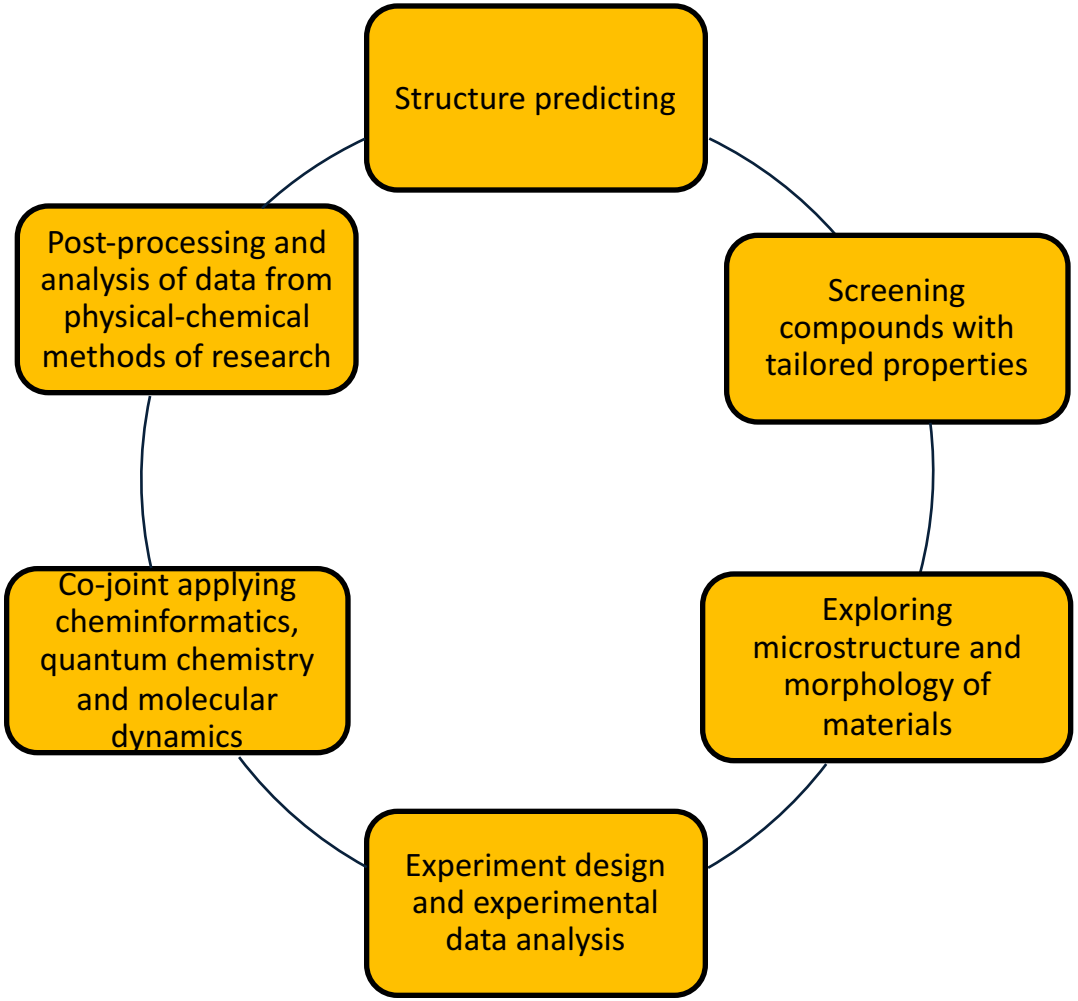
Введение в химическую информатику

Лекции 2-3

Data search
Data extraction
Data analysis
Data filtering

ХИМИЧЕСКАЯ ИНФОРМАТИКА: НЕКОТОРЫЕ ВОЗМОЖНОСТИ ПРИМЕНЕНИЯ МЕТОДОВ



ДОПОЛНИТЕЛЬНАЯ РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

Книги:

Virtual Screening. Principles, Challenges and Practical Guidelines. Edited by **C.Sotriffer** Wiley-VCH (2011), 398 c

Informatics for Materials Science and Engineering. Edited by **K. Rajan.** Butterworth-Heinemann (2013), 542 c

The Practice of Medicinal Chemistry 4th Edition. Edited by **David Aldous Pierre Raboisson and Didier Rognan,** Academic Press (2015), 902 c

---

D. Fourches et al **Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research**
Journal of Chemical Information and Modeling (2010) *50* (7), 1189-1204 DOI: 10.1021/ci100176x

A. Cherkasov et al **QSAR Modeling: Where Have You Been? Where Are You Going To?**
Journal of Medicinal Chemistry (2014) *57* (12), 4977-5010 DOI: 10.1021/jm4004285

A Varnek et al **Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis?**
Journal of Chemical Information and Modeling (2012) *52* (6), 1413-1437 DOI: 10.1021/ci200409x

I. Baskin et al **Artificial intelligence in synthetic chemistry: achievements and prospects**
Russ. Chem. Rev. (2017) 86 1127

S. Kalinin et al **Big–deep–smart data in imaging for guiding materials design**
Nature Mater (2015) 14 973

C. Zheng et al **Automated generation and ensemble-learned matching of X-ray absorption spectra**
npj Computational Materials (2018) 12 1

Korvigo et al **Putting hands to rest: e cient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules**
J Cheminform (2018) 10:28

ХИМИЧЕСКАЯ ИНФОРМАТИКА: ВЗАИМОСВЯЗЬ СТРУКТУРА – СВОЙСТВО

Сбор и подготовка данных

Сбор и проверка экспериментальных данных (поиск в базах данных или извлечение информации из различных источников, предварительный анализ корректности и полноты данных, одинаковые экспериментальные условия получения или введение дополнительных параметров, удаление нежелательных соединений, нормализация специфических хемотипов, таутомеров, восполнение пропущенных данных)

АВТОМАТИЧЕСКИЕ ИЗВЛЕЧЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

Information Retrieval and Text Mining Technologies for Chemistry
Martin Krallinger et al *Chemical Reviews* 2017 *117* (12), 7673-7761 DOI: 10.1021/acs.chemrev.6b00851

ОПТИЧЕСКОЕ РАСПОЗНАВАНИЕ (OPTICAL COMPOUND RECOGNITION)

Извлечение структурной информации из рисунков и изображений химических структур, представленных в виде цифрового растрового изображения.

Основные этапы:

1. Перевод изображения в оттенки серого и бинаризация, классификация каждого пиксела

2. Сегментация изображения для разделения текста и графики, сканирование и определение сегментов как связанных компонентов, удаление шума

3. Распознавание атомов и связей основано на алгоритмах поиска сочетания линий и символов: перевод в векторное представление, распознавание обозначения атомов и зарядов с учетом фиксированного максимального значения высоты и ширины символа, существования двух символов с горизонтальным или вертикальным выравниванием и символов «+» «-», распознается тип связи.

Направления дальнейшего развития:

- **Обработка нечитаемых символов**
- **Работа с таблицами**
- **Работа с большими макромолекулярными структурами и сложными циклами**
- **Распознавание химических таблиц или реакций**

Information Retrieval and Text Mining Technologies for Chemistry
Martin Krallinger et al *Chemical Reviews* 2017 *117* (12), 7673-7761 DOI: 10.1021/acs.chemrev.6b00851

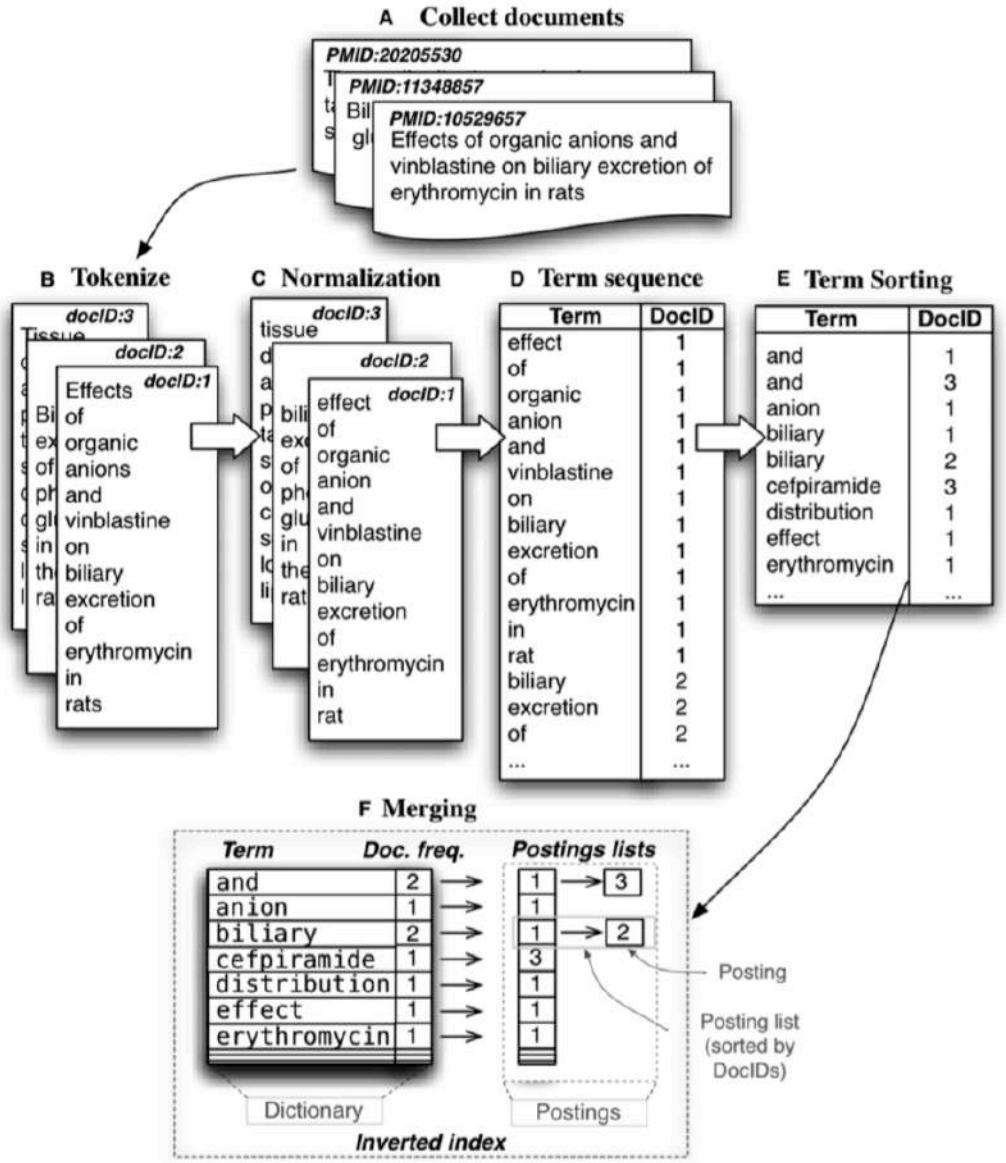
БАЗЫ ДАННЫХ: АВТОМАТИЧЕСКОЕ ПОПОЛНЕНИЕ

software	approach	input	outputs	availability	URL
AsteriX web server	reconstruction of 3D ligand coordinates from 2D images	PDF		free for academic use	http://swift.cmbi.ru.nl/bitmapb/
ChemEx	extracts compound, organism, and assay information; uses OSRA	reads full-text papers, and recovers SMILES and MOL from structure images	visualization via graphical user interface and exportation to XML	free	http://www.biotec.or.th/isl/ChemEx
ChemInfty	mathematical OCR (includes two engines), supporting Markush recognition	image formats	SDF, MOL	commercial	http://www.inftyproject.org/en/ChemInfty
chemOCR (chemical optical character recognition)	expert rules and supervised learning methods	BMP, GIF, PNG, and multipage TIF	SMILES, SDF	commercial	http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/products/chemocr.html
ChemReader	machine vision approach	image formats	ML, SMILES	commercial	http://www-personal.umich.edu/kazu/research-areas.html
CLiDE (chemical literature data extraction)	image processing and artificial intelligence	documents of the following types: PDF, DOC(X), and HTML; images files: BMP, GIF, JPEG, JPG, JPE, JIF, PBM, PGM, PNG, PNM, PPM, TIFF, TIF, XBM, and XPM	various export options (e.g., can be directly transferred into chemical editors), depends on the version	commercial	http://www.keymodule.co.uk/products/clide/index.html
D2S	supports CLiDE, OSRA, and Imago	documents in PDF, TXT, HTML, XML, and MS Office formats (e.g., DOC, DOCX, PPT, PPTX, XLS, XLSX), OpenOffice ODT, embedded structure objects (e.g., ChemDraw, SymyxDraw, MarvinSketch), and images in TIFF and BMP formats	MRV (Marvin documents), ML, SMILES, MOL	several commercial and free licenses	https://www.chemaxon.com/products/document-to-structure/
IBM OROCS (optical recognition of chemical graphics)	image processing, ability to recognize images containing structure diagrams in documents	image formats	MOL	commercial	not available
IMAGO OCR	image processing and lexicon-based abbreviation expansion	PNG, JPEG, BMP, DIB, TIFF, PBM, RAS	MOL	free (GPL-licensed but possible to purchase a commercial license)	http://lifescience.opensource.epam.com/imago/
Kekulé	image processing and rule-based logic, with manual marking of structure diagram regions	ISIS, MOLfile, ROSDAL, and Kekulé's native format	ISIS, MOL, SMILES, ROSDAL, and Kekulé's native format	no longer commercially available	http://aig.cs.man.ac.uk/research/kekule/
MLOCSR	combines a low-level processor with Markov logic (to reason about the low-level entities and relations); images with tables and/or reactions are not supported	image file (jpg, png, jpeg, and TIF supported)	MOL	system is available as a web server at http://mlocsr.dinfo.unifi.it; standalone distribution is currently under development	http://mlocsr.dinfo.unifi.it/
OSRA	image processing, character/string recognition, various connection table compilation, and confidence estimation	image file (gif, jpg, png, jpeg, tif, pdf, and ps supported)	SMILES, SD files	open source	https://cactus.nci.nih.gov/osra/

Information Retrieval and Text Mining Technologies for Chemistry
Martin Krallinger et al *Chemical Reviews* 2017 *117* (12), 7673-7761 DOI: 10.1021/acs.chemrev.6b00851

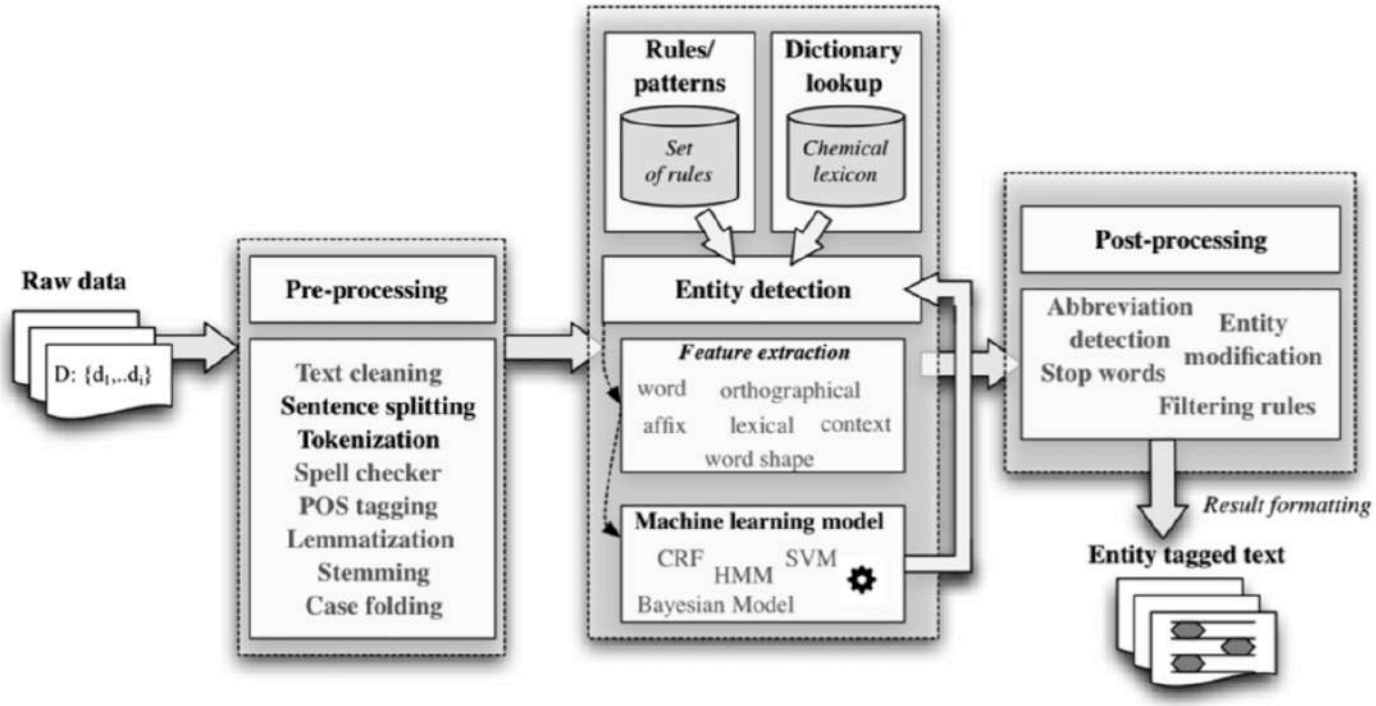
АВТОМАТИЧЕСКИЕ ИЗВЛЕЧЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

- Преобразование документа (программами оптического распознавания)
- Проверка типа кодировки
- Сегментация документа на подразделы и поиск необходимой информации внутри выделенных секций/параграфов
- Фрагментация на отдельные предложения (в частности, при помощи методов машинного обучения)
- Разделение на отдельные слова/термины – токенизация
- Индексация документа (выделение ключевых слов) и определение его приоритета при поиске по ключевым словам



Information Retrieval and Text Mining Technologies for Chemistry
Martin Krallinger et al *Chemical Reviews* 2017 *117* (12), 7673-7761 DOI: 10.1021/acs.chemrev.6b00851

АВТОМАТИЧЕСКИЕ ИЗВЛЕЧЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ



Information Retrieval and Text Mining Technologies for Chemistry
Martin Krallinger et al *Chemical Reviews* 2017 *117* (12), 7673-7761 DOI: 10.1021/acs.chemrev.6b00851

АВТОМАТИЧЕСКИЕ ИЗВЛЕЧЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

chemical NER/indexer	description	URL
BANNER-CHEMDNER	CRF-based systematic chemical tagger	https://bitbucket.org/tsendeemts/banner-chemdner
BC4-CHEMDNER Uni. Wuhan CER	CRF-based systematic chemical tagger	https://github.com/zuiwufenghua/biocreative_CHEMDNER
becas-chemicals	online CRF-based chemical/drug tagger	http://bioinformatics.ua.pt/becas-chemicals/
ChemEx	entity tagger integrating ChemicalTagger	http://www3a.biotec.or.th/isl/ChemEx/
chemicalize	commercial CER system	https://chemicalize.com
ChemicalTagger	chemical NLP tool	http://chemicaltagger.ch.cam.ac.uk
ChemSpot	hybrid (CRF and dictionary) chemical tagger	https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/chemspot/chemspot
chemxseer-tagger	CRF chemical tagger	https://github.com/SeerLabs/chemxseer-tagger
CheNER	CRF-based systematic chemical tagger	http://ubio.bioinfo.cnio.es/biotools/CheNER/
Cocoa	hybrid (manual rule/dictionary) chemical tagger	http://relagent.com/Tech.html
iice	online chemical entity and relation tagger	www.lasige.di.fc.ul.pt/webtools/iice/
LeadMine	hybrid (manual rule/dictionary) chemical tagger	https://www.nextmovesoftware.com/leadmine.html
MetaMap	tagger for UMLS metathesaurus concepts	https://metamap.nlm.nih.gov
NCBO Annotator (ChEBI ontology)	online tagger of OBO ontologies (incl. ChEBI)	http://bioportal.bioontology.org/annotator
OntoGene	online chemical tagger (incl. lexical lookup)	http://www.ontogene.org/webservices/
Oscar3	naïve Bayesian model-based CER tagger	http://www-pmr.ch.cam.ac.uk/wiki/Oscar3
Oscar4	modular adaptation and update of Oscar3	https://bitbucket.org/wwmm/oscar4
tmChem	CER and normalization using CRF	https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#tmChem
Whatizit	online tagger or entities (incl. chemicals)	http://www.ebi.ac.uk/webservices/whatizit/info.jsf

Information Retrieval and Text Mining Technologies for Chemistry
Martin Krallinger et al *Chemical Reviews* 2017 *117* (12), 7673-7761 DOI: 10.1021/acs.chemrev.6b00851

DEEP CNN-RNN ARCHITECTURE FOR CHEMICAL NAMED ENTITY RECOGNITION

CHEMDNER corpus

Ten thousand abstracts from eleven chemistry-related fields of science with over 84k manually annotated chemical entities (20k unique) of eight types:

- ABBREVIATION (15.55%)
- FAMILY (14.15%)
- FORMULA (14.26%)
- IDENTIFIER (2.16%)
- MULTIPLE (0.70%)
- SYSTEMATIC (22.69%)
- TRIVIAL (30.36%)
- NO CLASS (0.13%)

Neural networks

Three types of neural networks:

one-dimensional (1D) convolutional neural networks (CNN), recurrent neural networks (RNN) and

Detecting stitch points

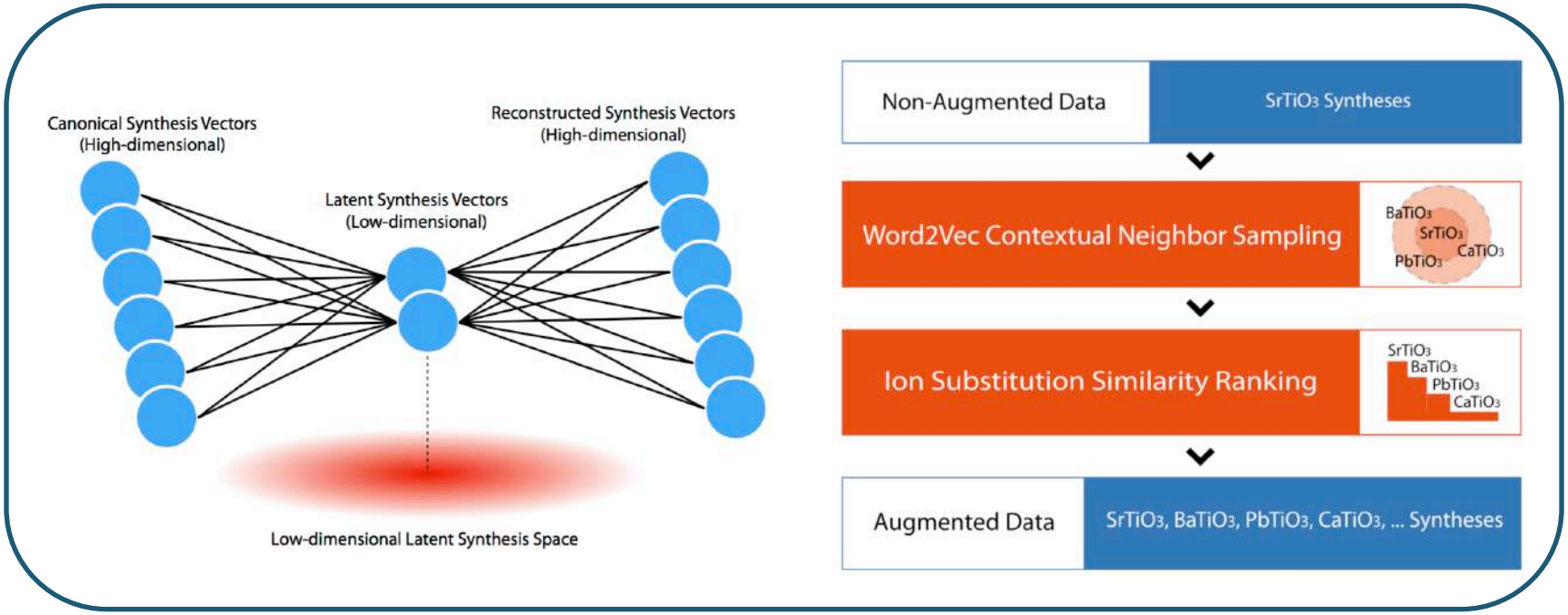
time-distributed dense (fully-connected) networks (TDD).

Classifier with binary output to recover target tokens

Korvigo et al **Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules** ***J Cheminform*** (2018) 10:28

VIRTUAL SCREENING OF INORGANIC MATERIALS SYNTHESIS PARAMETERS WITH DEEP LEARNING

Extracted synthesis data
(sintering and calcination temperature and time, method of synthesis, solvent)

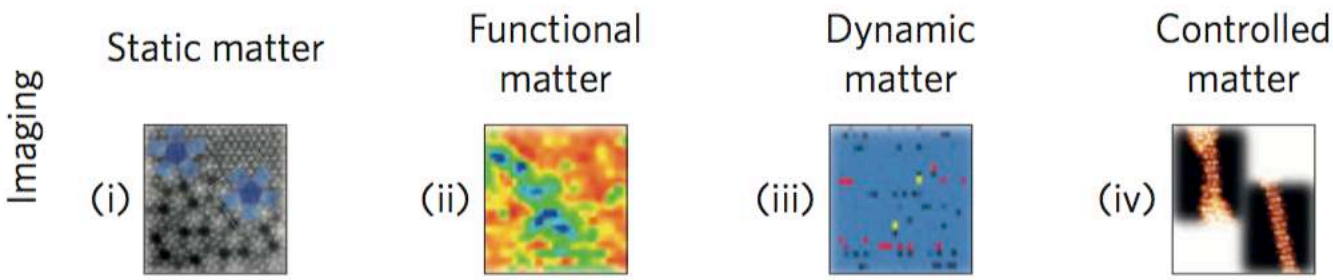


- **Defects engineering strategy**
- **Searching for common hidden materials behavior**
- **Structure predicting or elucidating**
- **Synthesis optimization, experiment design**

Kim et al Virtual screening of inorganic materials synthesis parameters with deep learning
npj Computational materials (2017)3:53

BIG–DEEP–SMART DATA IN IMAGING FOR GUIDING MATERIALS DESIGN

Различные виды сканирующей зондовой микроскопии:



Анализ локальной структуры материала
Анализ поведения материала в динамике
Анализ электронной и атомной структур

S. Kalinin et al **Big–deep–smart data in imaging for guiding materials design**
Nature Mater (2015) 14 973

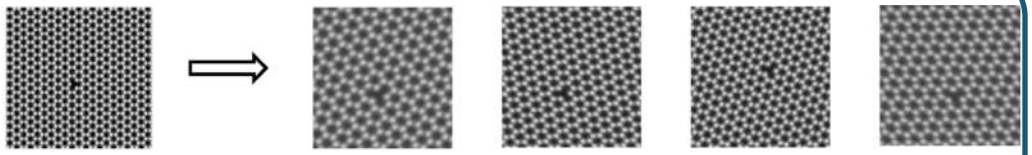
АНАЛИЗ ЛОКАЛЬНОЙ СТРУКТУРЫ МАТЕРИАЛА: ИСПОЛЬЗОВАНИЕ МЕТОДОВ ГЛУБОКОГО ОБУЧЕНИЯ

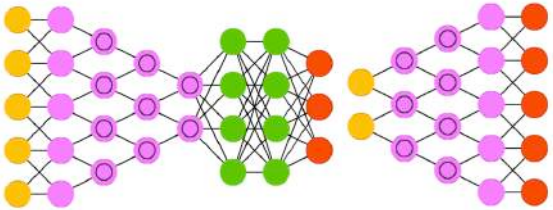
Возможности:

- Анализ кристаллических решеток и структуры и типа дефектов
- Уточнение типа структуры
- Извлечение значимых структурных/химических параметров

Обучающий набор данных (искусственные данные, мимикрирующие данные STEM):

- Идеальная кристаллическая решетка
- Решетка с присутствием вакансии
- Атом-допант с большим атомным



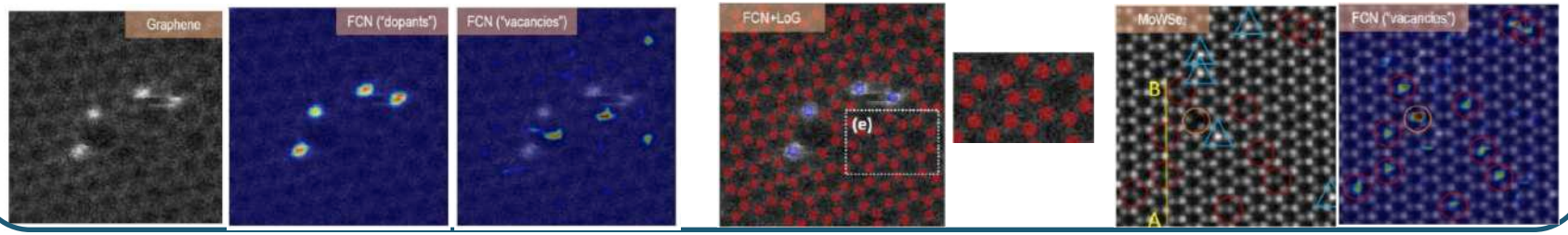


- Finding atomic positions from raw experimental data
- Pixel-wise recognition of atomic defects and lattice reconstructions
- Possible chemical structure-related interpretation of output

Identifying “not-seen” defects:

Stone-Wales defects in graphene

Discriminating vacancy

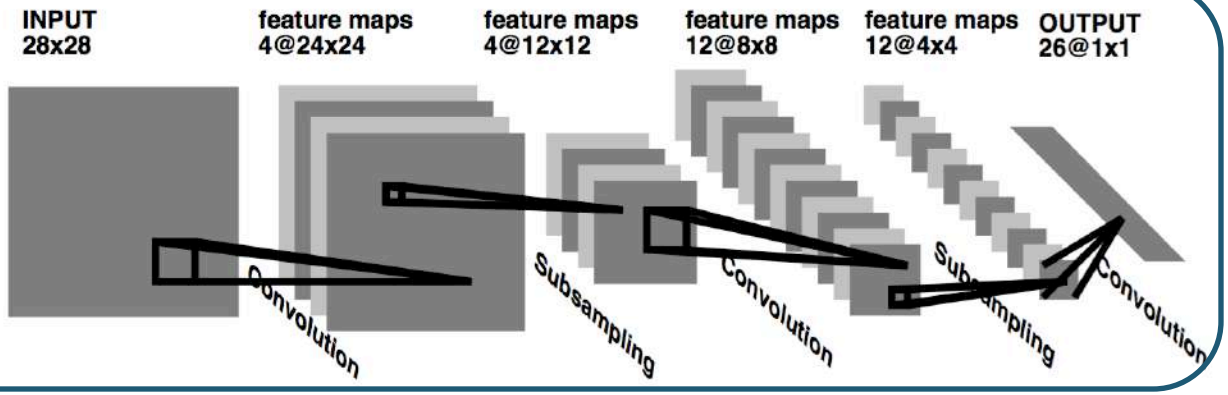


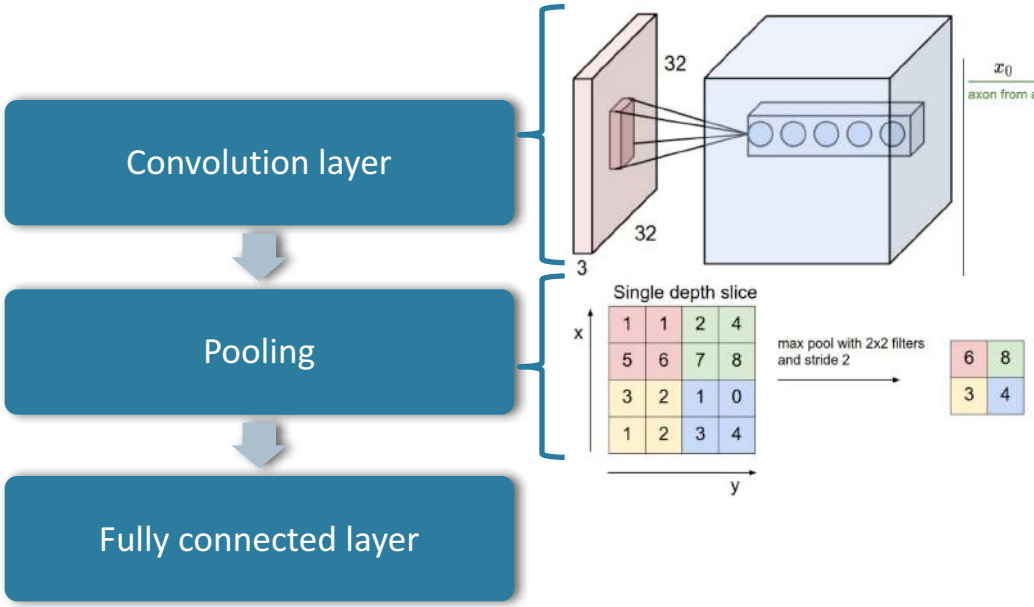
Maxim Ziatdinov et al **Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations** *ACS Nano* **2017** *11* (12), 12742-12752

CONVOLUTIONAL NEURAL NETWORKS

Specific characteristics:

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





Параметры метода:

- Число фильтров
- Размер фильтров
- Шаг
- Чередование слоев

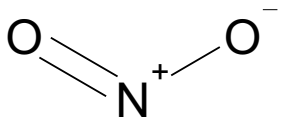
РАБОТА С ХИМИЧЕСКИМИ БАЗАМИ ДАННЫХ

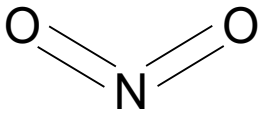
Виды поиска в химических базах данных

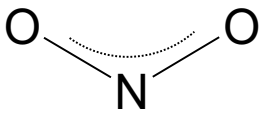
Поиск идентичной химической структуры
Подструктурный поиск
Надструктурный поиск
Поиск по молекулярному подобию
Поиск по заданному фармакофору
Поиск по заданным критериям, ключевым словам

ПОИСК ИДЕНТИЧНОЙ ХИМИЧЕСКОЙ СТРУКТУРЫ

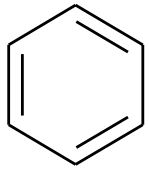
Альтернативное представление функциональных групп



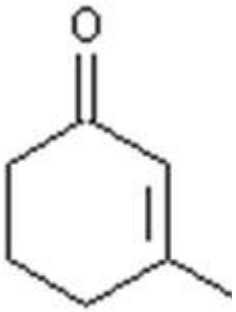


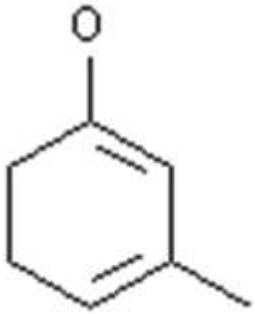


Ph

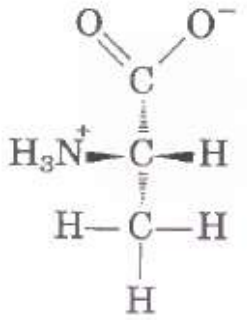


Таутомерия

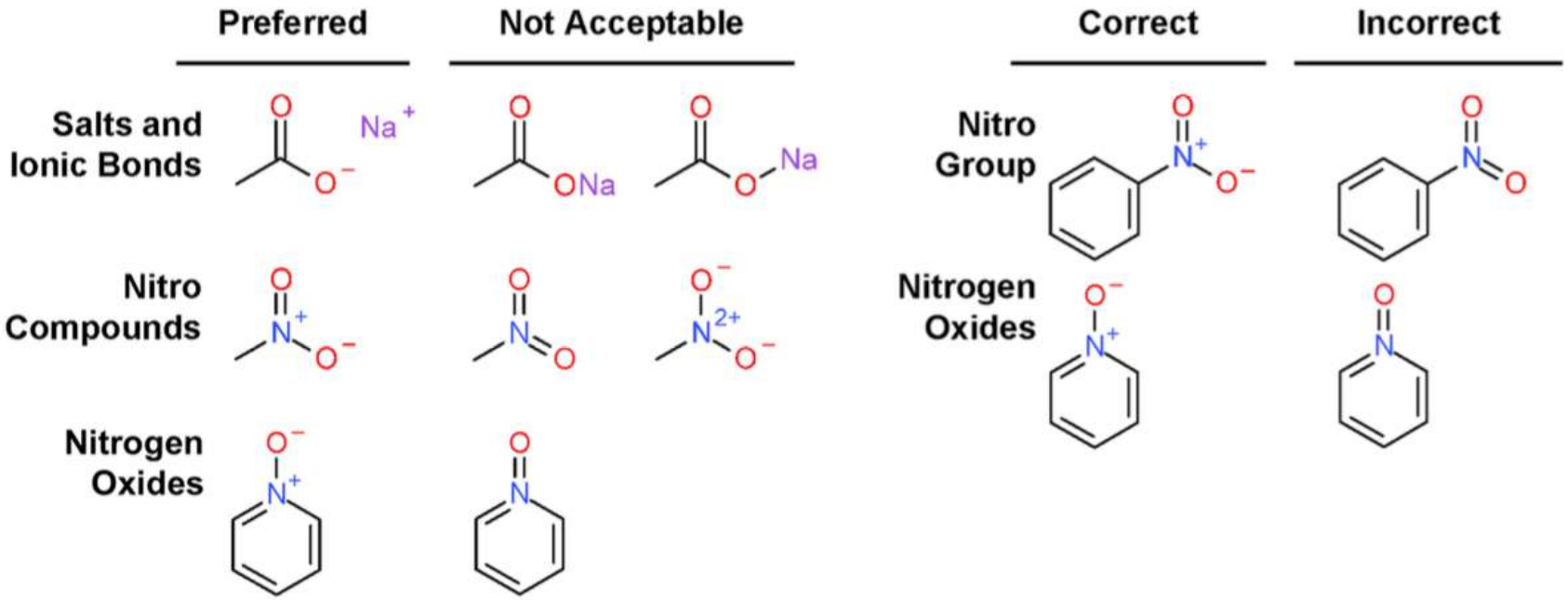




Стереоизомерия



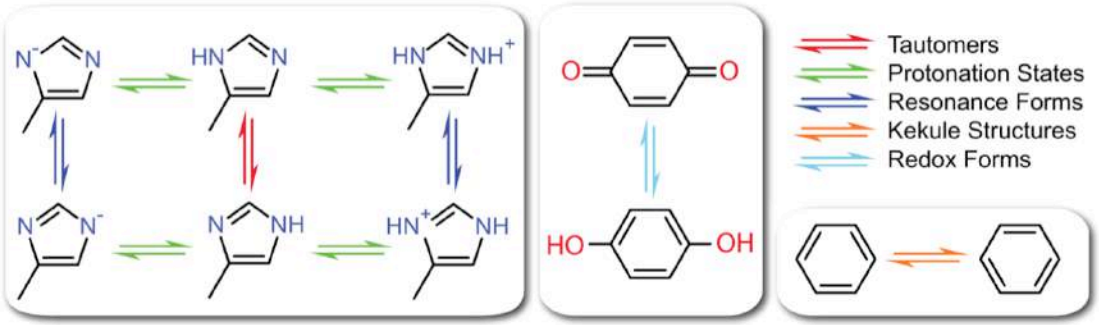
ПОИСК ИДЕНТИЧНОЙ ХИМИЧЕСКОЙ СТРУКТУРЫ: АЛЬТЕРНАТИВНОЕ ПРЕДСТАВЛЕНИЕ ХИМИЧЕСКИХ ГРУПП

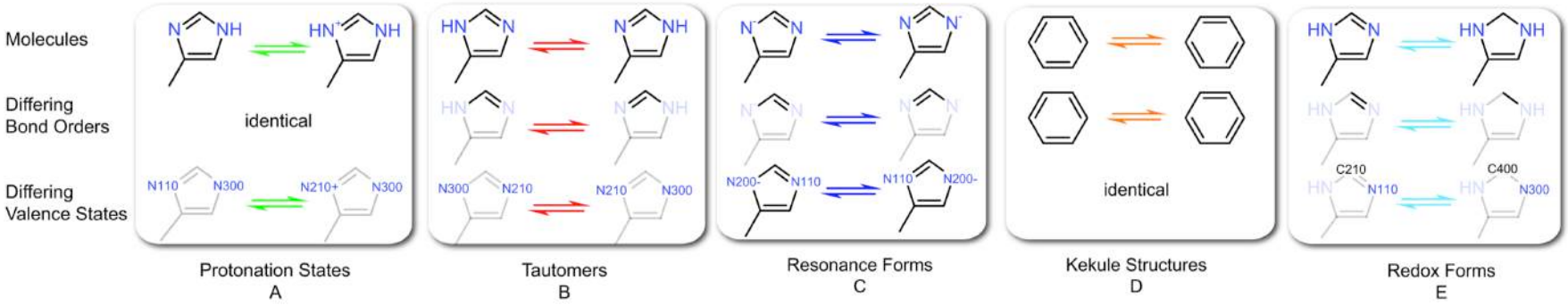


Hahnke et al PubChem chemical structure standardization J Cheminform (2018) 10:36

ПОИСК ИДЕНТИЧНОЙ ХИМИЧЕСКОЙ СТРУКТУРЫ: ТАУТОМЕРИЯ

Different valence bond structures





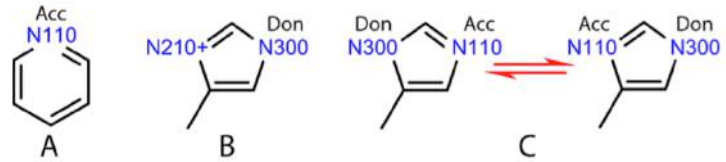
relation	substitution type	condition
kekule	none	
ionization	protonation	$\Delta(D \to A) \neq \Delta(A \to D)$
protonation	protonation	$\Delta(D \to A) = \Delta(A \to D)$
mesomer	resonance	$\Delta(D \to A) = \Delta(A \to D)$
tautomer	tautomer	$\Delta(D \to A) = \Delta(A \to D)$
redox	resonance	$\Delta(D \to A) \neq \Delta(A \to D)$
	tautomer	$\Delta(D \to A) \neq \Delta(A \to D)$

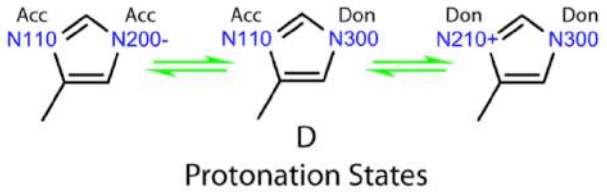
Urbaczek et al The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States JCIM (2014), 54, 756−766

https://docs.chemaxon.com/display/docs/Tautomerization+and+tautomers

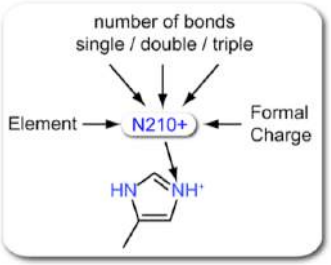
ПОИСК ИДЕНТИЧНОЙ ХИМИЧЕСКОЙ СТРУКТУРЫ: ТАУТОМЕРИЯ

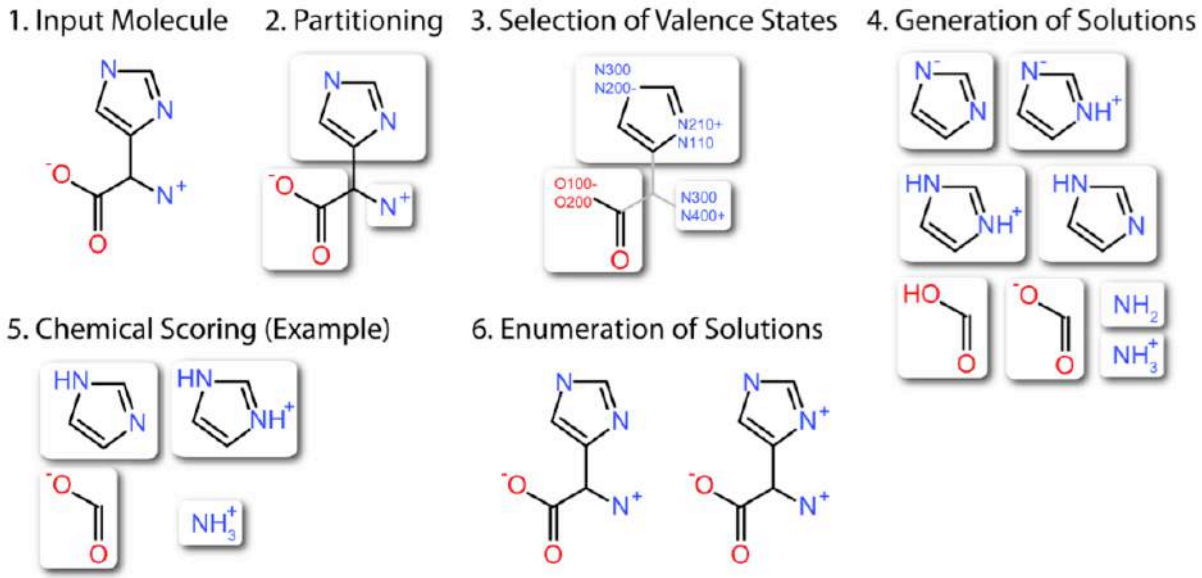
Criteria for the generation of additional states



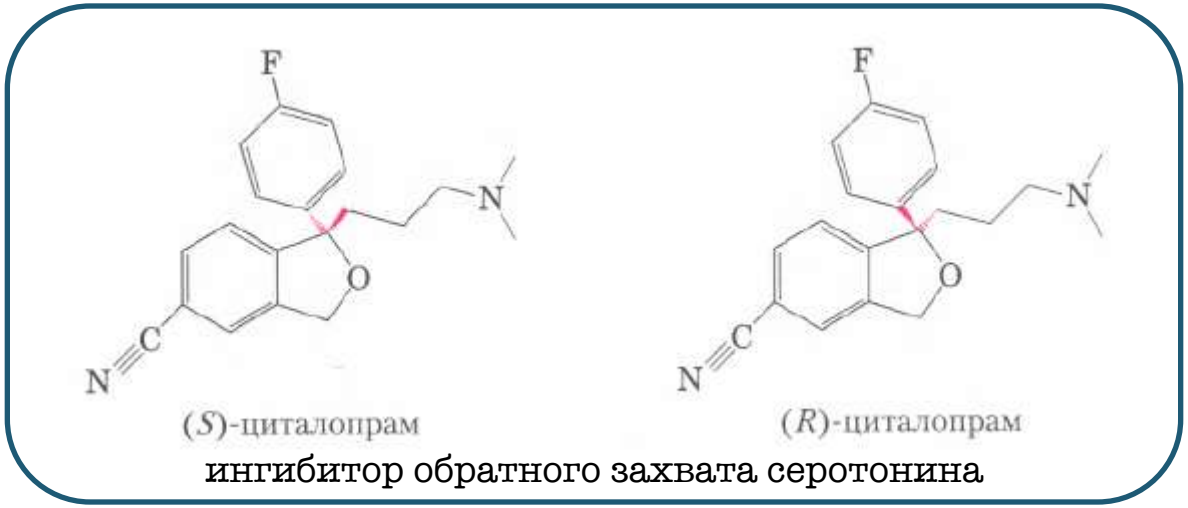


The generation of tautomers requires at least one tautomer acceptor and one tautomer donor in a zone.





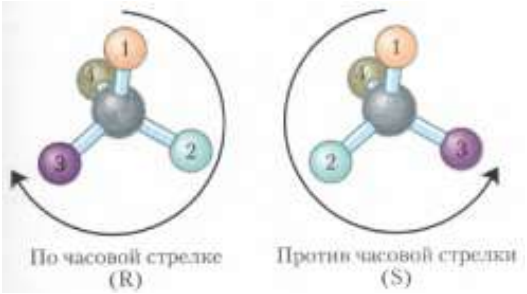
Оптическая изомерия в драг-дизайне



ингибитор обратного захвата серотонина

Целекса™ – рацемическая смесь двух стереоизомеров, из которых только (S)-циталопрам оказывает терапевтическое воздействие;
Чистый (S)-циталопрам - Лексапро™

R/S обозначения абсолютной конфигурации определяются в соответствии с правилом Кана-Ингольда-Прелога, регулирующих старшинство заместителей у хиральных атомов



R – *rectus* (лат) – правый
S – *sinister* (лат) – левый

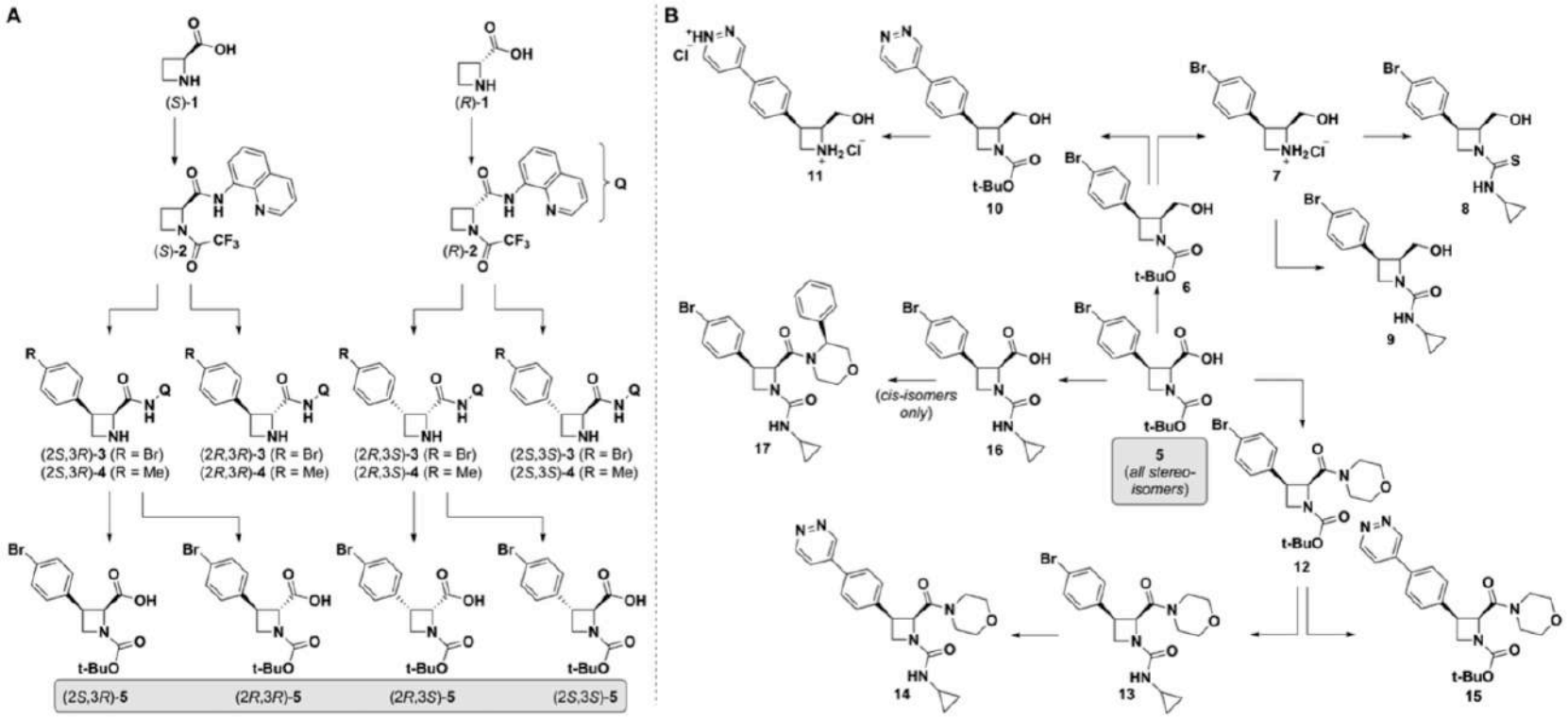
Приоритеты атомов: -OCH2 > -OH > -NH2 > -COOH > -CHO > -CH2OH > -CH3 > -H

Д. Нельсон, М. Кокс "Основы биохимии Ленинджера" Бином 2012

STEREOCHEMISTRY IMPACT ON THE PERFORMANCE DIVERSITY OF A COLLECTION OF SYNTHETIC COMPOUNDS

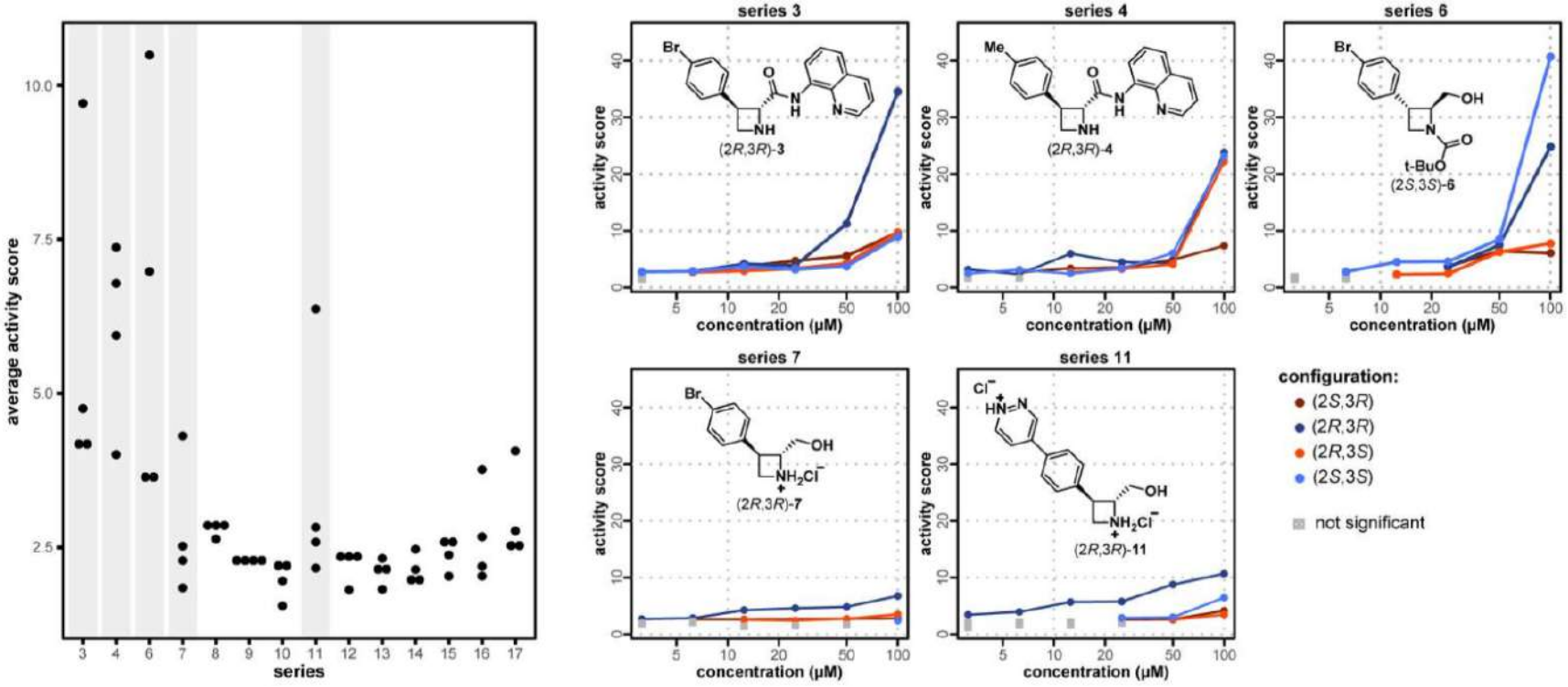
Stereospecific C–H arylation

Appendage diversification



Melillo et al Synergistic effects of stereochemistry and appendages on the performance diversity of a collection of synthetic compounds JACS (2018) Just accepted manuscript

STEREOCHEMISTRY IMPACT ON THE PERFORMANCE DIVERSITY OF A COLLECTION OF SYNTHETIC COMPOUNDS



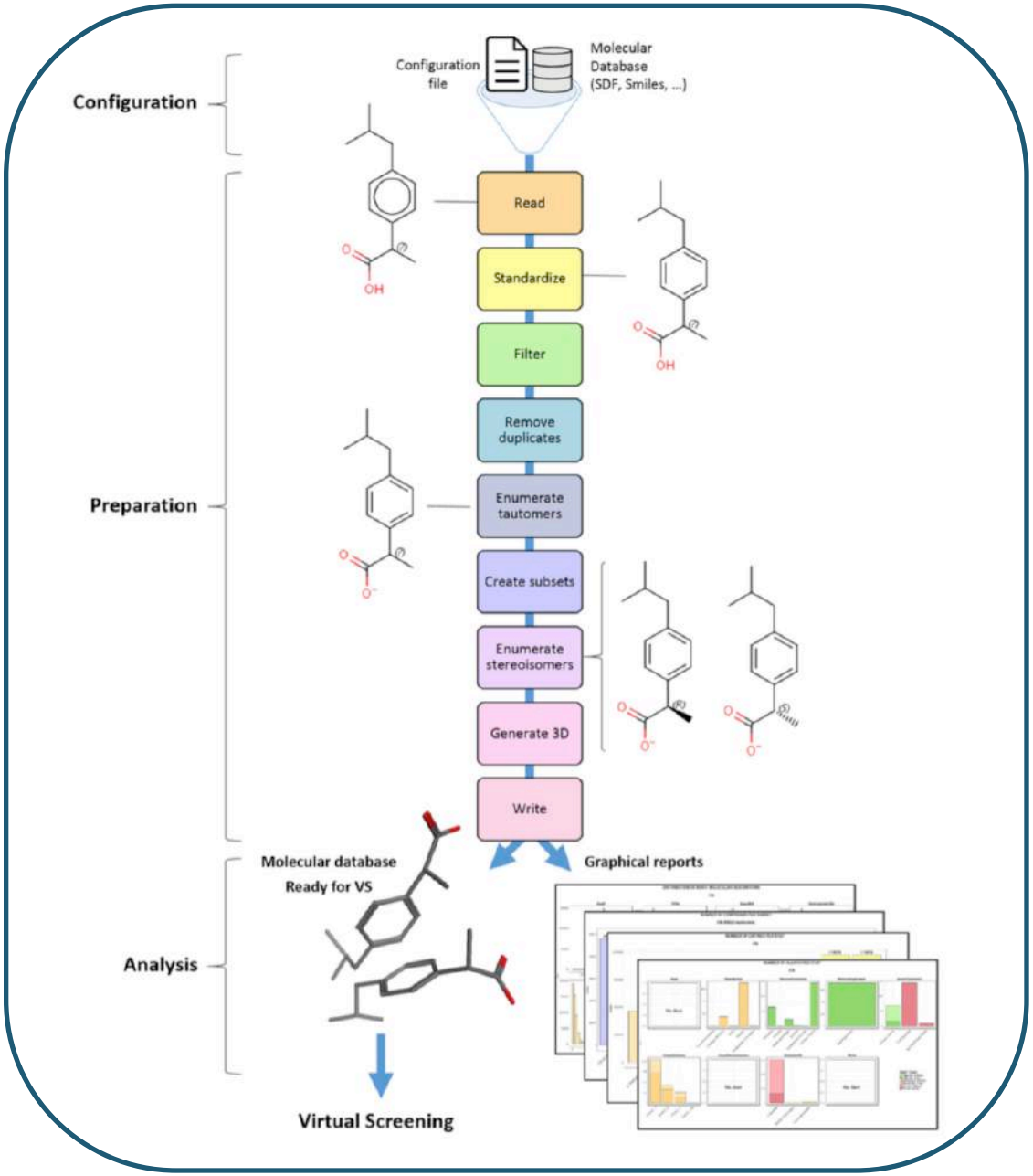
Melillo et al Synergistic effects of stereochemistry and appendages on the performance diversity of a collection of synthetic compounds JACS (2018) Just accepted manuscript

ПОДГОТОВКА ДАННЫХ: ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

procedures	software
inorganics removal	ChemAxon/Standardizer OpenEye/Filter
structure normalization (fragment removal, structural curation, salt neutralization)	ChemAxon/Standardizer OpenBabel Molecular Networks/CHECK,TAUTOMER
duplicate removal	ISIDA/Duplicates HiT QSAR CCG/MOE
SDF management/viewer file format converter	ISIDA/EdiSDF Hyleos/ChemFileBrowser OpenBabel ChemAxon/MarwinView CambridgeSoft/ChemOffice Schrödinger/Canvas ACD/ChemFolder Symyx/Cheminformatics CCG/MOE Accelrys/Accord Tripos/Benchware Pantheon

Fourches et al Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research *JCIM* 2010 *50* (7), 1189-1204

ПОДГОТОВКА ДАННЫХ



Reading:
removing entries with missing or ambiguous structures, removing salts, mixtures processing

Standartization:
bonds aromaticity normalization, determining canonical tautomer form, unique representation of chemical groups

Filtering:
removing charged carbonic atoms, isotopes and inorganic atoms

Removing duplicate structures: InChiKey as criteria

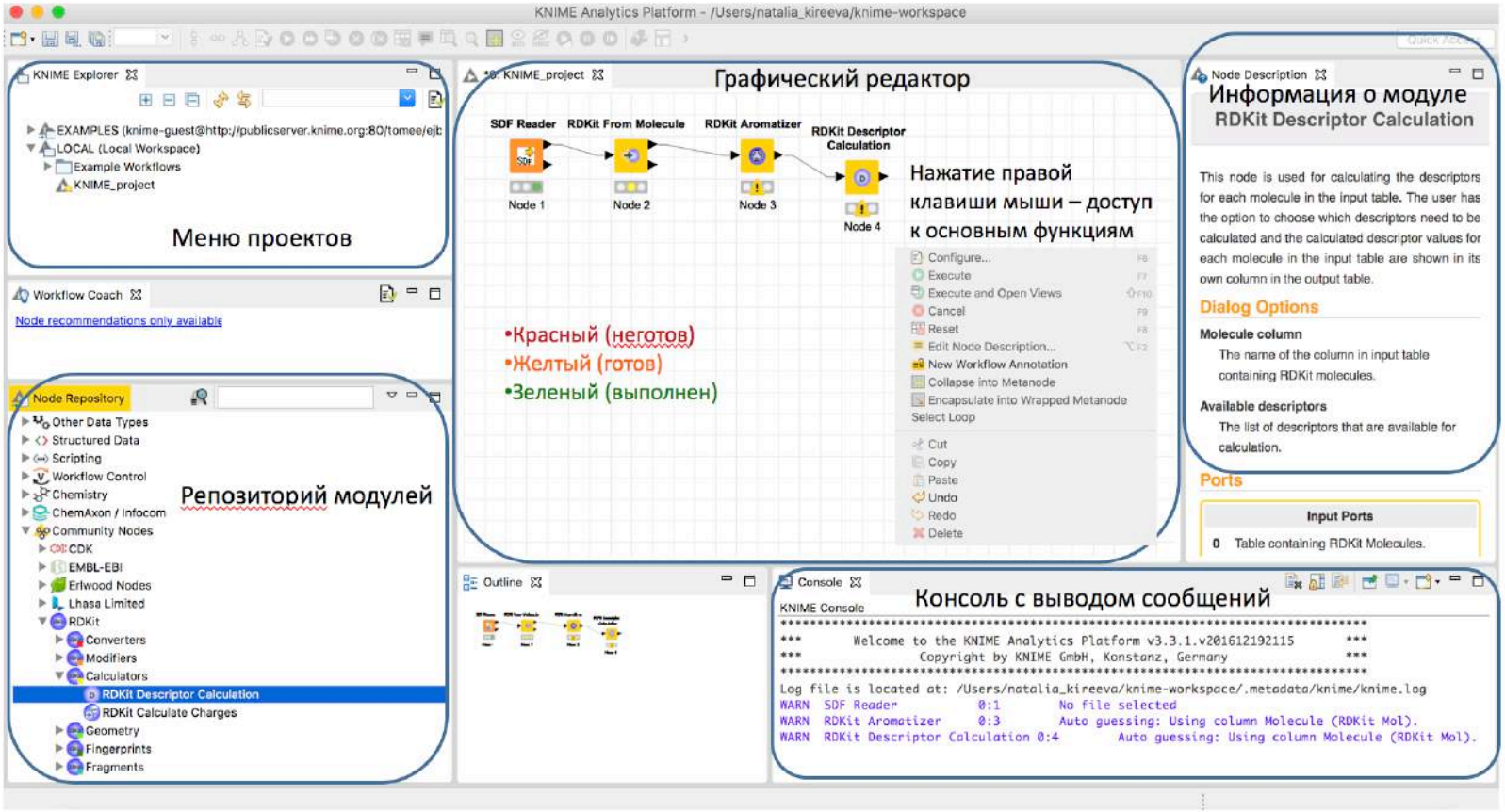
Tautomers enumerating

Stereoisomers enumerating

Gally et al VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening Mol Inf (2017) 36, 1700023

ПОДГОТОВКА ДАННЫХ: ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

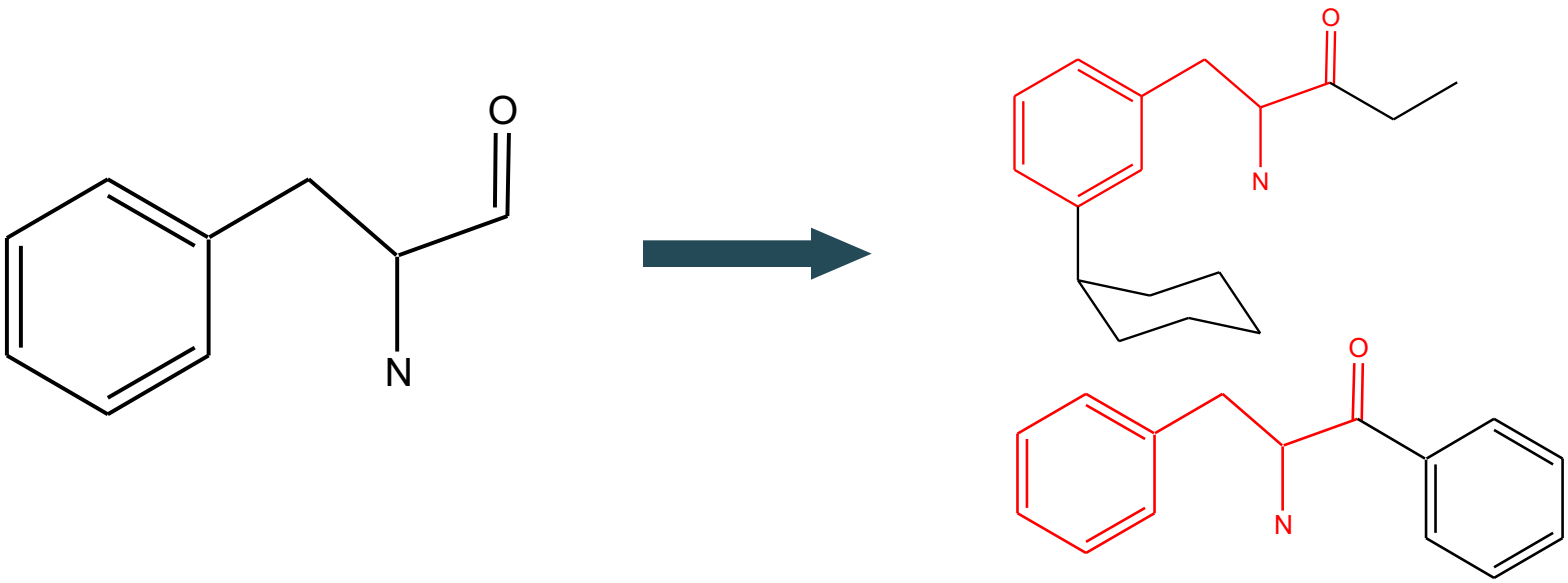
Open for Innovation ®
KNIME



https://www.knime.com

ПОДСТРУКТУРНЫЙ ПОИСК (SUBSTRUCTURAL SEARCH)

Поиск соединений, содержащих данную молекулу как подструктуру



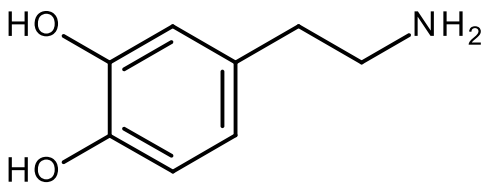
NB: на рис. представлена скелетная структурная формула, где не отображаются связи углерод—водород, а также атомы водорода и углерода. На атом углерода указывает излом цепи или ее окончание, если к нему не присоединена какая-либо не углеводородная группа.

Реализация подструктурного поиска:

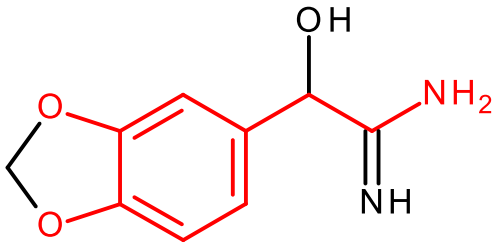
- Поиск изоморфного подграфа (рекурсивный обход с отсечением, смыкающееся дерево)
- Подструктурный поиск с использованием молекулярных отпечатков пальцев

ПОДСТРУКТУРНЫЙ ПОИСК С ИСПОЛЬЗОВАНИЕМ МОЛЕКУЛЯРНЫХ ОТПЕЧАТКОВ

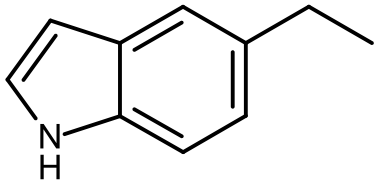
запрос

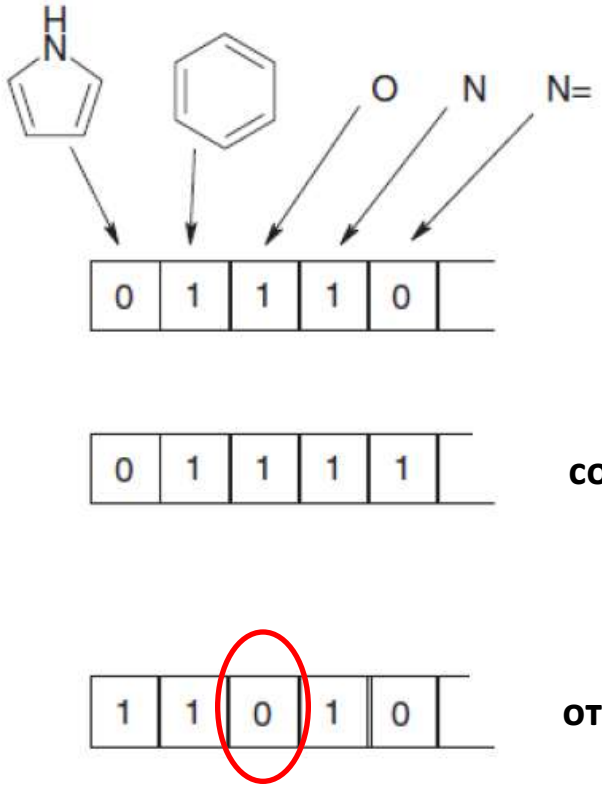


Соединение 1



Соединение 2



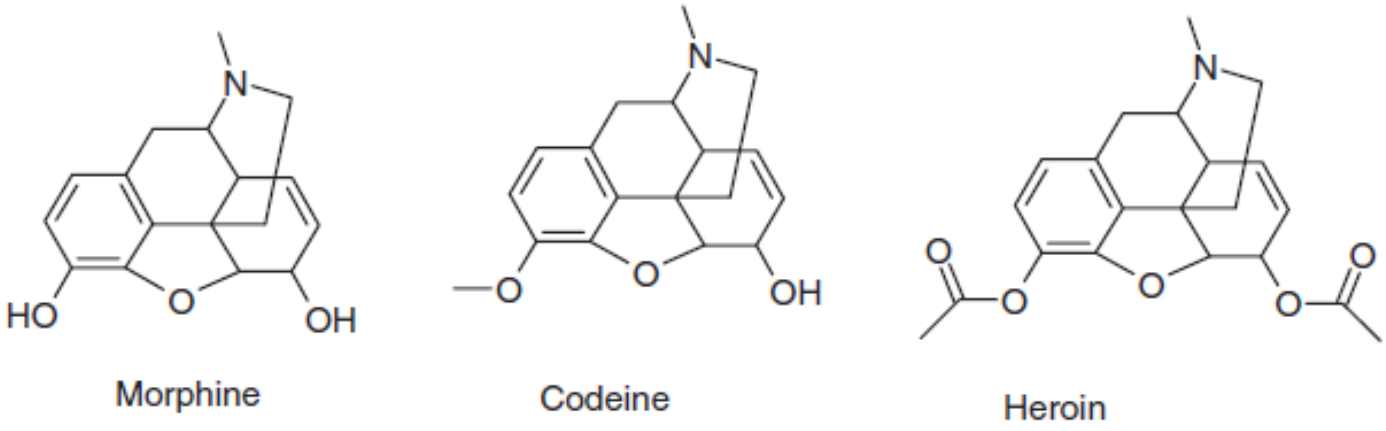


Слайд из курса лекций профессора А. Варнека

ПОИСК ПО МОЛЕКУЛЯРНОМУ ПОДОБИЮ

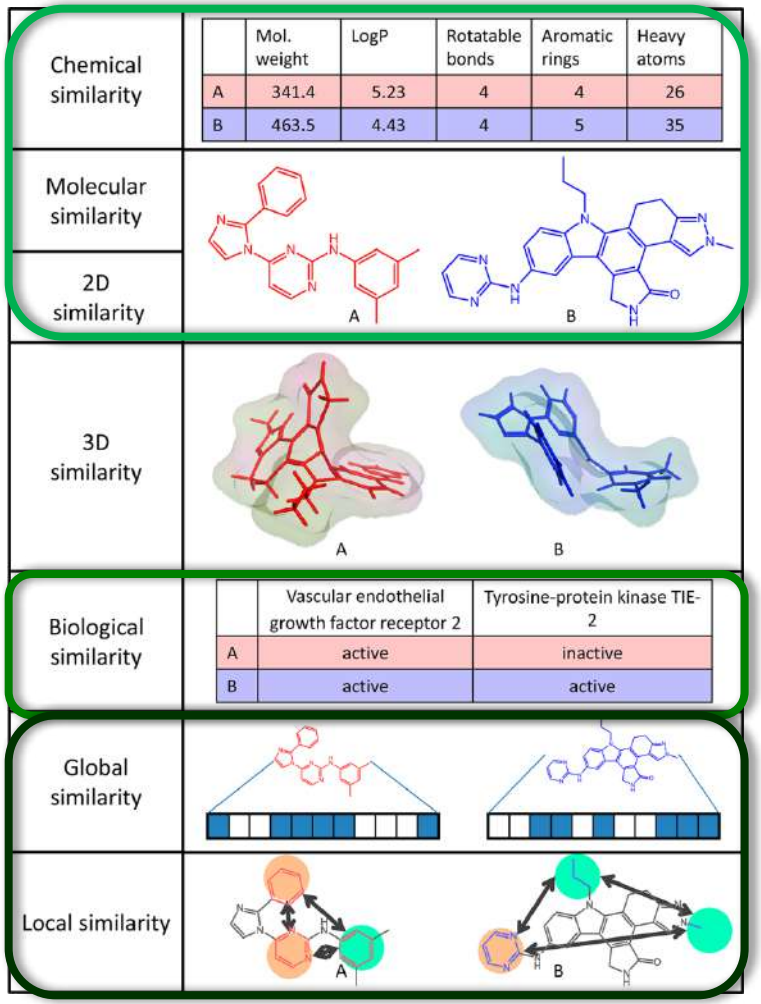
Более общий критерий структурного сходства молекул основан на количестве различных фрагментов, которые присутствуют одновременно в обеих молекулах.
Поиск молекул по такому критерию называется поиском по молекулярному подобию (Similarity Search).

Принцип молекулярного подобия (сходства): структурно схожие молекулы предположительно обладают сходными биологическими свойствами



Опиаты

АНАЛИЗ СХОДСТВА/РАЗНОРОДНОСТИ



	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

	Vascular endothelial growth factor receptor 2	Tyrosine-protein kinase TIE-2
A	active	inactive
B	active	active

Химическое и молекулярное сходство

Химическое сходство – основано на физико-химических характеристиках соединений (растворимость, температура кипения, молекулярный вес, дипольные моменты, электронная плотность)

Молекулярное сходство – структурная характеристика (общие группы, топология)

Биологическое сходство

Различие «биологических подписей» («biological signatures») – оценка биологической активности соединений по ряду мишеней (химическое пространство -> пространство мишеней)

Глобальное и локальное сходство

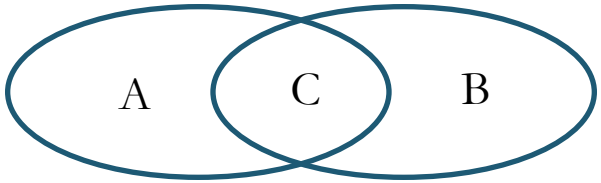
Локальное сходство – присутствие определенных групп
Глобальное – сходство по общему описанию всей структуры

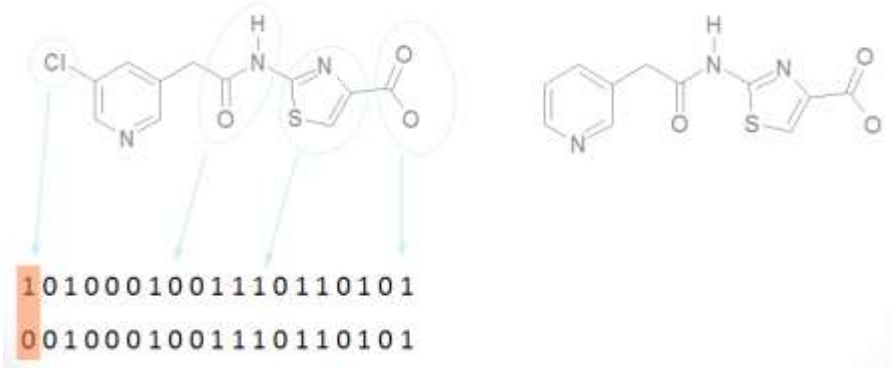
Maggiora *et al., JMC, 2014, 57, 3186-3204*

СТРУКТУРНОЕ СХОДСТВО: НАИБОЛЕЕ ЧАСТО ИСПОЛЬЗУЕМЫЕ В ХИМИЧЕСКОЙ ИНФОРМАТИКЕ КОЭФФИЦИЕНТЫ СХОДСТВА (ПОДОБИЯ)

Структурное сходство

- Дескрипторное описание данных
- Коэффициент подобия
- Метрика





$$s(x,y) = \frac{c}{a+b-c}$$

Tanimoto

$$s(x,y) = \frac{2c}{a+b}$$

Dice

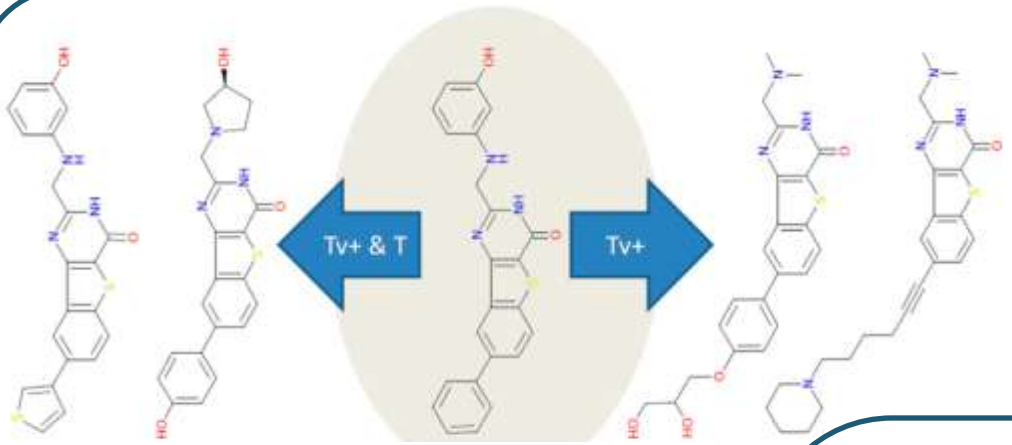
$$s(x,y) = \frac{c}{\sqrt{ab}}$$

Cosine

$$s(x,y) = \frac{c}{\alpha(a-c)+\beta(b-c)+c}$$

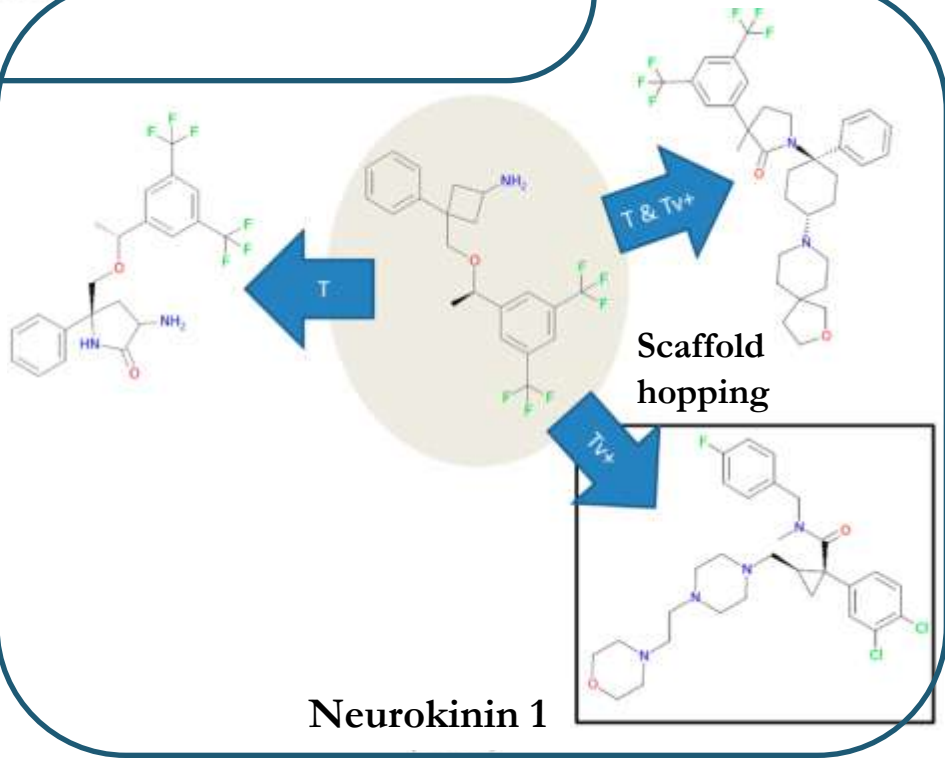
Tversky

DO NOT HESITATE TO USE TVERSKY AND OTHER HINTS FOR SUCCESSFUL ACTIVE ANALOGUE SEARCHES WITH FEATURE COUNT DESCRIPTORS



Лиганды с вариативностью в топологии и фармакофорном описании

PIM1 kinase



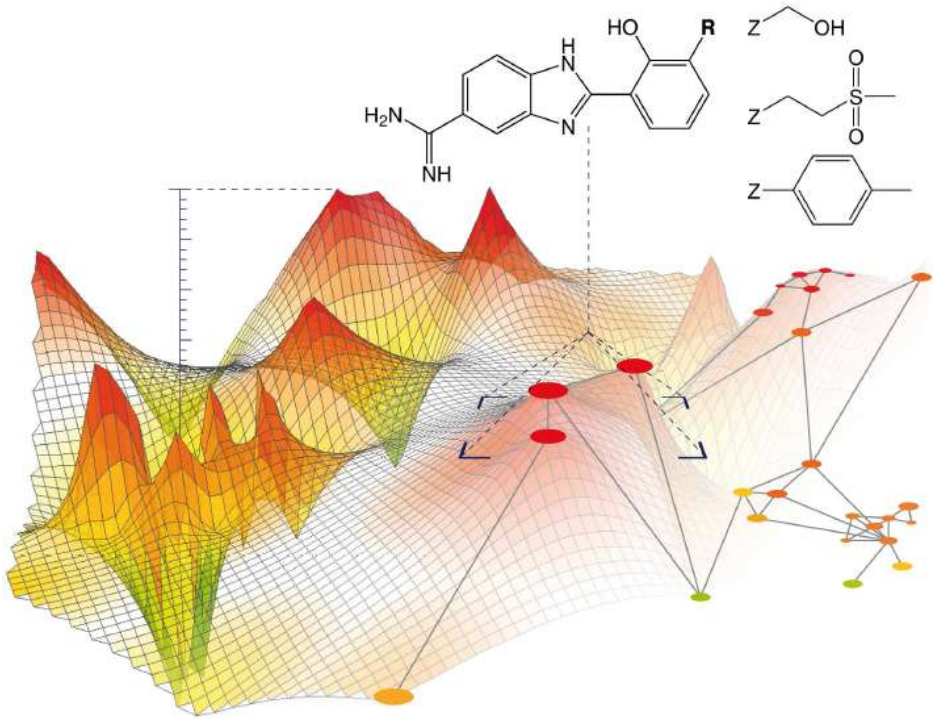
Neurokinin 1

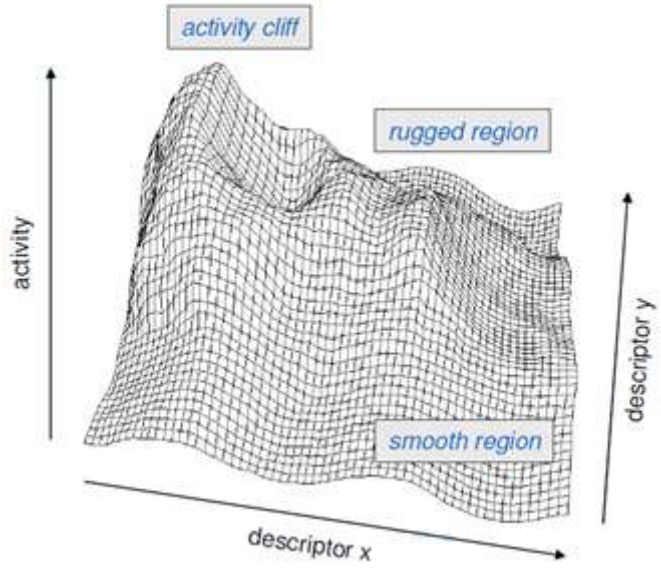
dx.doi.org/10.1021/ci400106g J. Chem. Inf. Model. 2013, 53, 1543 − 1562

ВСЕГДА ЛИ РАБОТАЕТ ПРИНЦИП МОЛЕКУЛЯРНОГО ПОДОБИЯ?

Ландшафт активности

обычно определяется как любое графическое представление, объединяющее молекулярное подобие и взаимосвязь между значениями активности соединений

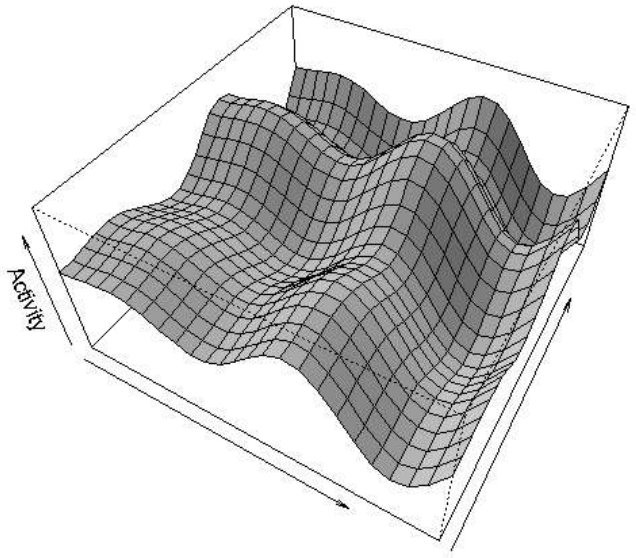




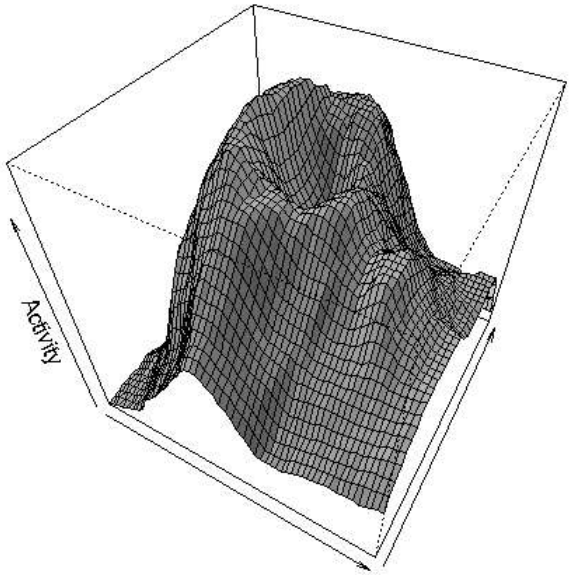
Слайд из курса лекций проф. Ю. Байората

ЛАНДШАФТ АКТИВНОСТИ

continuous SARs



discontinuous SARs



Постепенное изменение в структуре приводит к постепенному изменению в активности

"*холмы*" (G. Maggiora)

Небольшие изменения в структуре приводят к значительным изменениям в значении активности

"*обрывы*" (activity cliffs)

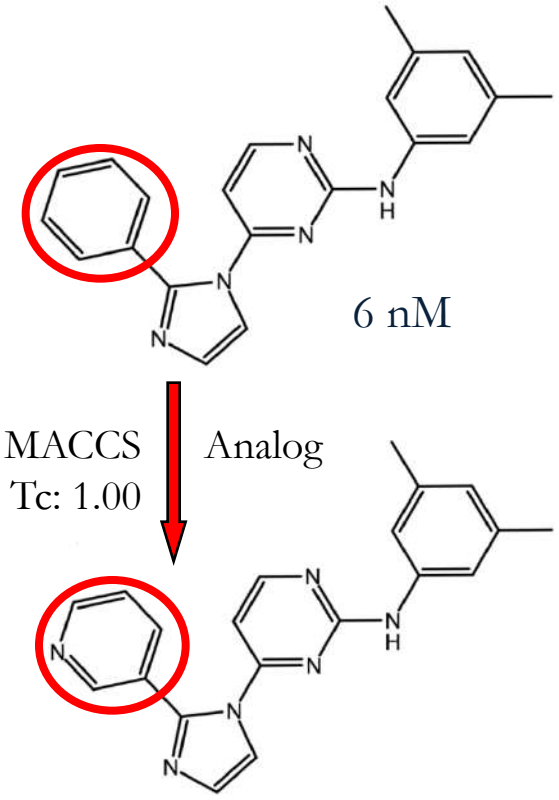
Индекс ландшафта активности (Structure-Activity Landscape Index): $\mathbf{SALI_{ij} = DA_{ij} / DS_{ij}}$

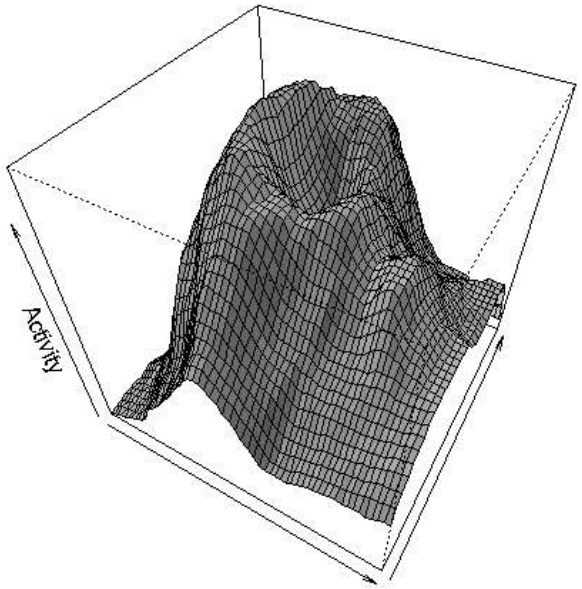
DA_{ij} (DS_{ij}) различия в активности (молекулярном подобии) молекул *I* и *j*

R. Guha et al. J.Chem.Inf.Mod., 2008, ***48****, 646*

ЛАНДШАФТ АКТИВНОСТИ: ОБРЫВЫ (ACTIVITY CLIFFS)

VEGFR-2 tyrosine kinase inhibitors

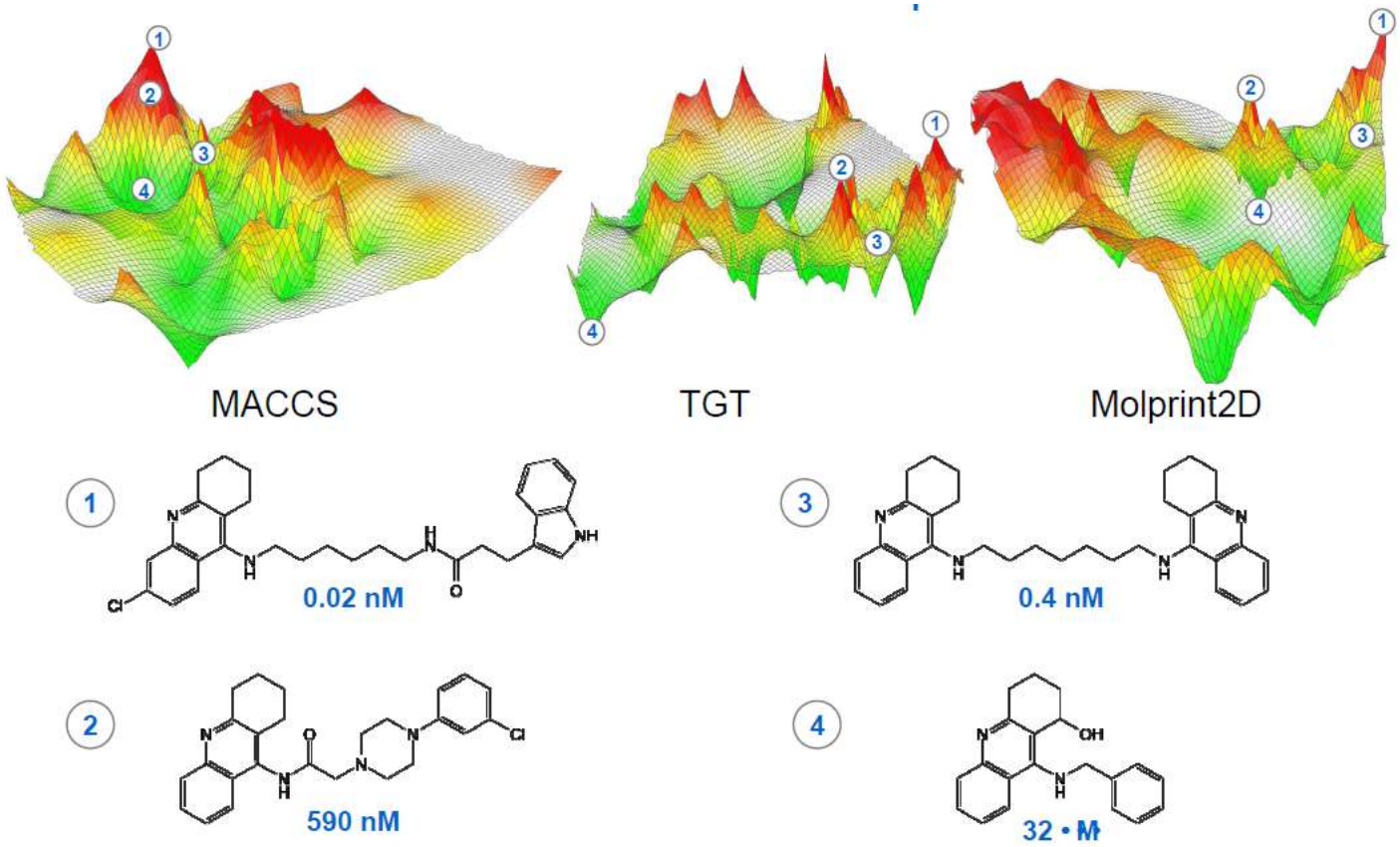




Небольшие изменения в структуре приводят к значительным изменениям в значении активности "*обрывы*" (activity cliffs)

Слайд из курса лекций проф. Ю. Байората

ЛАНДШАФТ АКТИВНОСТИ: ВЫБОР ТИПА ДЕСКРИПТОРОВ



Слайд из лекций проф. Ю. Байората

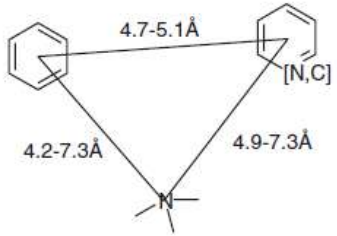
ПОИСК ПО ЗАДАННОМУ ФАРМАКОФОРУ

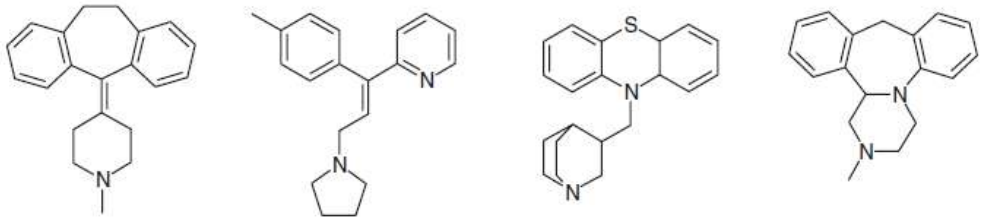
«Фармакофор — это набор пространственных и электронных признаков, необходимых для обеспечения оптимальных супрамолекулярных взаимодействий со специфической биологической мишенью, которые могут вызывать (или блокировать) ее биологический ответ»

«Фармакофор – наибольший общий деноминатор группы лигандов, обладающих сходным биологическим эффектом, распознаваемый тем же сайтом связывания»

Фармакофоры – отдельные функциональные группы или классы веществ, обладающих биологической активностью (сульфонамиды, дигидропиридины), понятие связанное с термином привилегированные структуры

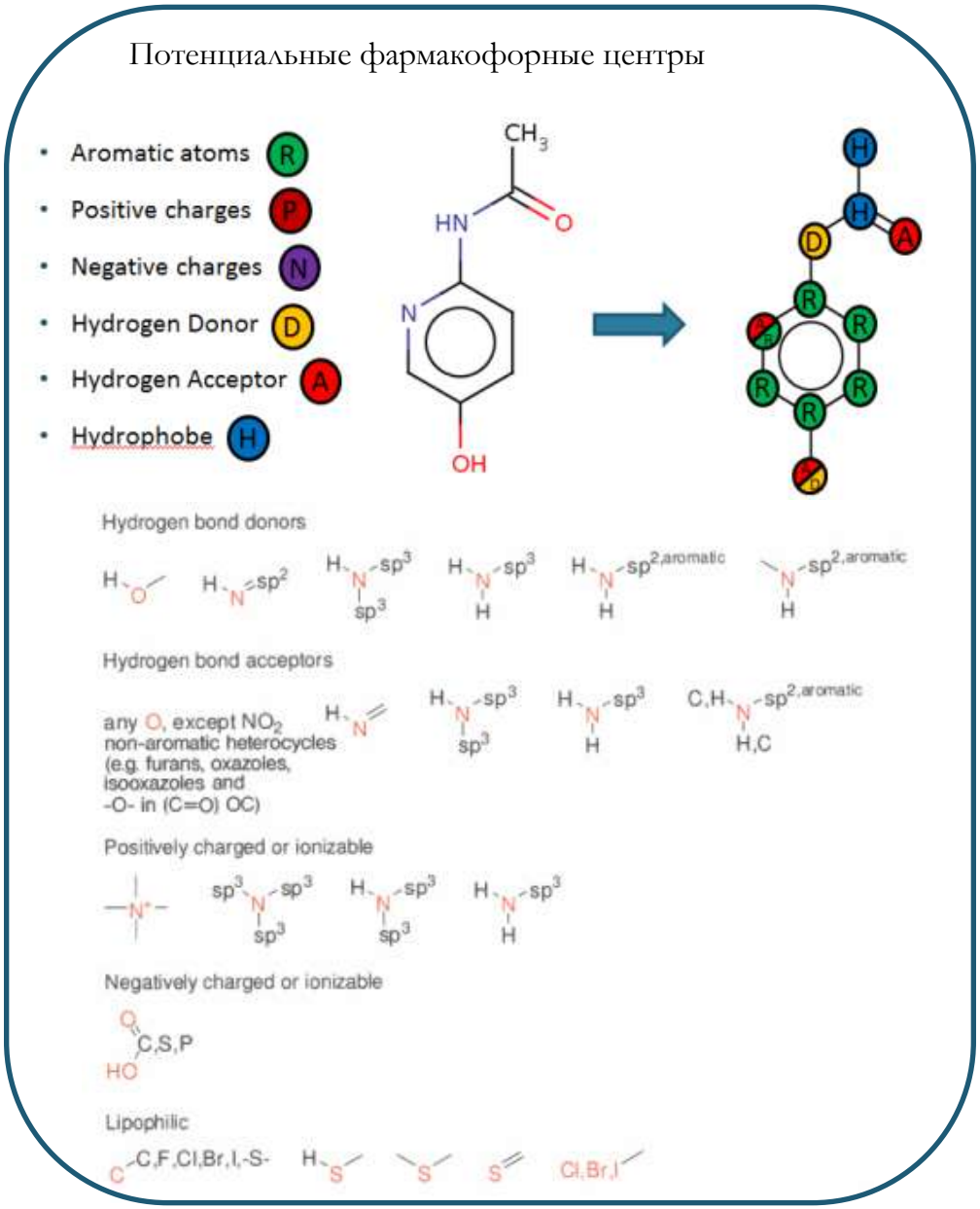
Фармакофорные признаки:
фармакофорные центры и интервалы расстояний между ними, необходимые для проявления данного типа биологической активности.

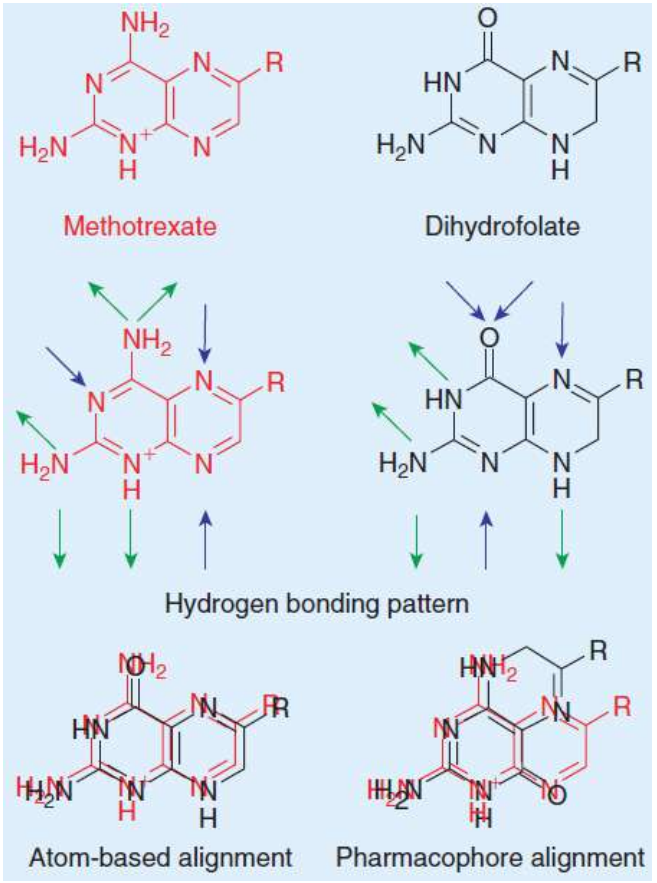




ПОИСК ПО ЗАДАННОМУ ФАРМАКОФОРУ

Потенциальные фармакофорные центры





G. Schneider and K.-H. Baringhaus Molecular Design: Concepts and Applications, Wiley, 2008

C. Wermuth The Practice of Medicinal Chemistry: Third Edition, Elsevier (2008)

ХИМИЧЕСКИЕ БАЗЫ ДАННЫХ

SciFINDER®
A CAS SOLUTION

- Ссылки из более чем 10,000 в настоящее время публикуемых журналов и патентов из более чем 60 различных источников
- Ежедневное обновление
- Крупнейшая мировая база данных по органическим и неорганическим веществам

- CAS REGISTRY – химические соединения
- CAS REACT – химические реакции
- CAplus – интегрированный источник публикаций и патентов
- CHEMCATS – поставщики
- CHEMLIST – регулируемые вещества
- CIN – информация, связанная с промышленным производством
- MARPAT – структуры Маркуша

SciFinder®
A CAS SOLUTION

CAS REGISTRY

CAS REGISTRY: > 120 миллионов органических и неорганических веществ (пополняется ежедневно)
Охватывает информацию по различным типам веществ: сплавы, координационные соединения, минералы, смеси, полимеры, соли, белки и т.д.

CASREACT

Более 77 миллионов химических реакций.
Обновляется ежедневно. Еженедельно обновление на приблизительно 150 000 реакций из публикаций и патентов

CAplus

Интегрированный источник публикаций и патентов: 43 миллиона записей в 80 различных категориях. Обновляется ежедневно. Все патенты, соответствующие критериям отбора CAS, из 9 основных патентных ведомств, доступны через Интернет в CAplus в течение 2 дней с момента выдачи патентов, полностью индексируются менее чем за 27 дней с даты появления

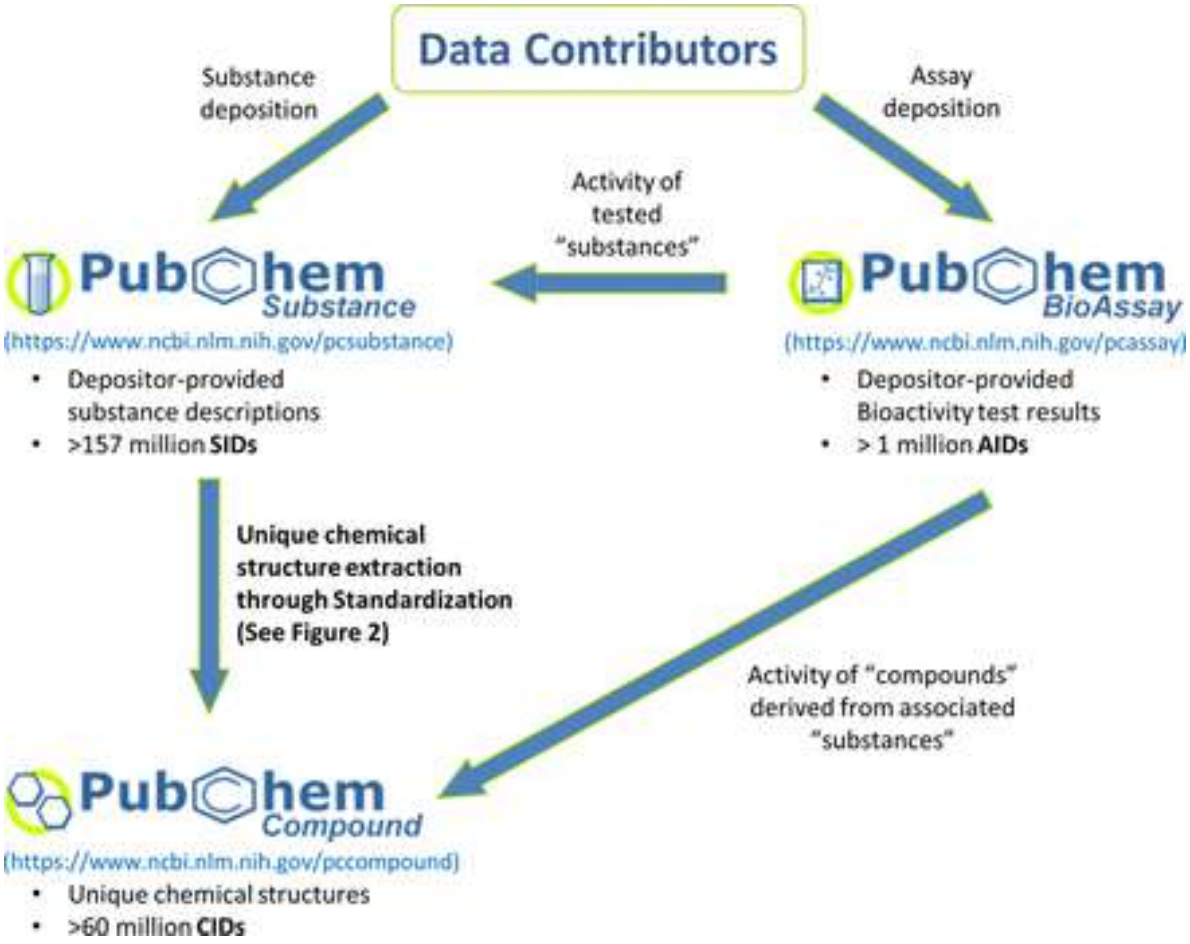
PubChem

Создана в 2004 как часть Molecular Libraries Roadmap Initiatives Национального Института здоровья США

База данных из > 60 миллионов соединений с расширенными возможностями поиска: по номеру, по названию, по структурной формуле, по подструктуре и по сходству, дополнительными критериями поиска. База постоянно пополняется из более чем 350 источников (другие базы данных (CheMBL, BindingDB, DrugBank) , гос. организации, университеты, фармацевтические компании и т.д.).

URL: http://pubchem.ncbi.nlm.nih.gov/

PubChem



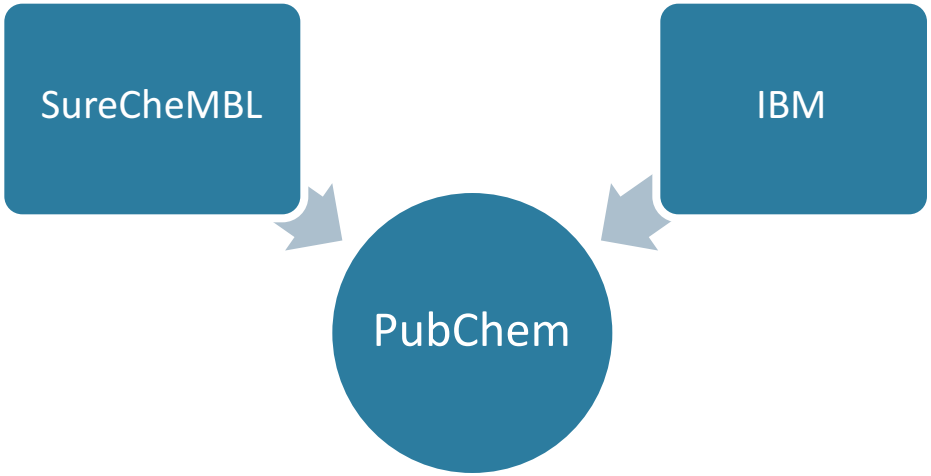
- 157 million depositor-provided chemical substance descriptions,
- 60 million unique chemical structures
- 1 million biological assay descriptions, covering about 10 thousand unique protein target sequences.

PubChem

SureChEMBL: 17 миллионов соединений, извлеченных из более чем 14 млн патентов

https://www.surechembl.org/search/

IBM (http://www.almaden.ibm.com/)

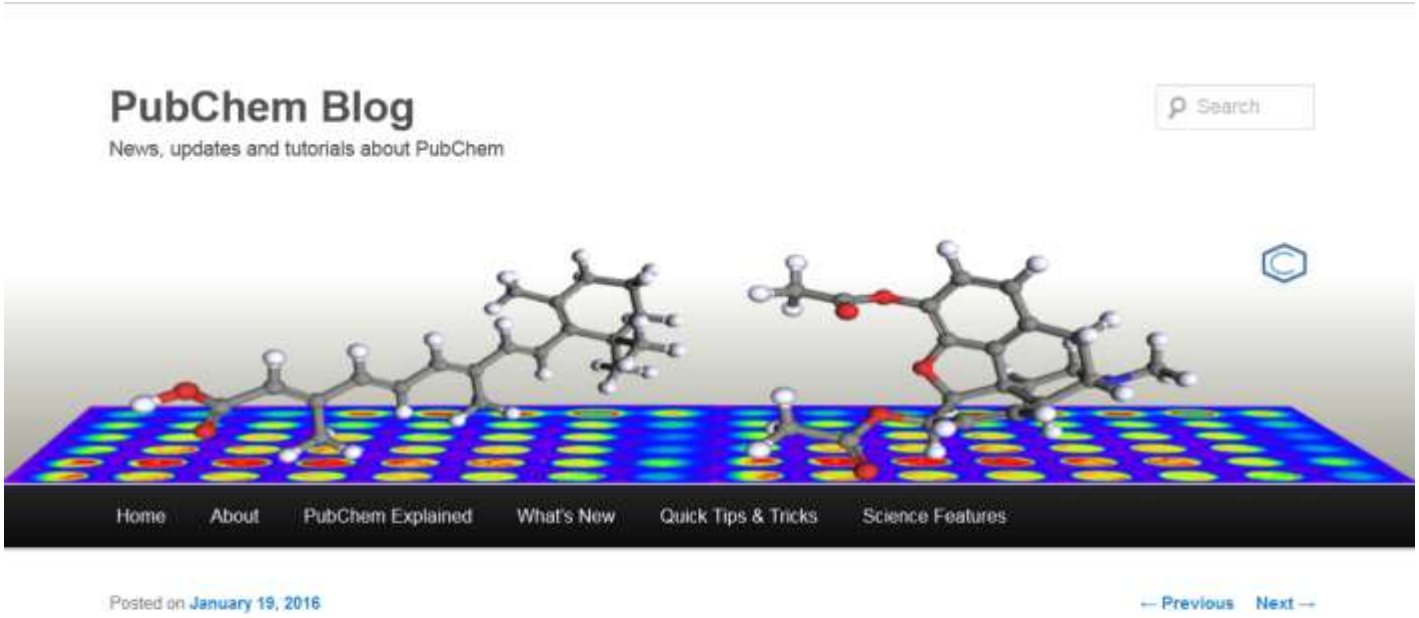


6 млн патентов, более 16 миллионов уникальных химических структур, более 329 миллионов источников химическое соединение-патент (US, European and World Intellectual Property)

PubChem: некоторые инструменты и сервисы

•**Chemical structure search (https://pubchem.ncbi.nlm.nih.gov/search/search.cgi)**
•Allows users to query the PubChem Compound database by chemical structure or chemical structure pattern.
•**PubChem Upload (https://pubchem.ncbi.nlm.nih.gov/upload/)**
A data submission system that enables one to contribute substance or assay data to PubChem.
•**Classification browser (https://pubchem.ncbi.nlm.nih.gov/classification/)**
Allows users to browse PubChem data using a classification of interest, or search for records annotated with the desired classification/term.
•**Score matrix service (https://pubchem.ncbi.nlm.nih.gov/score_matrix/)** Computes matrices of 2-D and 3-D similarity scores for a given set of compounds.
•**Widgets (https://pubchem.ncbi.nlm.nih.gov/widget/docs/)**
Provides a rapid way to display some commonly requested PubChem data views.
•**Pc3D viewer (https://pubchem.ncbi.nlm.nih.gov/pc3d/)**
An interactive 3-D molecular viewer that can be downloaded and installed on local machines.
•**Structure download (https://pubchem.ncbi.nlm.nih.gov/pc_fetch/)** Downloads a set of substance or compound records in PubChem.
•**PubChemRDF (https://pubchem.ncbi.nlm.nih.gov/rdf/)**
the RDF-based resource compatible with Semantic Web standards and technologies.
Nucleic Acids Research, 2016, Vol. 44, Database issue
D1203
• **Chemical structure sketcher (https://pubchem.ncbi.nlm.nih.gov/edit/)**
A platform-independent 2-D molecule drawer, compatible with major web browsers.
• **Standardization service (https://pubchem.ncbi.nlm.nih.gov/standardize/)**
Validates and normalizes an input chemical structure in the same way as PubChem standardization process.
• **Identifier exchange service (https://pubchem.ncbi.nlm.nih.gov/idexchange/)**
Converts one type of identifiers for a given set of chemical structures into a different type of identifiers for identical or similar chemical structures.
• **Structure clustering (https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=clustering)**
Clusters compounds/substances based on their structural similarity using the Single Linkage Algorithm.
• **Web-based 3D viewer (https://pubchem.ncbi.nlm.nih.gov/vw3d/)**
An interactive web-based viewer for 3-D conformations of molecules, which visualizes 3-D information available within PubChem.

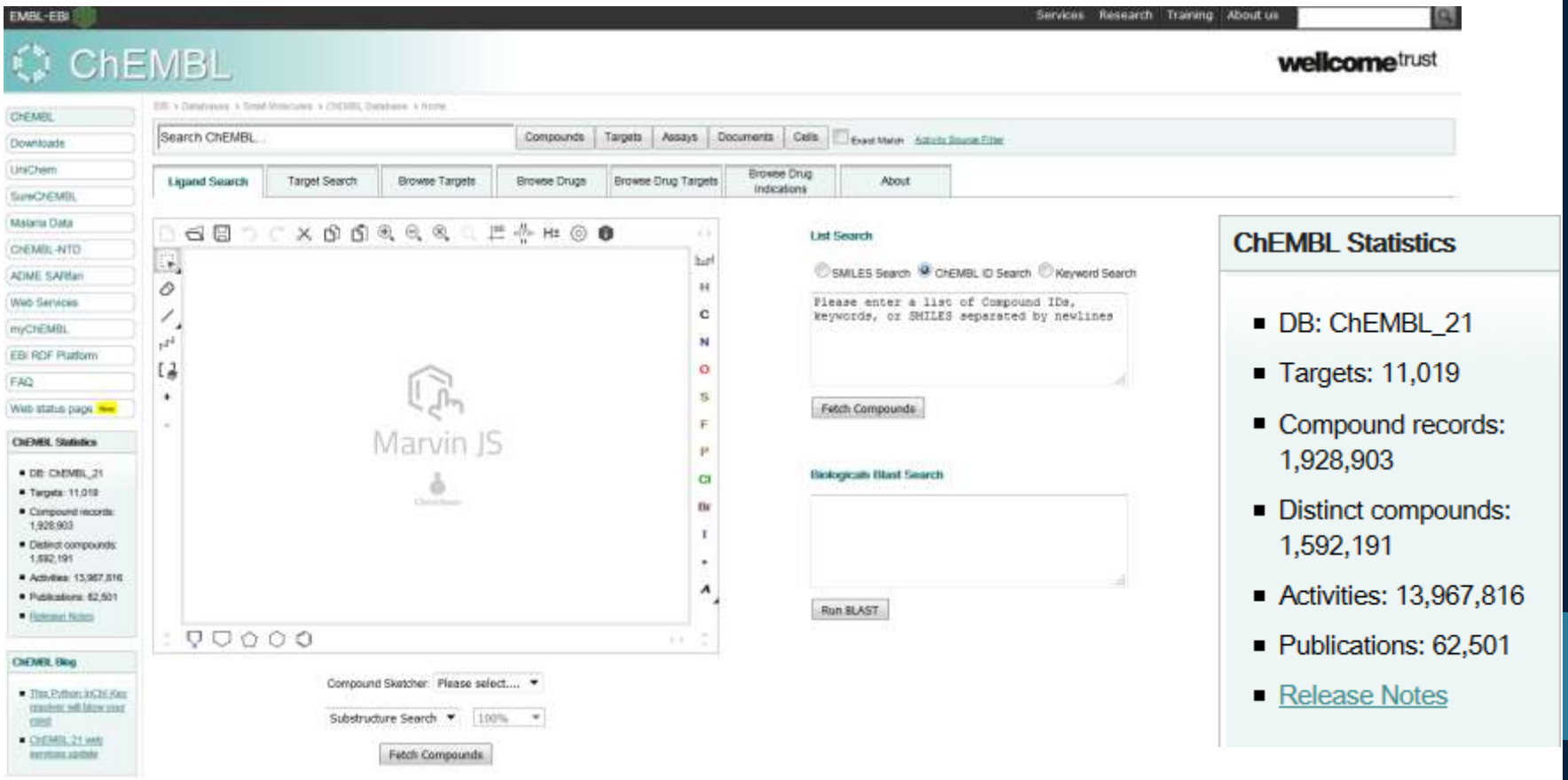
PubChem



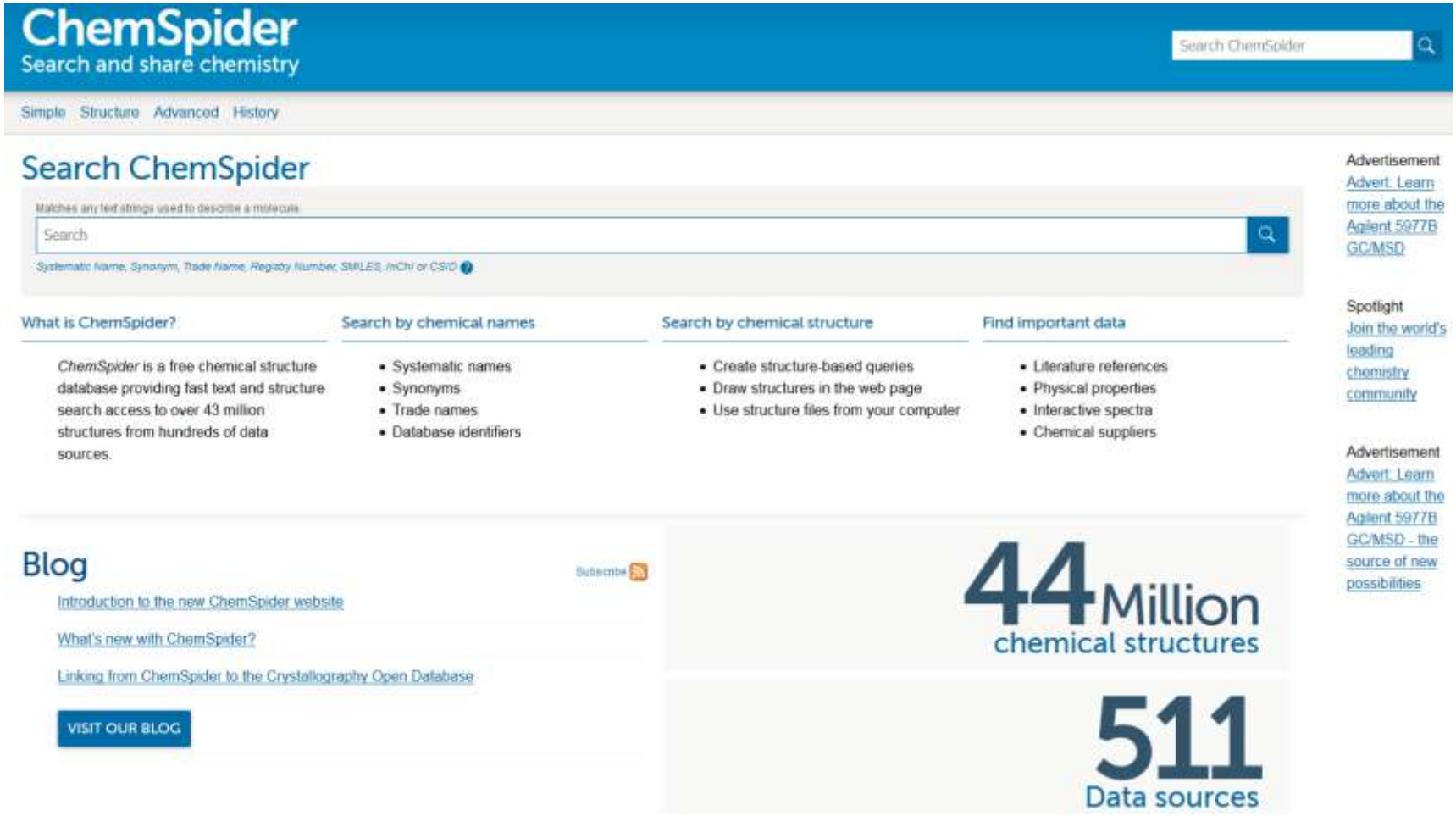
https://pubchemblog.ncbi.nlm.nih.gov/2016/01/19/recent-pubchem-publications-read-about-whats-new/

ChEMBL

ChEMBL – база данных биологически активных «drug-like» молекул, мишеней, лекарственных средств. Содержит 2D структуры, рассчитанные значения свойств (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) и указанные активности.



ChemSpider



URL: http://www.chemspider.com/

BindingDB

The Binding Database

Home | Info | Download | About us | Email us | Contribute data | Web Services

myBDB logout

- **Search and Browse**
  - **Target**
    - Sequence
    - Name &
      - Ki IC50 Kd EC50
      - Rate constants
      - ΔG° ΔH° -TΔS°
      - pH (Enzymatic Assay)
      - pH (ITC)
      - Substrate or Competitor
      - Compound Mol. Wt.
      - Chemical Structure
    - Pathways
    - Source Organism
    - Number of Compounds
    - Monomer List in csv
    - Het List in SDF
  - **Compound**
    - FDA Drugs
    - Important Compounds
    - Chemical Structure
    - Name
    - SMILES
    - Number of Data / Targets
  - **Special tools**
    - 3D Structure Series

BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules. BindingDB contains 1,186,317 binding data, for 6,191 protein targets and 519,724 small molecules.

There are 2291 protein-ligand crystal structures with BindingDB affinity measurements for proteins with 100% sequence identity, and 5816 crystal structures allowing proteins to 85% sequence identity.

Simple Search
Article Titles, Authors, Assays, Compound Names, Target Names

Go

Use ? for single-letter wild-card or * for general wild-card.
For example, "adeny*" or "adeny?". Query cannot start with wild card.

Advanced Search — Combine multiple search criteria, such as chemical structures, target names, and numerical affinities; restrict searches by data source, such as BindingDB, ChEMBL, PubChem, and Patents.

Messages

We are delighted to announce that Elsevier's Science Direct journals now include links from articles to BindingDB datasets, where available! For an example, go to this article, and see the "Data for this Article, BindingDB" link on the right. (December 2015)

The BindingDB survey is now over. Thanks so much for your help in filling it out! (January 2016)

Journal Curation by BindingDB

BindingDB continually curates a set of journals not covered by other public databases. As of January 2016, the status of our current curation effort is as follows:

- **ACS Chemical Biology** 2006-2015 (vol 1-10)
- **ACS BioChemistry** 1962-1970 (vol 1-9), 1991-2015 (vol 30-54)
- **Bioorganic Chemistry** 1971-2015 (vol 1-62)
- **BMC Chemical Biology** 2001-2012 (vol 1-12)
- **ChemBioChem** 2000-2015 (vol 1-16)
- **Chemical Biology & Drug Design** 2006-2015 (vol 67-86)
- **Chemistry & Biology** 1994-2014 (vol 1-20)
- **Journal of Biological Chemistry** 1988-2013 (vol 264-288)
- **Journal of Chemical Biology** 2008-2013 (vol 1-6)

BindingDB News

September 2015. We plan to post the latest BindingDB user survey on or about October 5. We would greatly appreciate your feedback and suggestions! Note, however, that you are always free to click through the survey to the regular home-page.

September 2015. All compounds in BindingDB have now been assigned a BindingDB Molecule ID, such as BDBM50183711. (The numeric component is also known, internally, as a Monomer ID.)

September 2015. BindingDB should give you faster performance now, as we have upgraded the main server. Please let us know immediately if you notice any problems.

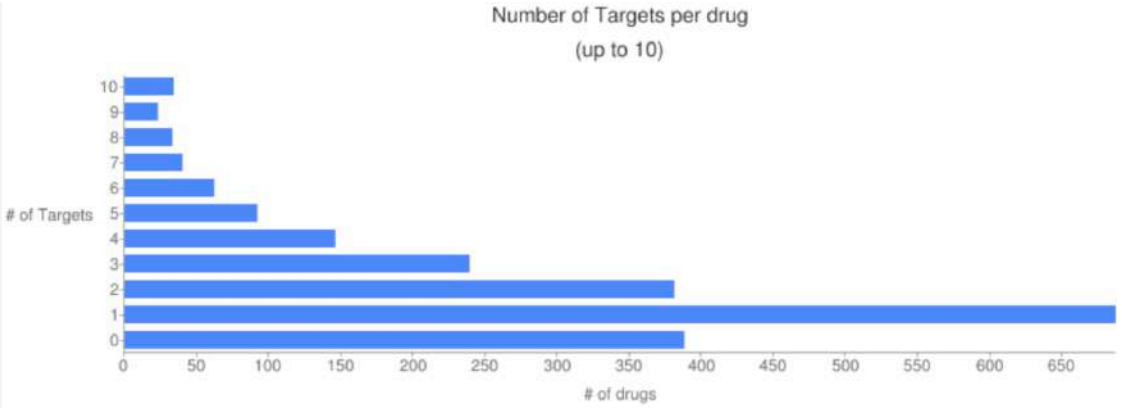
July 2015. Please try our new tool to map from one or more proteins of known sequence to known or potential ligands: Find Compounds for My Targets.

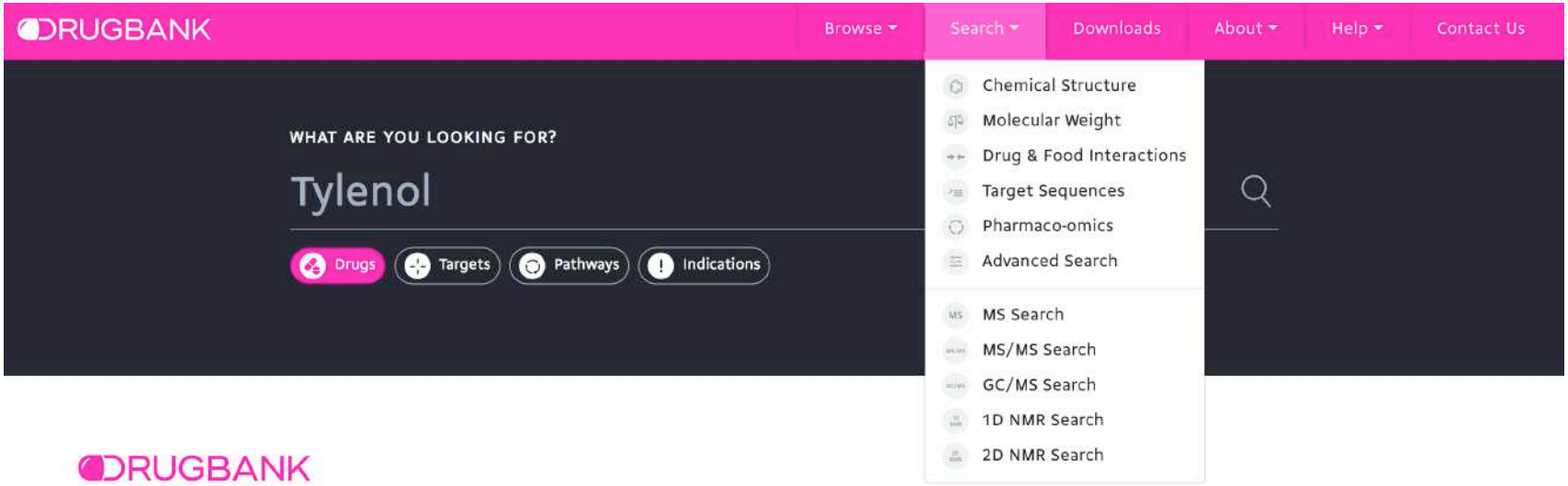
April 2015. BindingDB has

http://bindingdb.org/bind/index.jsp

DrugBank

Total Number of Small Molecule Drugs	9041
Total Number of Biotech Drugs	549
Total Number of Approved Drugs	2278
Total Number of Approved Small Molecule Drugs	2037
Total Number of Nutraceutical Drugs	96
Total Number of Experimental Drugs	4953
Total Number of Illicit Drugs	202
Total Number of Withdrawn Drugs	210
Total Number of Drugs	9590



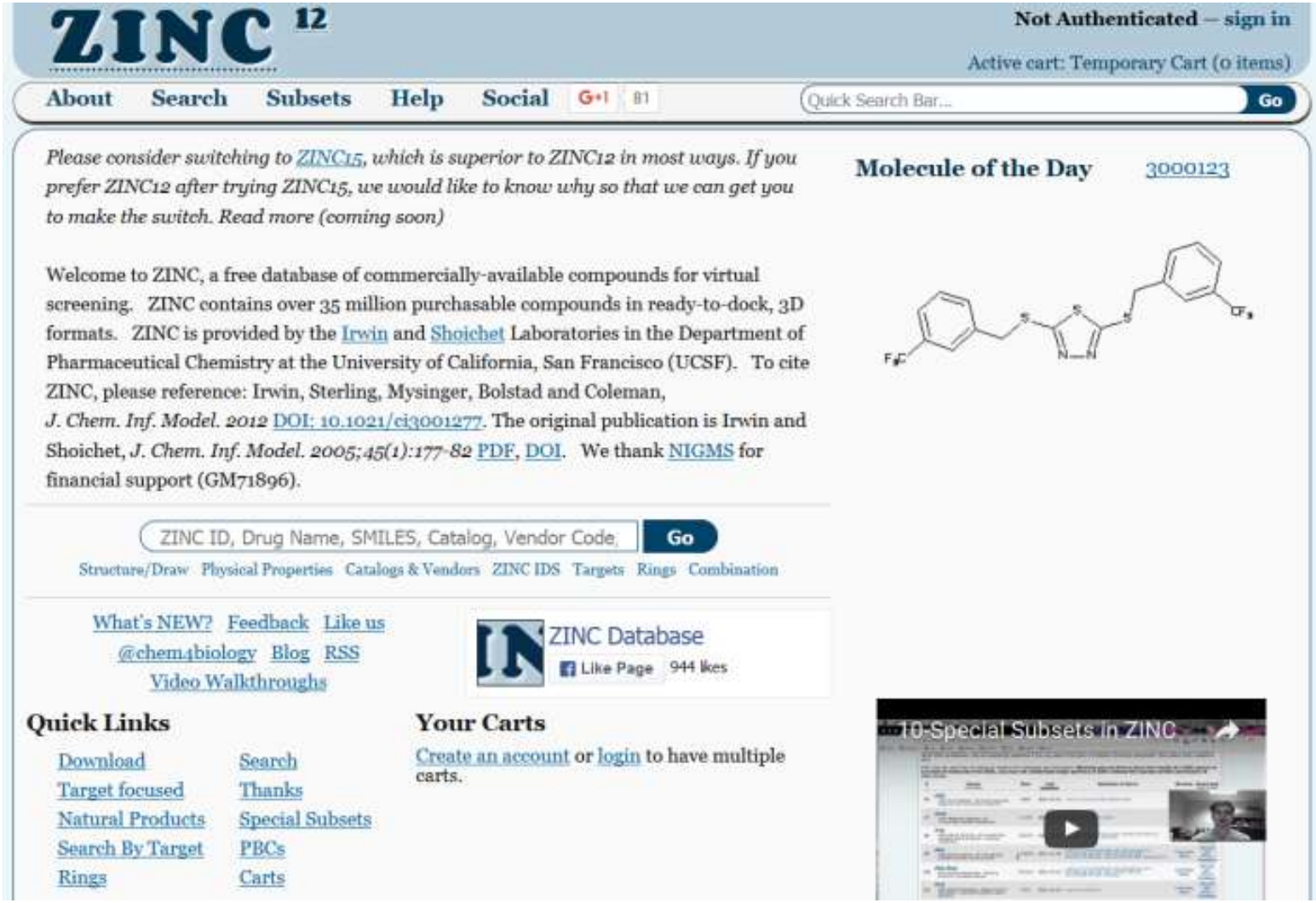


DRUGBANK

http://www.drugbank.ca/

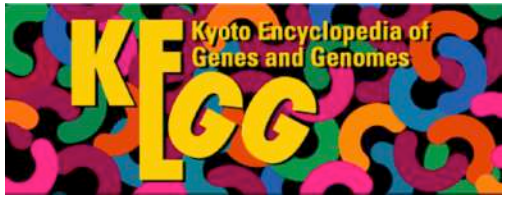
ZINC

ZINC, свободнодоступная база данных коммерчески доступных соединений для виртуального скрининга, содержит около 35 миллионов соединений в 3D формате.



http://zinc.docking.org/

KEGG DATABASE



https://www.kegg.jp/kegg/

KEGG Database as of 2018/9/25

Systems information		
KEGG PATHWAY	Pathway maps, reference (total)	530 (566,553)
KEGG BRITE	Functional hierarchies, reference (total)	203 (213,814)
KEGG MODULE	KEGG modules, reference (total)	801 (487,867)
Genomic information		
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	22,290
KEGG GENOME	KEGG organisms and selected viruses (457 eukaryotes, 4872 bacteria, 278 archaea, and 328 viruses)	5,935
KEGG GENES	Genes in KEGG organisms and other categories (including 4,350 addendum, 318,764 viral) (see annotation statistics)	27,309,951
KEGG SSDB	Best hit relations within GENES Bi-directional best hit relations within GENES	255,897,345,859 13,648,361,607
Chemical information		
KEGG COMPOUND	Metabolites and other small molecules	18,384
KEGG GLYCAN	Glycans	11,040
KEGG REACTION	Biochemical reactions	11,012
KEGG RCLASS	Reaction class	3,147
KEGG ENZYME	Enzyme nomenclature	7,354
Health information		
KEGG NETWORK	Disease-related network elements	448
KEGG VARIANT	Human gene variants	169
KEGG DISEASE	Human diseases	2,194
KEGG DRUG	Drugs	10,538
KEGG DGROUP	Drug groups	2,115
KEGG ENVIRON	Crude drugs and health-related substances	856
Drug labels		
KEGG MEDICUS	Japanese prescription drug labels from JAPIC Japanese OTC drug labels from JAPIC	14,494 11,151
KEGG MEDICUS	FDA prescription drug labels linked to DailyMed FDA OTC drug labels linked to DailyMed	32,085 46,469

Reaxys (Beilstein + Gmelin)

reaxys®

Innovation from CrossFire Beilstein

Reaxys = Beilstein + Gmelin Databases

Beilstein Database:

10 миллионов соединений, 10 миллионов реакций и 35 миллионов химических, физических, экологических и токсикологических данных из Beilstein Handbook of Organic Chemistry и 180 журналов в области органической химии

Gmelin Database:

Содержит неорганические и металлоорганические соединения начиная с 1772 по наше время. На данный момент содержит более 2.5 миллионов соединений, 1.9 миллиона реакций и 1.3 миллиона ссылок.

http://www.info.reaxys.com/

SCUBIDOO (SCREENABLE CHEMICAL UNIVERSE BASED ON INTUITIVE DATA ORGANIZATION): DATABASE OF COMPUTATIONALLY GENERATED SYNTHETIC TRACTABLE COMPOUNDS

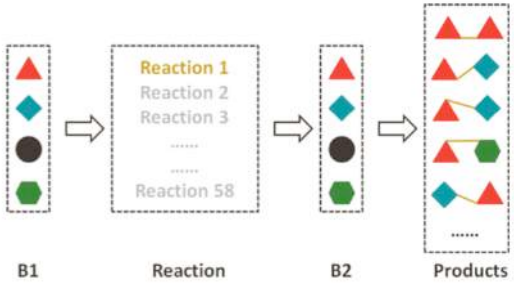
~18000 «строительных блоков»

58 наиболее распространенных в медицинской химии реакций

21 миллион новых соединений

Предварительная фильтрация:

- Удаление дубликатов
- Удаление противоионов
- Контроль «молекулярной сложности»:
  - MW ≤ 250 Da. (для продуктов с большей вероятностью MW ≤ 500 Da).
  - Количество конформационных связей ≤ 2. (для продуктов ≤ 6 => возможность применения методов молекулярного докинга) Количество хиральных центров ≤ 1 (для продуктов ≤ 3 => упрощает синтез)

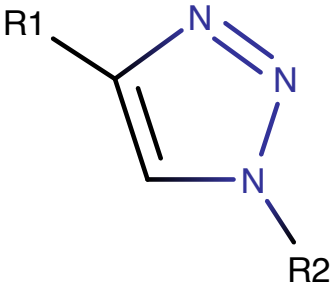


Для каждого соединения в базе данных представлена информация по методу синтеза, потенциальным побочным реакциям, и альтернативным способам получения

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

ZINClick: База данных 1,4-дизамещенных триазолов

~16 миллионов 1,4-дизамещенных-1,2,3-триазолов

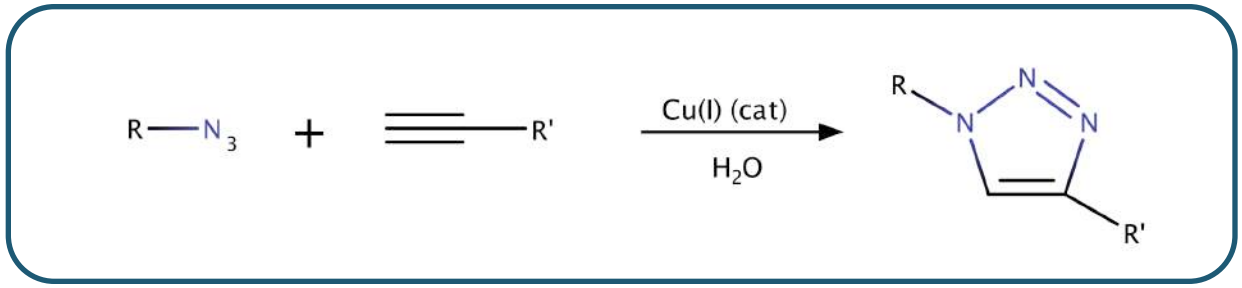


Соединения активно участвующие во взаимодействиях лиганд-рецептор, относящиеся к новому для медицинской химии классу соединений

Понятие «клик-химия» было введено Барри Шарплесом в 2001 для описания реакций, направленных на быстрое и надёжное получение химических веществ путём соединения между собой отдельных маленьких элементов, характеризующихся:

- ❖ высоким выходом
- ❖ модульностью
- ❖ стереоспецифичностью
- ❖ протеканием в физиологических условиях
- ❖ образованием единственного продукта (быть термодинамически выгодными)
- ❖ широкой областью применения

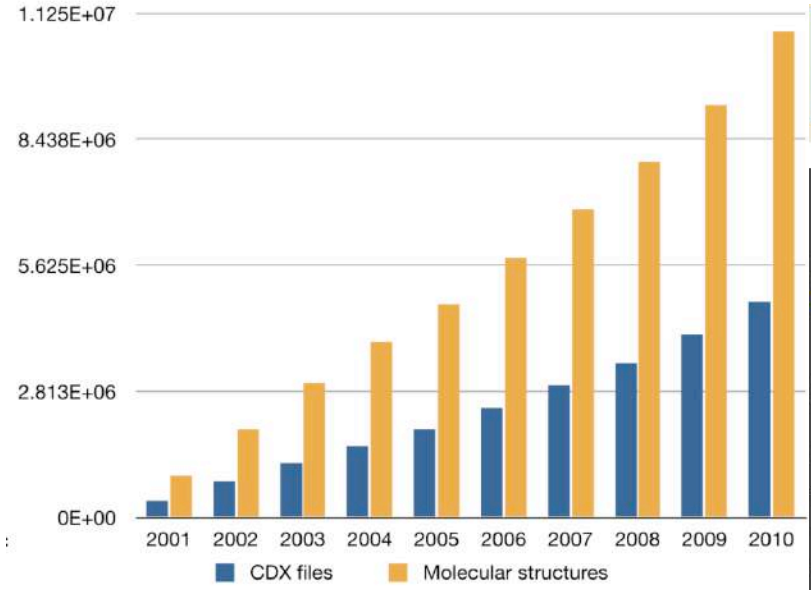
Реакция азид-алкинового циклоприсоединения:

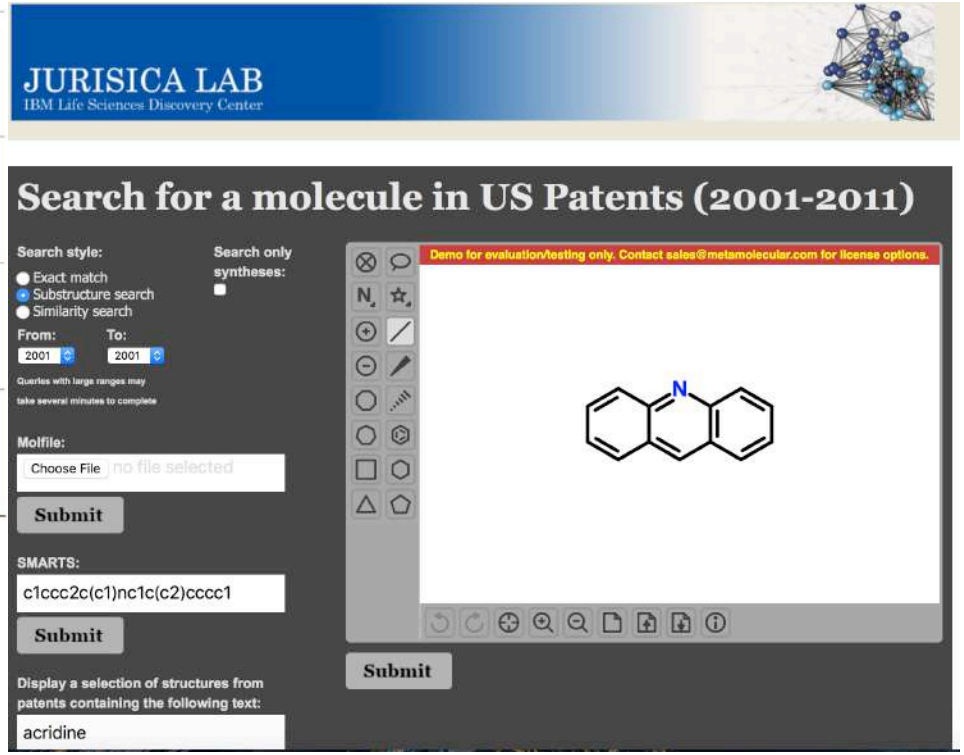


Massarotti et al ZINClick: A Database of 16 Million Novel, Patentable, and Readily Synthesizable 1,4-Disubstituted Triazoles *JCIM* 2014 *54* (2), 396-406

www.symech.it/ZINClick

SCRIPDB: БАЗА ДАННЫХ ХИМИЧЕСКИХ СТРУКТУР С АССОЦИИРОВАННОЙ ПАТЕНТНОЙ ИНФОРМАЦИЕЙ

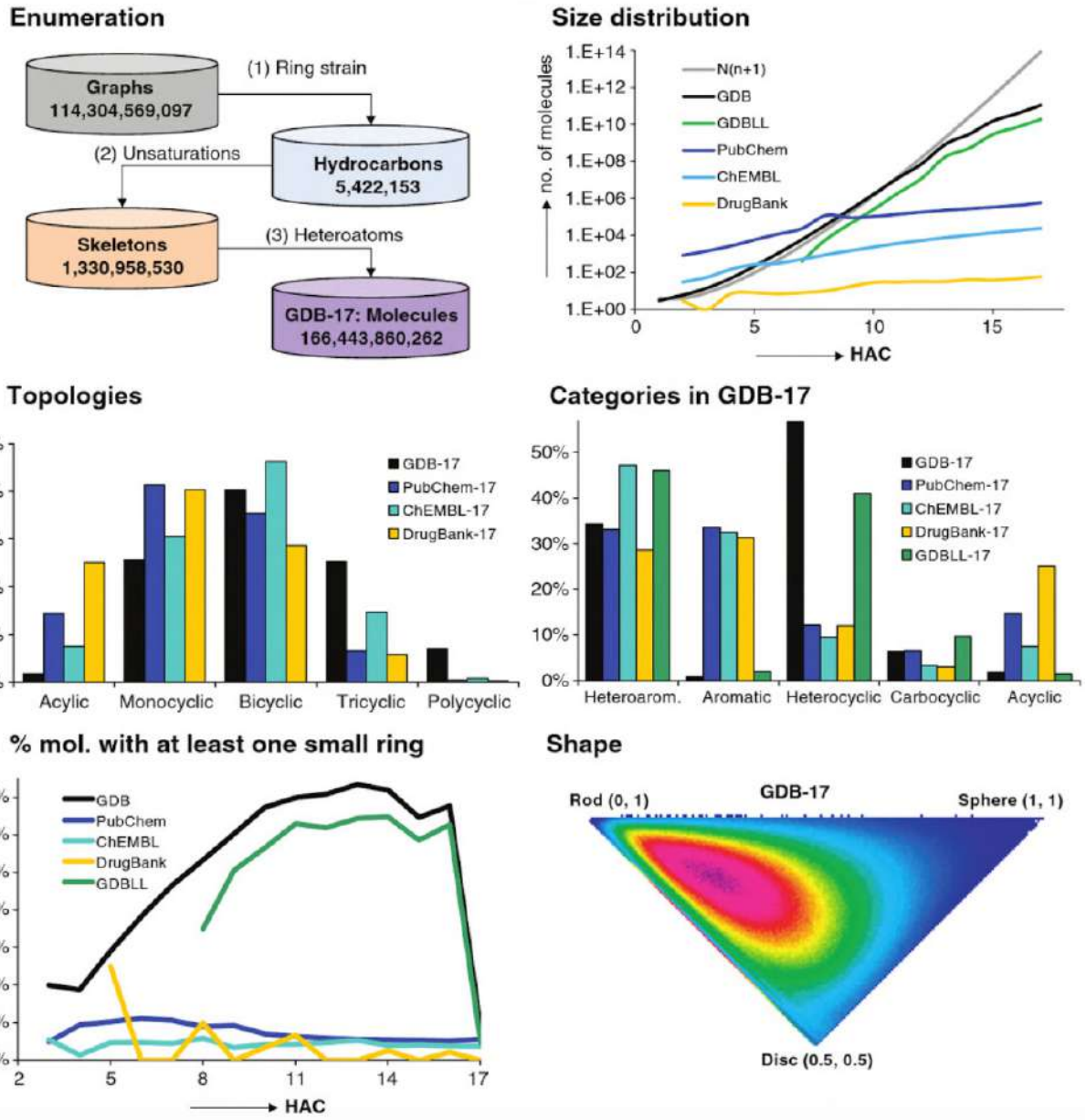




http://dcv.uhnres.utoronto.ca/SCRIPDB

GDB-17 DATABASE

166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens



Fragment-based Drug Discovery: Lessons and Outlook Edited by Daniel A. Erlanson, Wolfgang Jahnke, Raimund Mannhold, Hugo Kubinyi, Gerd Folkers Wiley 2006

Некоторые другие базы данных

Databases for Virtual Screening

ChemBank	http://chembank.broadinstitute.org
eMolecules	https://www.emolecules.com
DUD	http://dud.docking.org
MUV	https://www.tubraunschweig.de/pharmchem/forschung/baumann/muv
CoCoCo	http://cococo.isof.cnr.it
Wombat	http://dud.docking.org/wombat/
TCM	http://tcm.cmu.edu.tw
DNP	http://dnp.chemnetbase.com/intro/index.jsp
SPRESI	https://www.spresi.com
SORD	http://www.sord.nl/SOR-Database.html
Bioinfo	http://bioinfo-pharma.u-strasbg.fr/bioinfo/
SC PDB Frag	http://bioinfo-pharma.u-strasbg.fr/scPDBFrag/

Базы данных в области неорганической химии и материаловедения

Materials Project
https://www.materialsproject.org

AFLOWLIB
http://www.aflowlib.org

Harvard Clean Energy
https://cepdb.molecularspace.org

NIMS materials database
http://mits.nims.go.jp/index_en.html

База данных по информационным ресурсам в области неорганической химии и материаловедения
http://iric.imet-db.ru

MatWeb
http://www.matweb.com

MatBase
http://www.matbase.com

Pauling File
http://paulingfile.com

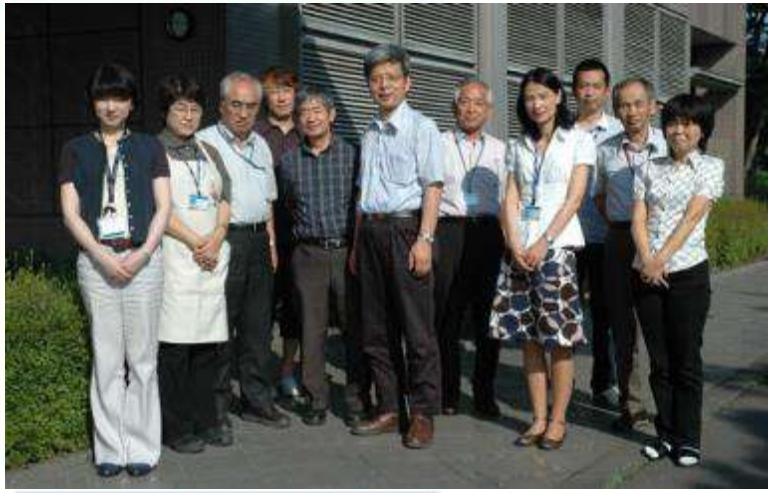
MatNavi
NIMS Materials Database

Japanese | For New User | National Institute for Materials Science, Materials Information Station

Home | About us | MITS Symposium | Link | Contact us | NIMS

MatNavi

The use of "MatNavi" is free.
All you need to do is register.



Database

Basic Properties

- Polymer Database (PoLyInfo)
- Inorganic Material Database (AtomWork)
- Computational Phase Diagram Database (CPDDB)
- Computational Electronic Structure Database (CompES)
- Database of Promising Adsorbents for Decontamination of Radioactive Substances (READS)
- Neutron Transmutation Database (NeuTran)
- Interfacial Thermal Conductance Database (ITC)
- Diffusion Database (Kakusan)
- Superconducting Material Database (SuperCon)

NIMS Structural Materials Data Sheet Online

- Creep Data Sheet (CDS)
- Fatigue Data Sheet (FDS)
- Corrosion Data Sheet (CoDS)
- Space Use Materials Strength Data Sheet (SDS)
- Metallic Material Microstructure Database (Kinso)

[Printed copy]

Engineering

- Metallic Material Database (Kinzoku)
- CCT Diagram Database (CCTD)
- Materials Risk Information Platform (MRiP)
- FGMs Database

Applications

- Composite Design & Property Prediction System (CompoTherm)
- Polymer Properties Prediction System
- Metal Segregation Prediction System (SurfSeg)
- Interface Bonding Prediction System (InerChemBond)

MatNavi Search

MatNavi Search

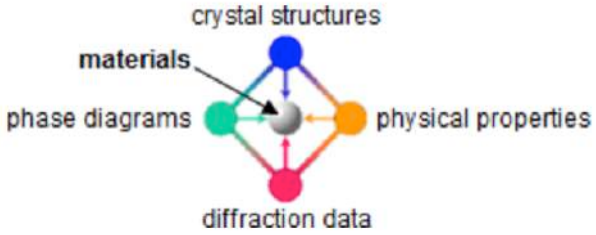
Keyword Search

search

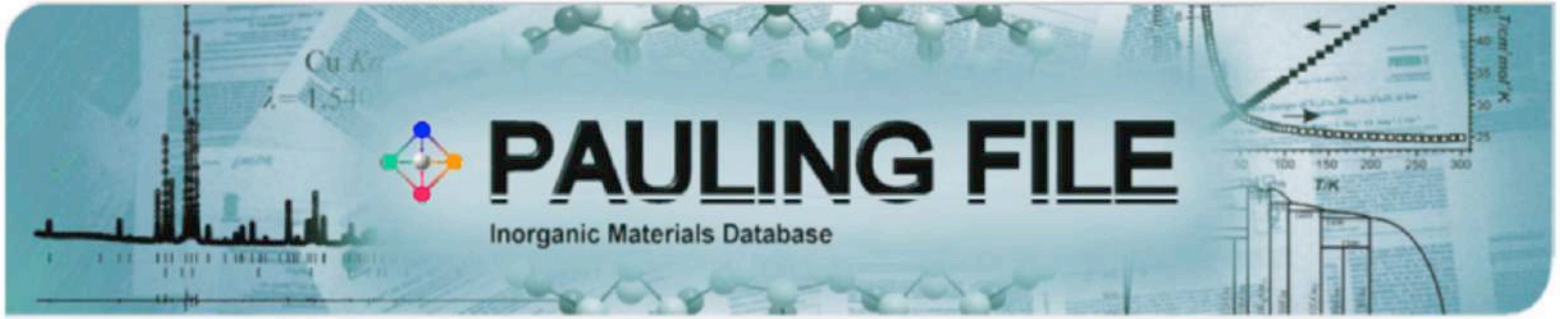
AND OR Contain

Tree Search

- Material
  - Element
  - Alloy
  - Ceramic
  - Polymer
- Property



phase diagrams + crystal structures + physical properties
together in the world largest database for inorganic compounds



Dear Dr. Villars: 27 August 1993

It is surely fine that you are starting a very large scale project to extend the database to cover all non-organic solid state materials. Also, it is fine that you want to call it the PAULING'S FILE.

This letter can serve as my permission.

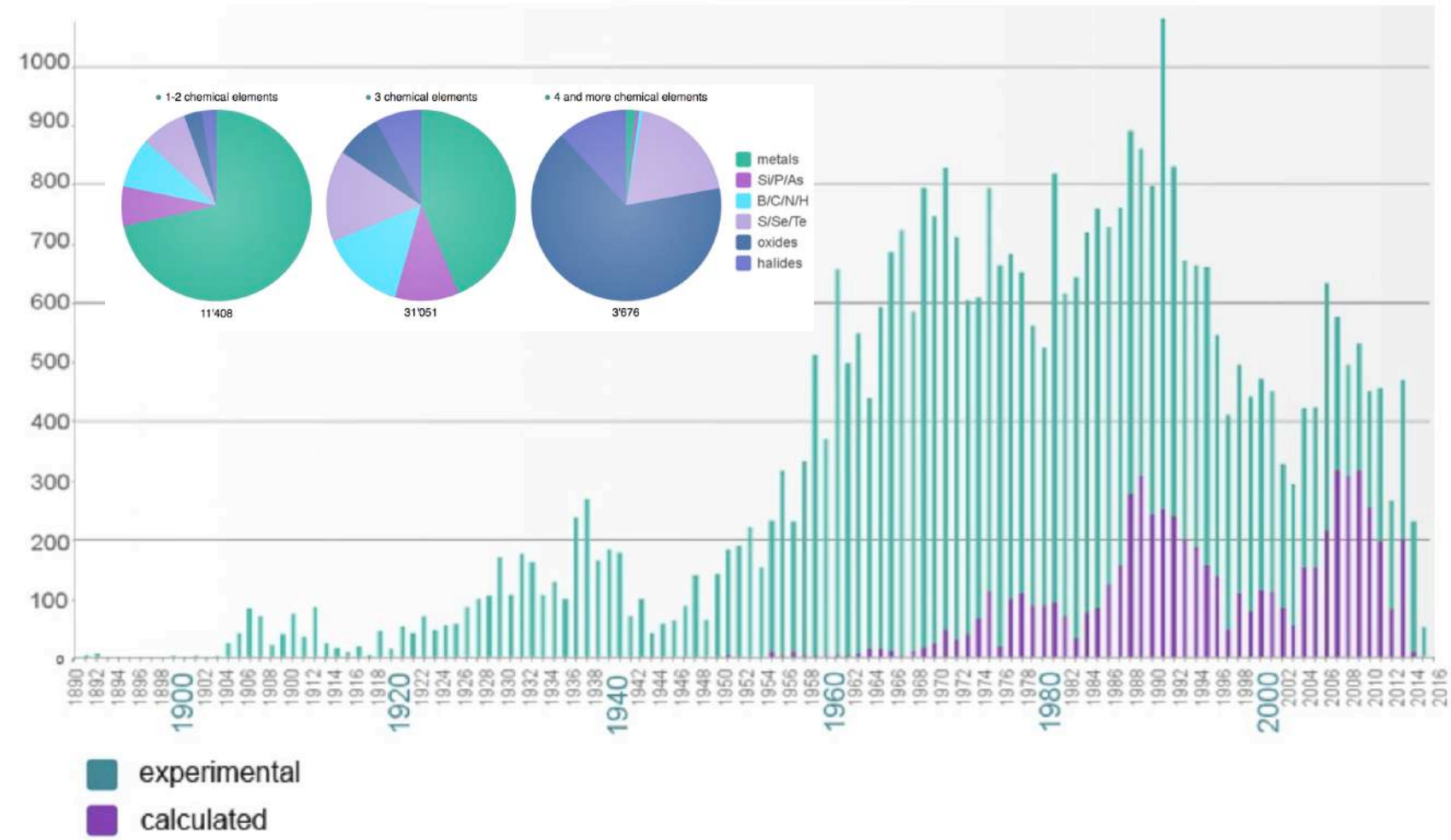
Sincerely,

Linus Pauling

Linus Pauling

http://paulingfile.com/

PAULING FILE: фазовые диаграммы

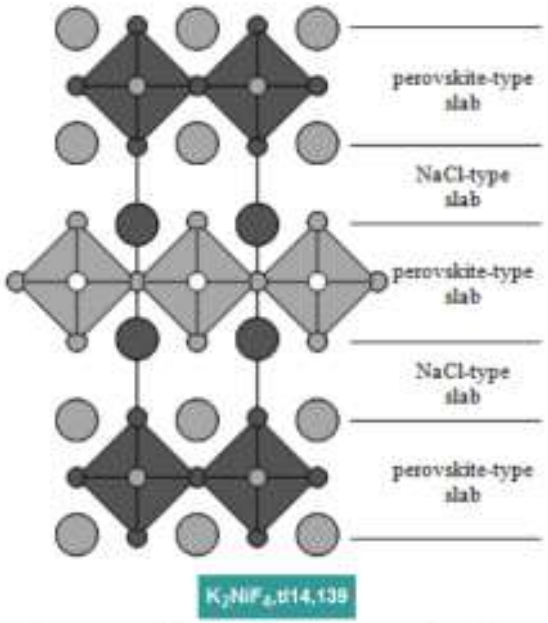


46'135 database entries processed 01.01.2017

PAULING FILE: кристаллические структуры

select a prototype

K2NiF4



Arrangement of NiF_6 octahedra (F atoms small) and K atoms (large) viewed along [100]; slabs containing light and dark octahedra are shifted by a/2

1299 entries for this prototype in the PAULING FILE (2012)

• structural features

NiF_6 octahedra share vertices to form infinite slabs. Intergrowth of perovskite-type slabs (KF-NiF_2-KF) and NaCl-type (rocksalt) slabs (KF-KF).

• prototype names

K_2NiF_4 [1], Ruddlesden-Popper phase n = 1, $SrLaCuO_4$ [2], T phase, cuprate 0201, Ca_4Sb_2O [3]

• crystallographic data [4]

space group (139) *I*4/*mmm*

a = 0.4013, c = 1.3088 nm, c/a = 3.261, V = 0.2108 nm^3, Z = 2

site	WP	sym.	*x*	*y*	*z*	occ.	atomic environment
F1	4e	4*mm*	0	0	0.15331		single atom Ni
K	4e	4*mm*	0	0	0.35377		tricapped trigonal prism F_9
F2	4c	*mmm*	0	1/2	0		colinear Ni_2
Ni	2a	4/*mmm*	0	0	0		octahedron F_6

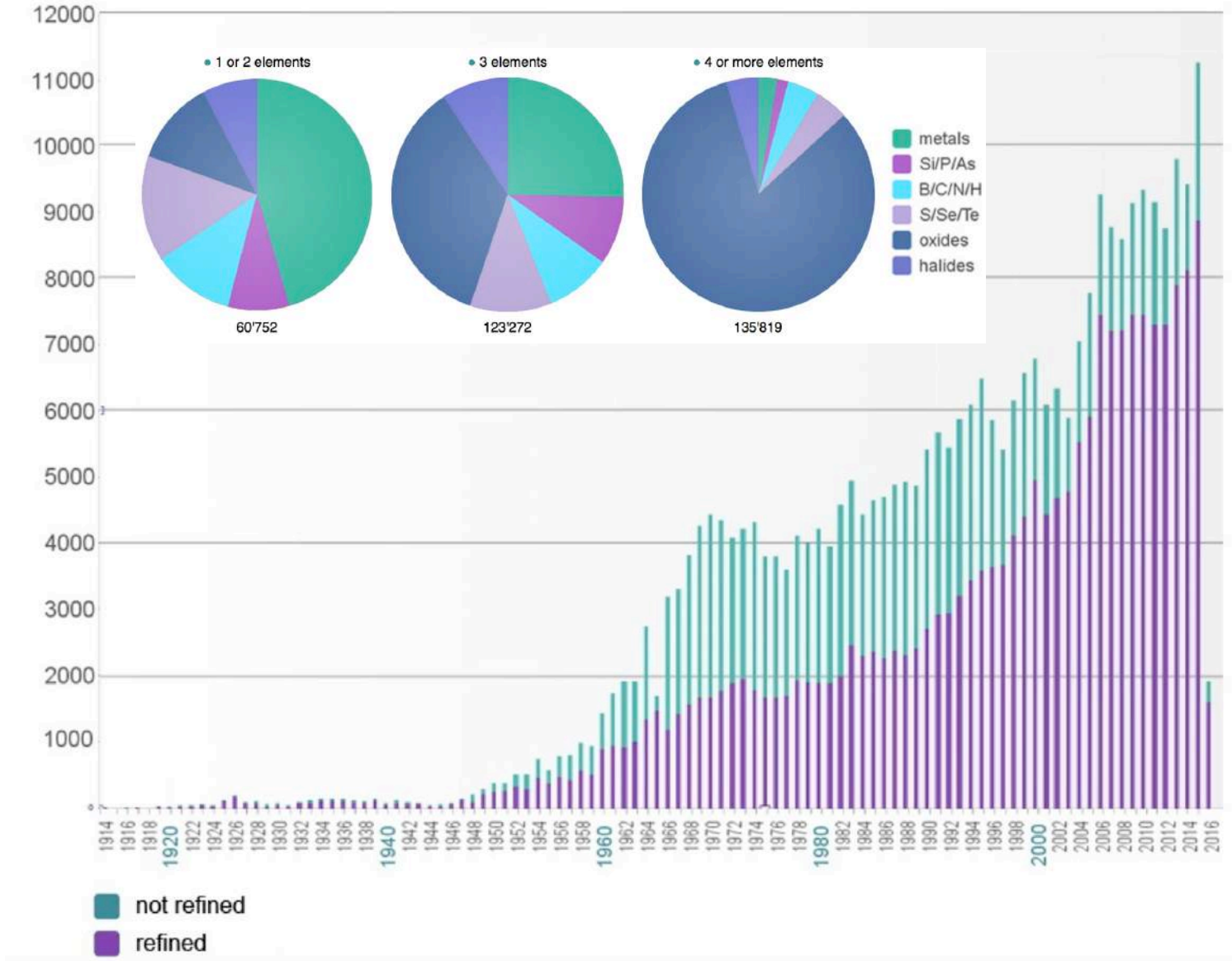
• experimental

single crystal, diffractometer, X-rays, R = 0.018

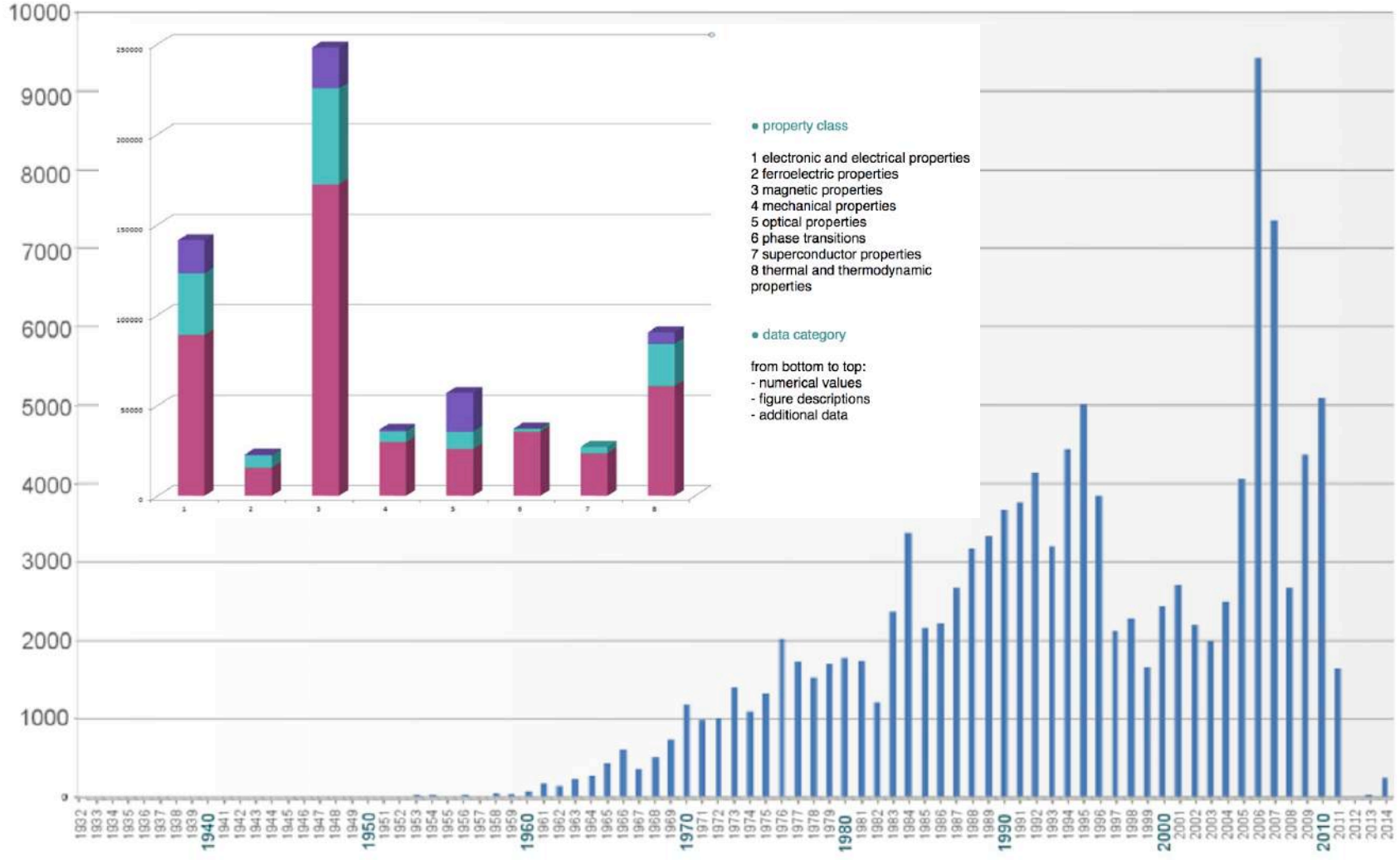
• references

[1] Balz D. (1953), Naturwissenschaften 40, 241
[2] Goodenough J.B., Demazeau G., Pouchard M., Hagenmuller P. (1973), J. Solid State Chem. 8, 325-330
[3] Eisenmann B., Limartha H., Schäfer H., Graf H.A. (1980), Z. Naturforsch. B 35, 1518-1524
[4] Yeh S.K., Wu S.Y., Lee C.S., Wang Y. (1993), Acta Crystallogr. B 49, 806-811

PAULING FILE: кристаллические структуры

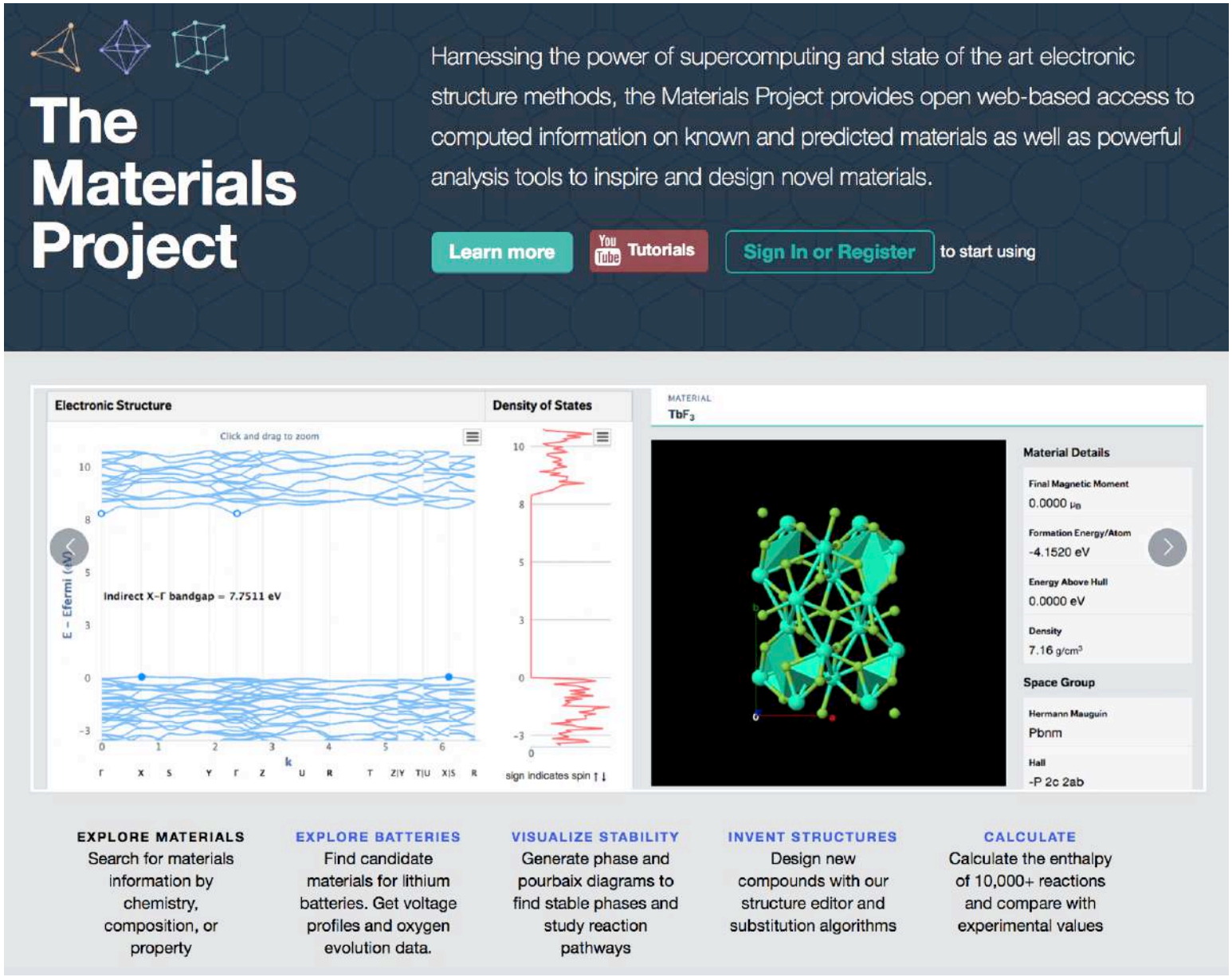


PAULING FILE: физические свойства



125'446 database entries processed 01.01.2017

MATERIALS PROJECT



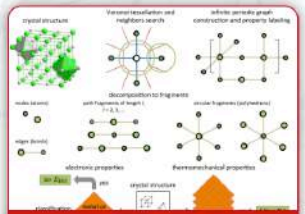
https://materialsproject.org

AFLOW

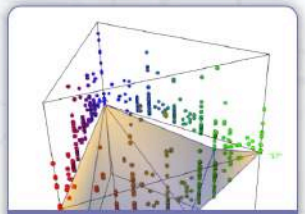
Automatic - FLOW for Materials Discovery

1702812 material compounds - with over 168578388 calculated properties

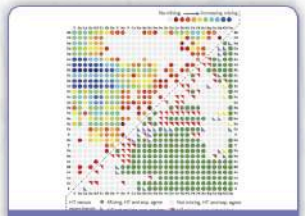
MendeLIB search

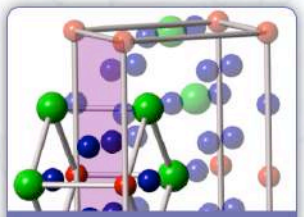

AFLOW-ML

Crystal prototypes


AFLOW hull beta


Aflow-online


Binary alloy library

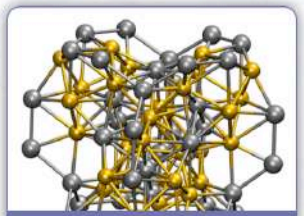

Superalloys search

AFLOW REST-API WIKI

Documentation

AFLOWπ


Geolocation data


Showcase material

Automatic (VASP, Quantum ESPRESSO) calculation of relaxed geometries, electronic and phonon band structures, magnetic properties, and thermodynamic properties

Duke UNIVERSITY · ONR · U.S. DEPARTMENT OF ENERGY · NSF · Homeland Security · CRAY THE SUPERCOMPUTER COMPANY · UNT

http://www.aflowlib.org