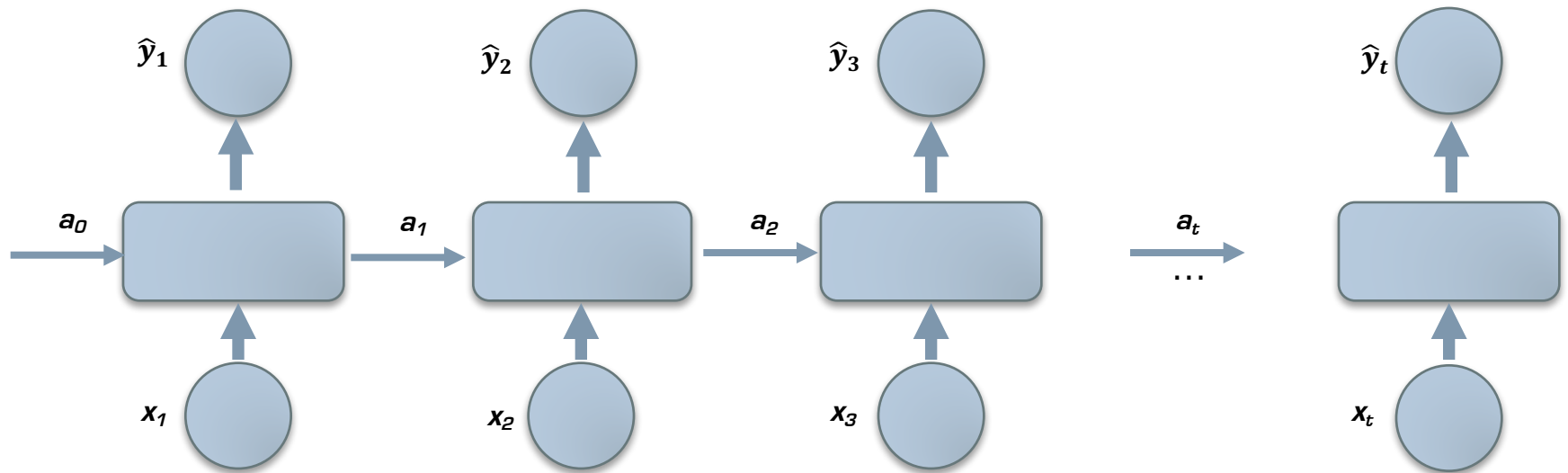


МОЛЕКУЛЯРНЫЙ ДИЗАЙН DE NOVO:  
ВОЗМОЖНОСТИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

---

# RECURRENT NEURAL NETWORKS

- Standard RNN – nonlinear system transforming sequences to sequences, sequences can vary in lengths
- RNN is parameterized with three weight matrices and three bias vectors

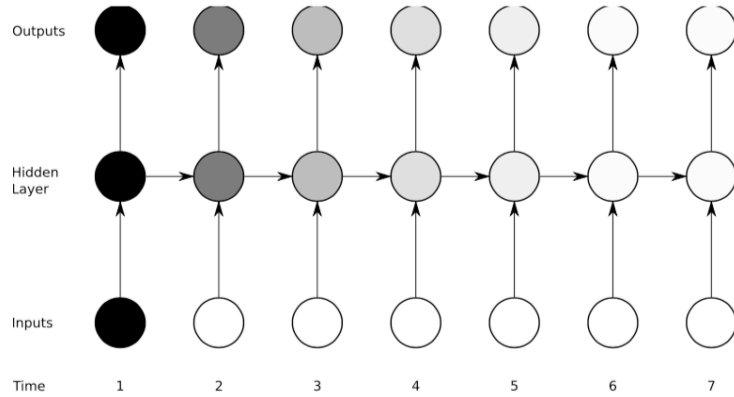


$$a^{(t)} = g(W_x x_t + U_x a_{t-1} + b_x)$$

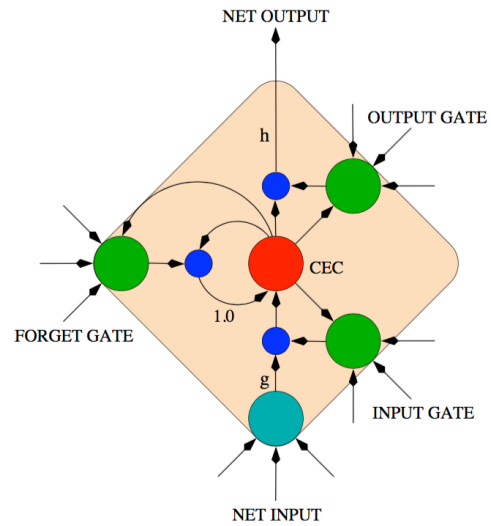
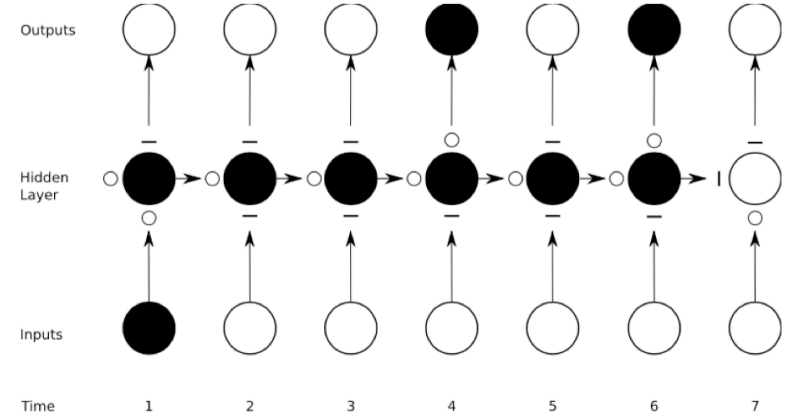
$$y^{(t)} = g(W_y a_t + b_y)$$

# LONG SHORT-TERM MEMORY NETWORKS

Рекуррентная нейронная сеть: проблема затухание градиентов



Сеть долговременной краткосрочной памяти (LSTM)





# BAYES THEOREM

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad \text{Теорема Байеса}$$

$P(c|d)$  - апостериорная вероятность принадлежности данному классу при данном значении признака

$P(c)$  – априорная вероятность данного класса

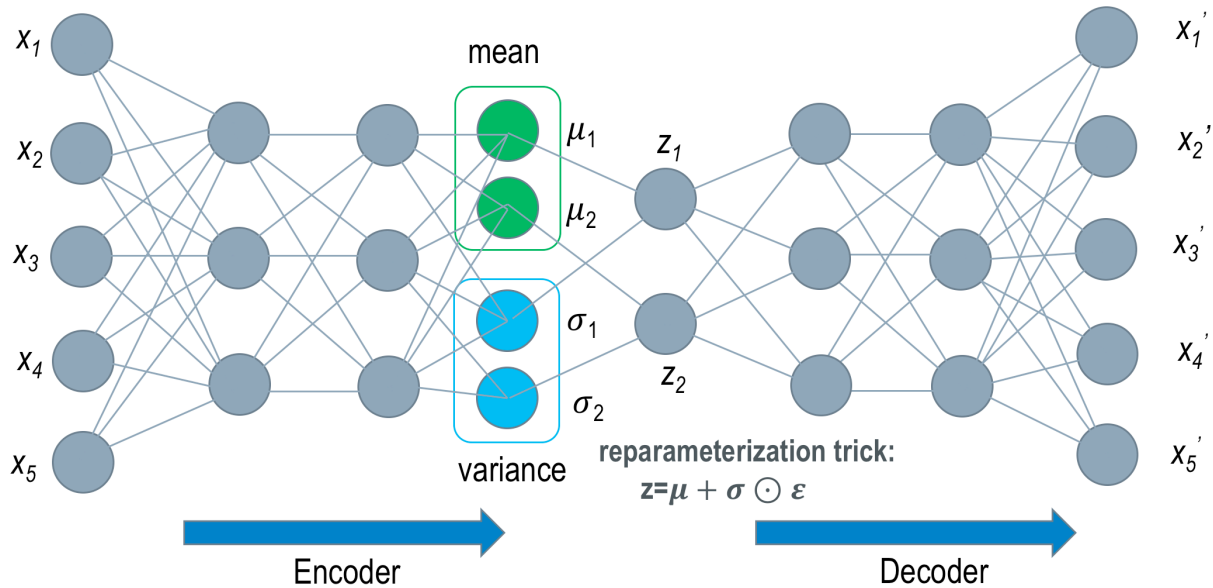
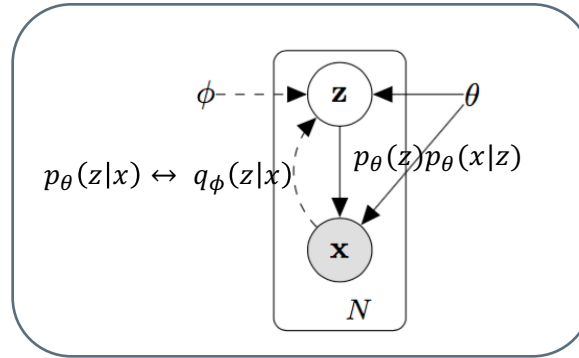
$P(d|c)$  – правдоподобие (вероятность данного значения признака при данном классе)

$P(d)$  – априорная вероятность данного значения признака

---

# VARIATIONAL AUTOENCODERS

Вариационный автокодировщик (Variational Autoencoder) – метод вариационного вывода моделей скрытых (латентных) переменных



# VARIATIONAL AUTOENCODERS

Задан некоторый набор данных  $X = \{X^{(i)}\}_{i=1}^N$

Введем пространство скрытых переменных  $Z = \mathbb{R}^d$

Предполагается, что данные порождены (сгенерированы) некоторым случайным процессом, включающим два шага:

- Из многомерного нормального априорного распределения данных  $P_{\theta^*}(z)$  генерируется скрытая переменная  $z \in Z$
- Объект  $x^{(i)}$  генерируется из условного параметрического распределения на каждый признак  $P_{\theta^*}(x|z)$

Введём вспомогательное параметрическое распределение  $q_{\phi}(z|x)$ , являющееся аппроксимацией апостериорного  $P_{\theta}(z|x)$

$q_{\phi}(z|x) \Rightarrow$  кодировщик

$P_{\theta}(x|z)$  – декодировщик

## Вариационный вывод

$$\log P_{\theta}(x_1, \dots, x_N) = \sum_{i=1}^N \log P_{\theta}(x^{(i)})$$

$$\log P_{\theta}(x^{(i)}) = \underbrace{D_{KL}(q_{\phi}(z|x^{(i)}) || P_{\theta}(z|x^{(i)}))}_{\text{Расхождение между истинным апостериорным и вспомогательным параметрическим распределениями}} + \underbrace{\alpha(\theta, \phi; x^{(i)})}_{\text{Вариационная нижняя оценка на логарифм максимального правдоподобия}}$$

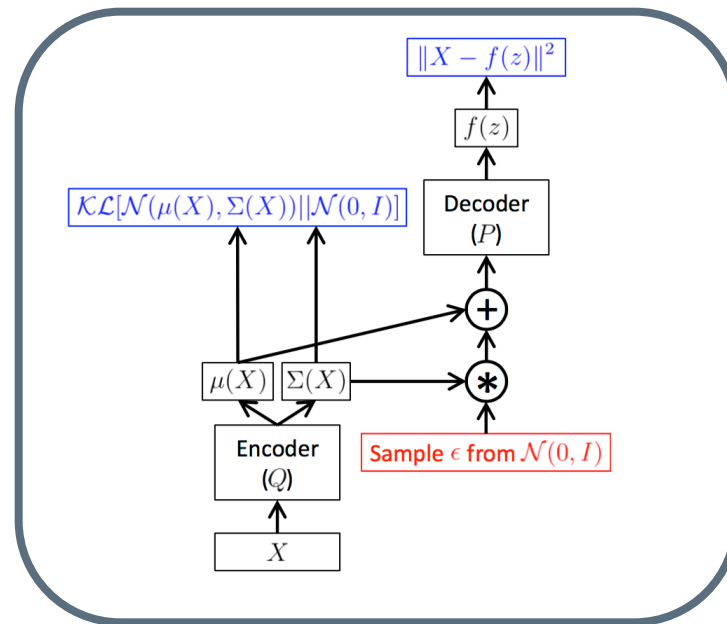
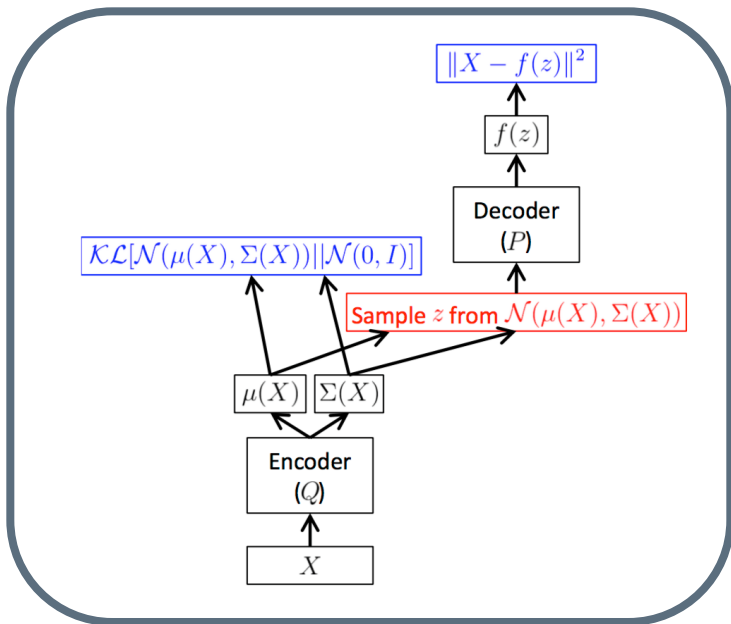
Расхождение между истинным апостериорным и вспомогательным параметрическим распределениями

Вариационная нижняя оценка на логарифм максимального правдоподобия

$$\alpha(\theta, \phi, x^{(i)}) = -D_{KL}(q_{\phi}(z|x^{(i)}) || P_{\theta}(z|x^{(i)})) + E_{q_{\phi}(z|x_i)}[\log P_{\theta}(x^{(i)}|z)]$$

# VARIATIONAL AUTOENCODERS

## Вариационный вывод



## Трюк репараметризации

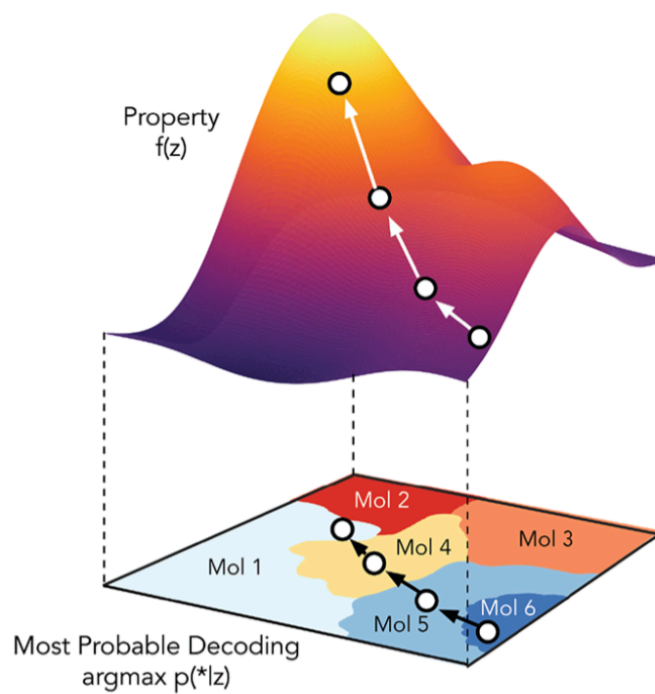
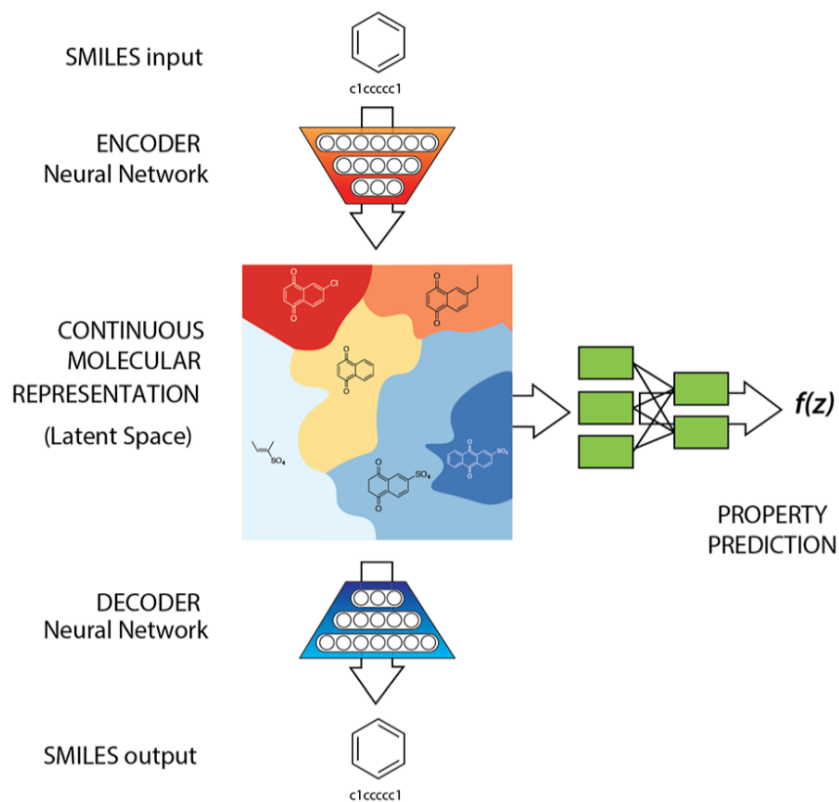
Латентная переменная  $Z$  может быть репараметризована при помощи случайной переменной шума (замена генерации точек из параметрического распределения на генерацию точек из распределения без настраиваемых параметров):

$$\tilde{z} = g_{\phi}(\epsilon, x)$$

Таким образом, латентная переменная является входными данными функции:

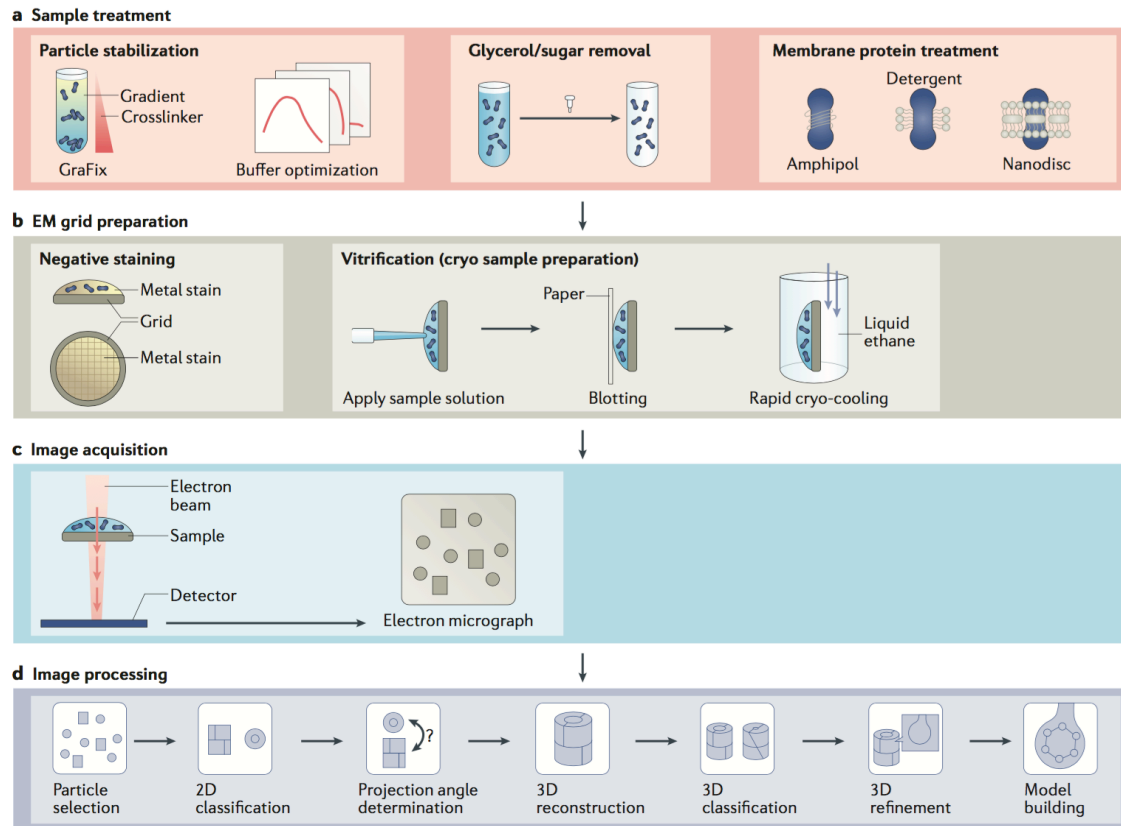
$$\log P_{\theta}(x^{(i)} | z^{i,l})$$

# VARIATIONAL AUTOENCODERS



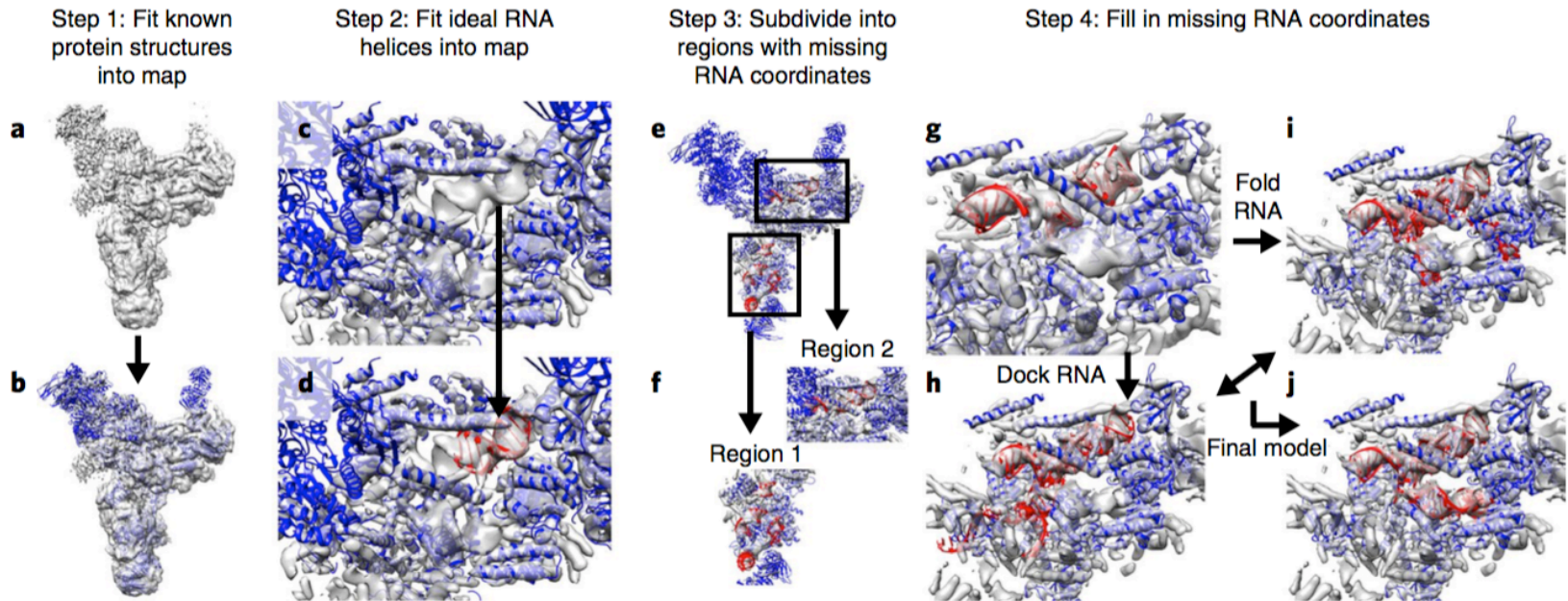
# DE NOVO COMPUTATIONAL MODELING OF RNA-PROTEIN COMPLEXES WITH ROSETTA SOFTWARE

## Cryo-EM workflow



- Most high-resolution maps contain regions of low resolution in which manual tracing of coordinates is not possible
- Missed atomic coordinates can be obtained by fitting of known structures of smaller components into the density, however due to difficulty of this procedure RNA coordinates are often omitted from models of RNP complexes
- Developing tools capable of modeling RNA into density maps to reproduce the complex is needed
- DRRAFTER automatically builds missing RNA coordinates into cryo-EM maps by means of fragment-based docking and folding

# DE NOVO COMPUTATIONAL MODELING OF RNA-PROTEIN COMPLEXES WITH ROSETTA SOFTWARE



## DRRAFTER include two stepwise stages:

### I. Low-resolution stage:

Monte-Carlo-based optimization – RNA fragment insertion, docking procedure is used to optimize the placement of RNA helices and proteins, the proteins are treated as rigid bodies.

Missing segments are built iteratively; starting from residues immediately N- or C-terminal to a missing segment, putative solutions are spawned that add one additional residue and sample the conformation of the most recently placed three residues guided by backbone segments from high-resolution structures with similar local sequence.

### II. Full atom refinement

Energy minimization both for protein and RNA followed by structure refinement [single residue fragment insertion, small rigid body perturbations, side chain packing] and second stage of energy minimization. Scoring is performed alternatively using energy function including terms that describe hydrogen bonding, electrostatics torsion angles, vDW interactions, solvation etc