UNC
ESHELMAN
SCHOOL OF PHARMACY

MML
UNC.EDU

Applications of machine learning and artificial intelligence to designing chemicals and materials with the desired properties

Alexander Tropsha

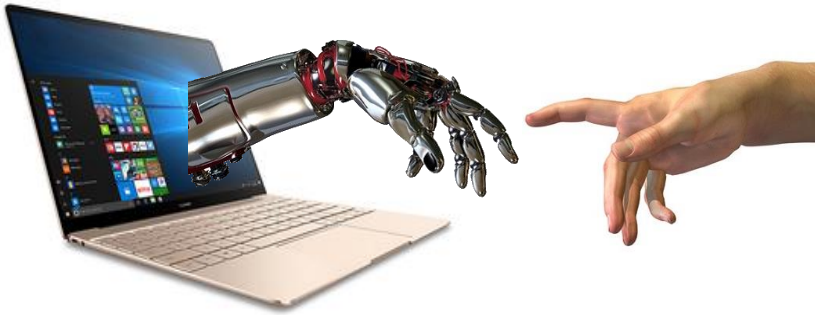
UNC Eshelman School of Pharmacy

Outline

- Brief notes on machine learning/QSAR
- Materials Informatics and Materials Design
- Design, development and application of the Reinforcement Learning for Structural Evolution (ReLeaSE)*
- Summary and future work: QSAR without borders

*Popova, Mariya, Olexandr Isayev, and Alexander Tropsha. "Deep reinforcement learning for de-novo drug design." *arXiv preprint arXiv:1711.10907* (2017); Science Advances, 2018, in press

Machine Learning Framework



$$\mathbf{y} = \mathbf{f}(\mathbf{x})$$

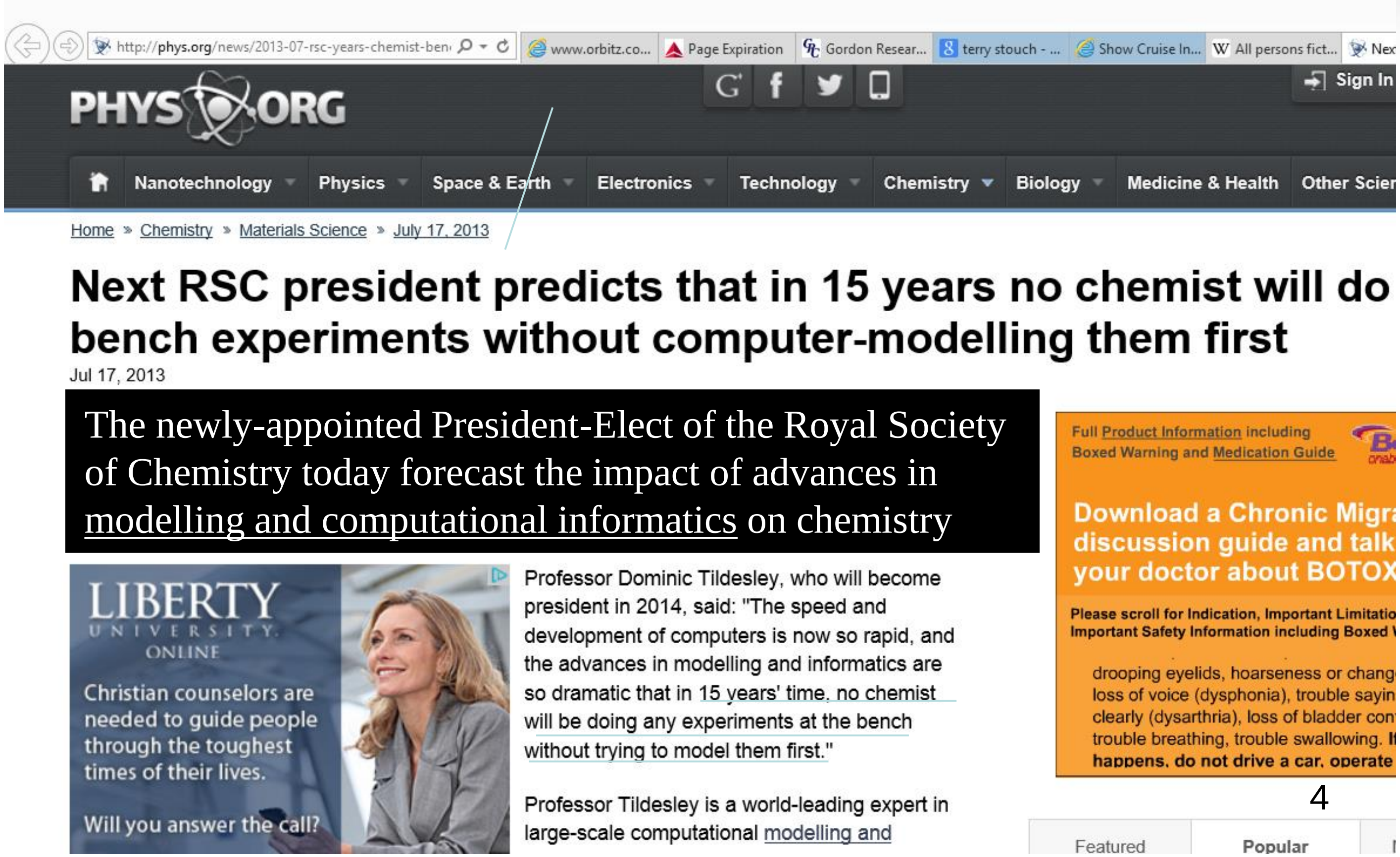
output

prediction function

Molecular features

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

The growing appreciation of molecular modeling and informatics



Home » Chemistry » Materials Science » July 17, 2013

Next RSC president predicts that in 15 years no chemist will do bench experiments without computer-modelling them first

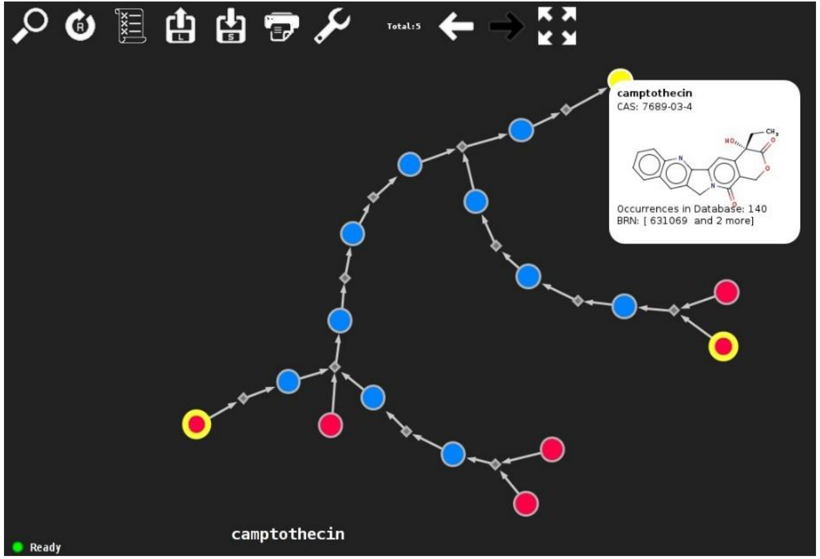
Jul 17, 2013

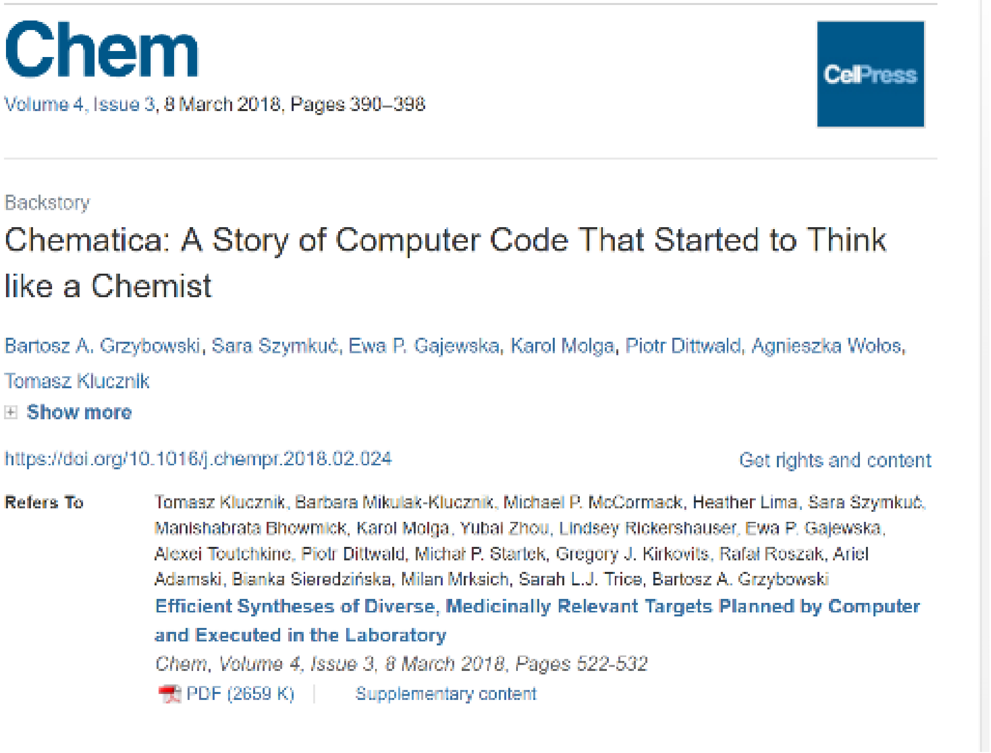
The newly-appointed President-Elect of the Royal Society of Chemistry today forecast the impact of advances in modelling and computational informatics on chemistry

Professor Dominic Tildesley, who will become president in 2014, said: "The speed and development of computers is now so rapid, and the advances in modelling and informatics are so dramatic that in 15 years' time, no chemist will be doing any experiments at the bench without trying to model them first."

Professor Tildesley is a world-leading expert in large-scale computational modelling and

Automated Retrosynthesis (Chematica)





Chem

Volume 4, Issue 3, 8 March 2018, Pages 390–398

CellPress

Backstory

Chematica: A Story of Computer Code That Started to Think like a Chemist

Bartosz A. Grzybowski, Sara Szymkuć, Ewa P. Gajewska, Karol Molga, Piotr Dittwald, Agnieszka Wołos, Tomasz Klucznik

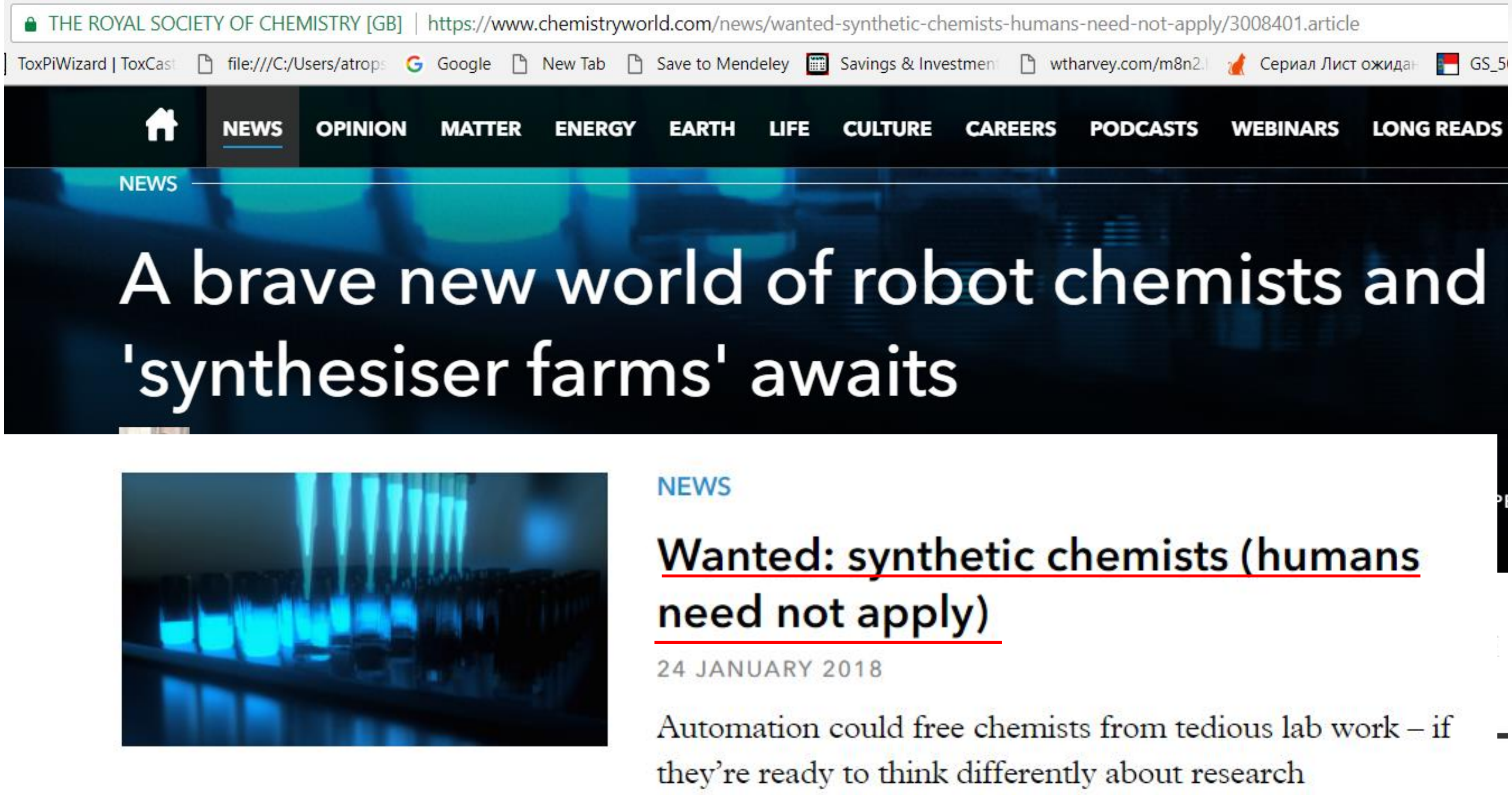
⊞ Show more

https://doi.org/10.1016/j.chempr.2018.02.024

Get rights and content

Refers To Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L.J. Trice, Bartosz A. Grzybowski
Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory
Chem, Volume 4, Issue 3, 8 March 2018, Pages 522-532
PDF (2659 K) | Supplementary content

The growing appreciation of molecular modeling and informatics



NEWS

A brave new world of robot chemists and 'synthesiser farms' awaits

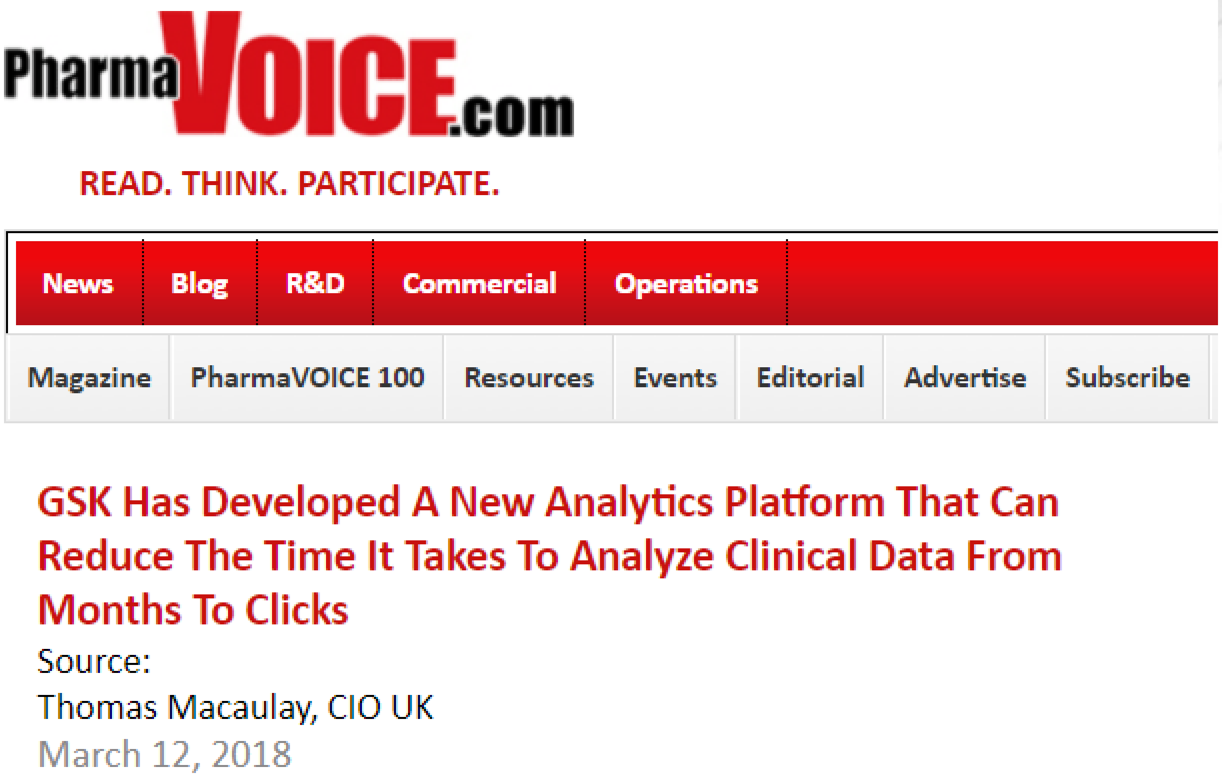
NEWS

Wanted: synthetic chemists (humans need not apply)

24 JANUARY 2018

Automation could free chemists from tedious lab work – if they're ready to think differently about research

Promise of dramatic acceleration of drug discovery

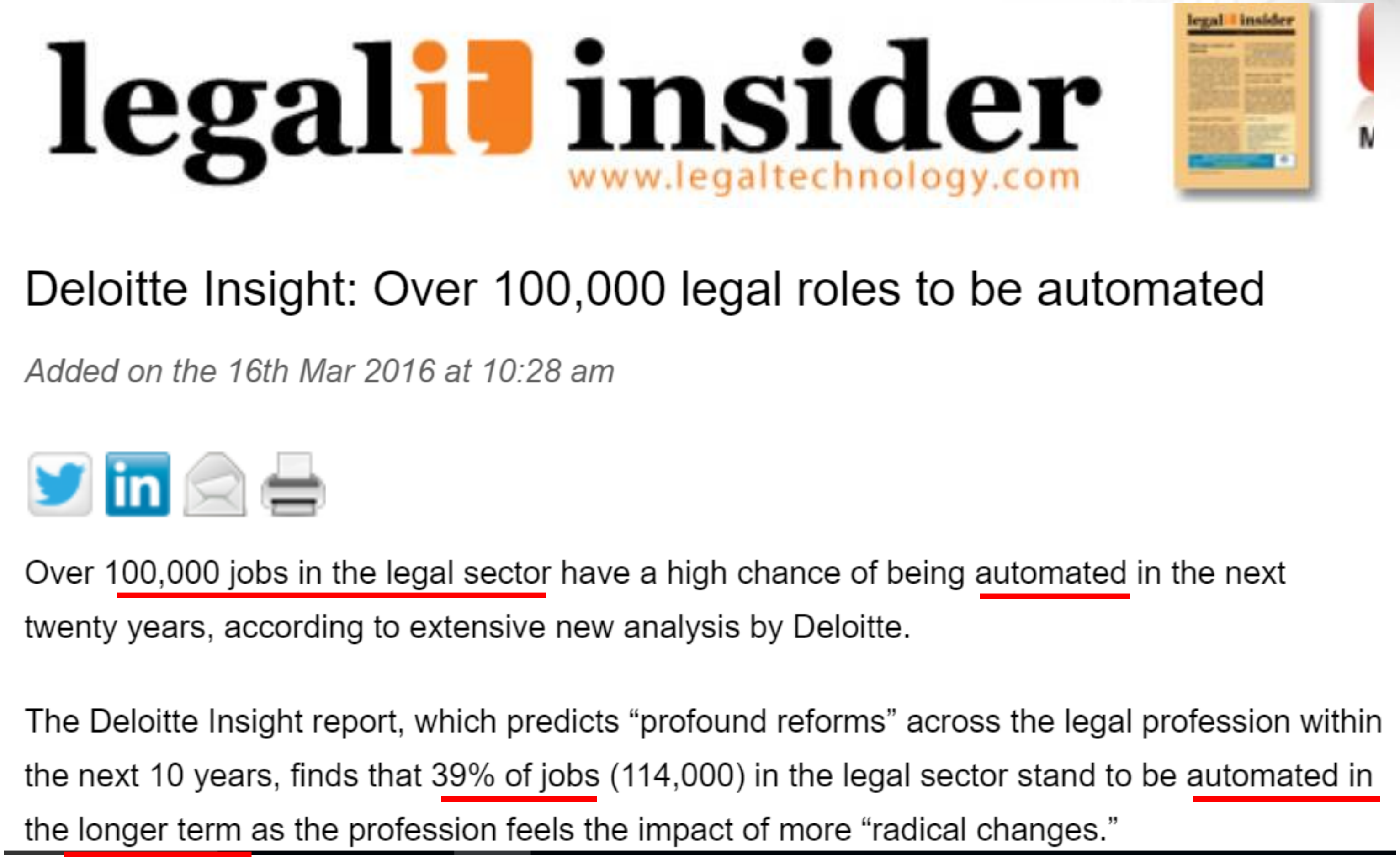


GSK Has Developed A New Analytics Platform That Can Reduce The Time It Takes To Analyze Clinical Data From Months To Clicks

Source:
Thomas Macaulay, CIO UK
March 12, 2018

The platform uses large-scale data analytics to drive better decisions about the drug discovery pipeline, by allowing the pharmaceuticals giant to test the potential for new drugs before it begins clinical trials.

Rise of the machines in legal industry



legal it insider
www.legaltechnology.com

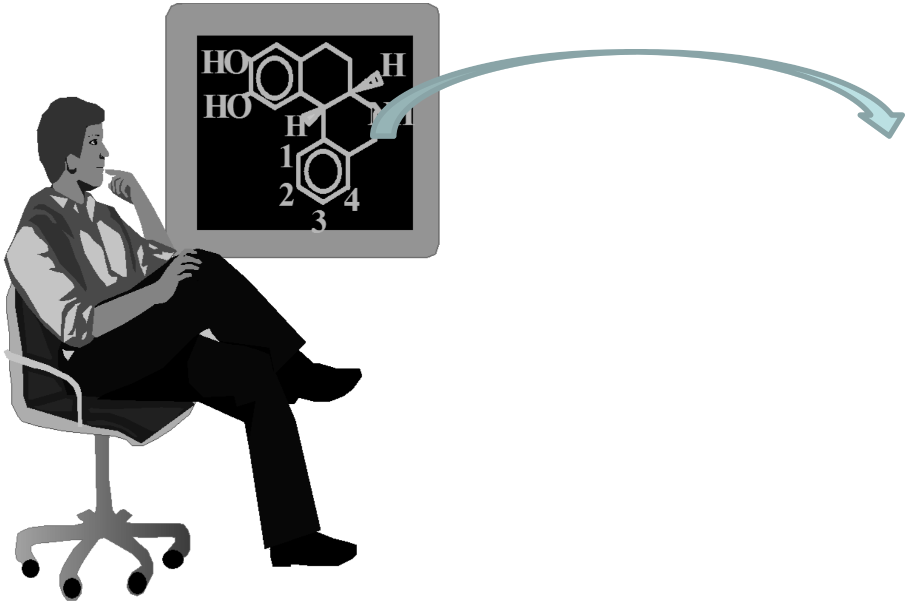
Deloitte Insight: Over 100,000 legal roles to be automated

Added on the 16th Mar 2016 at 10:28 am

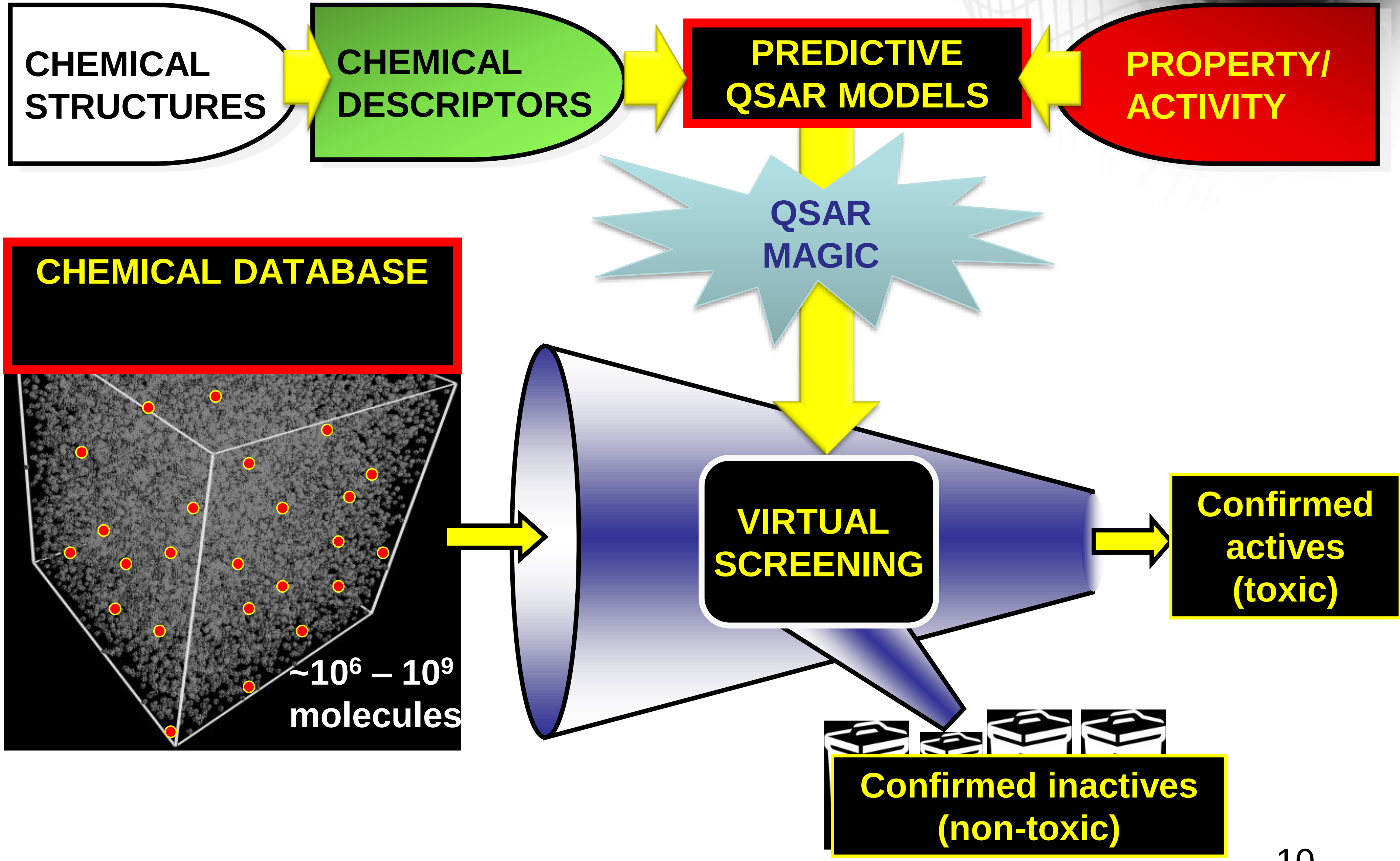
Over 100,000 jobs in the legal sector have a high chance of being automated in the next twenty years, according to extensive new analysis by Deloitte.

The Deloitte Insight report, which predicts “profound reforms” across the legal profession within the next 10 years, finds that 39% of jobs (114,000) in the legal sector stand to be automated in the longer term as the profession feels the impact of more “radical changes.”

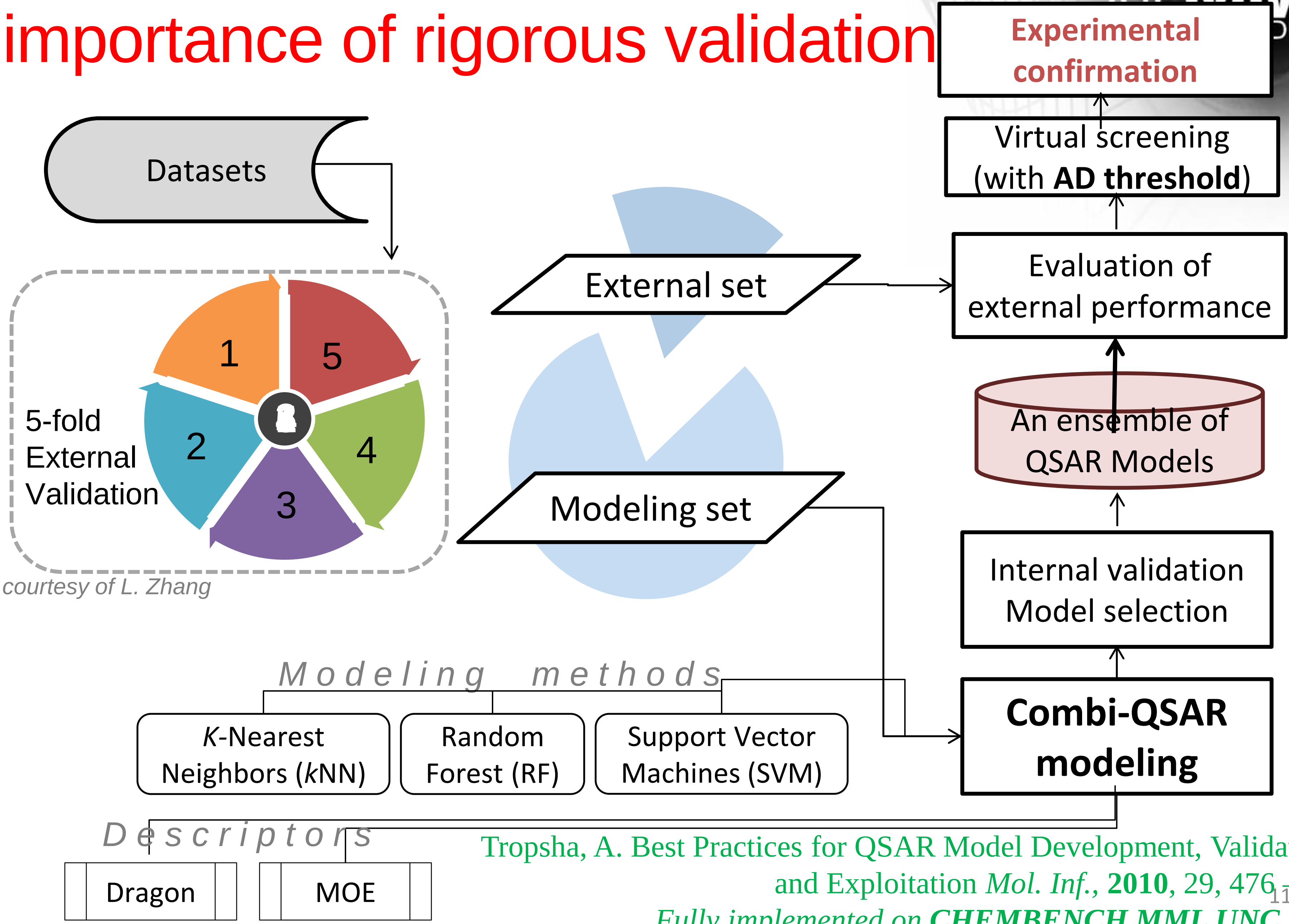
The ultimate dream of a computational chemist



The chief utility of computational models: Annotation of new compounds



QSAR Modeling Workflow: the importance of rigorous validation



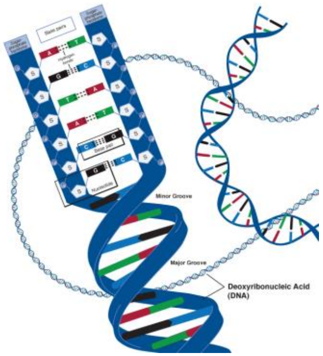
Experimental confirmation
Virtual screening (with **AD threshold**)
Datasets
Evaluation of external performance
External set
1
5
2
4
3
5-fold External Validation
An ensemble of QSAR Models
Modeling set
Internal validation Model selection
courtesy of L. Zhang
Modeling methods
Combi-QSAR modeling
K-Nearest Neighbors (*k*NN)
Random Forest (RF)
Support Vector Machines (SVM)
Descriptors
Dragon
MOE

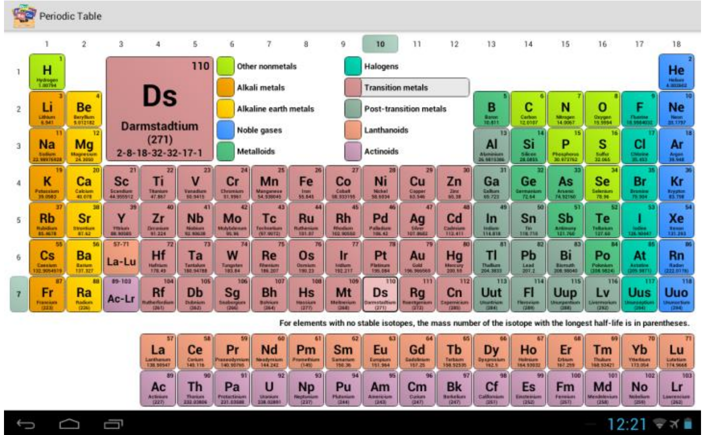
Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation *Mol. Inf.*, **2010**, 29, 476 – 488

*Fully implemented on **CHEMBENCH.MML.UNC.EDU***

Material Science and the Rise of Materials Informatics

- Explosive growth of materials data, both experimental databases and computational repositories.
  - <u>Structural data</u>: 160,000 entries in the Inorganic Crystal Structure Database (ICSD)
  - <u>Experimental data</u>: Numerous commercial and open experimental databases NIST, MatWeb, MatBase etc.
  - <u>Computational data</u>: Huge databases such as AFLOWLIB, Materials Project, and Harvard Clean Energy
  - Chemical space of possible materials is HUGE ~10^{100} candidates [*Nat. Chem*. 7, 274-275 (**2015**)]
- Materials Genome Initiative or MGI (US Govt): Need for new high performance materials





110 Ds Darmstadtium (271) 2-8-18-32-32-17-1

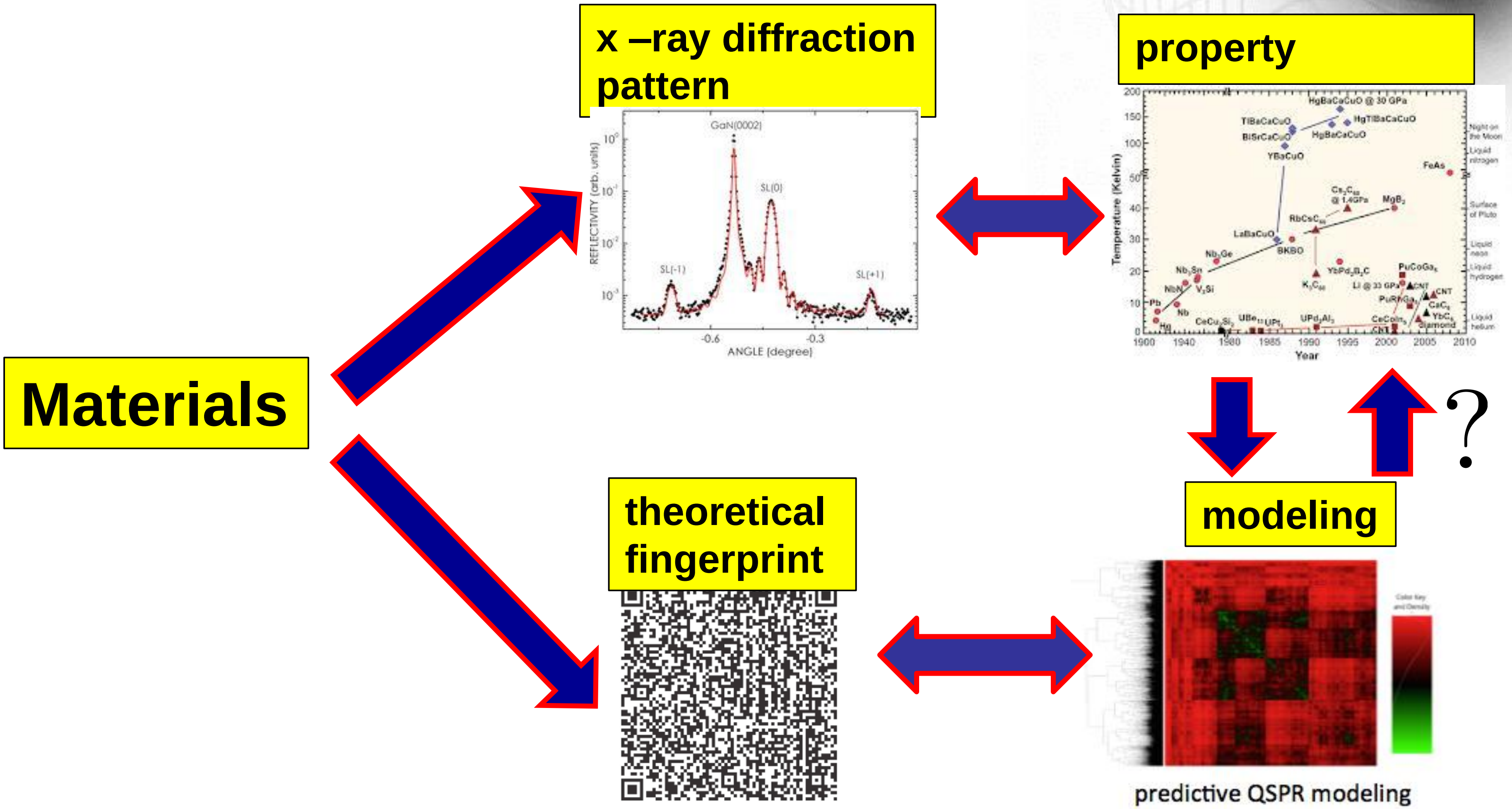
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57-71 La-Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89-103 Ac-Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo

Other nonmetals
Alkali metals
Alkaline earth metals
Noble gases
Metalloids
Halogens
Transition metals
Post-transition metals
Lanthanoids
Actinoids

For elements with no stable isotopes, the mass number of the isotope with the longest half-life is in parentheses.

57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Closing the gap: materials structure-property relationships



Materials

x –ray diffraction pattern

property

?

modeling

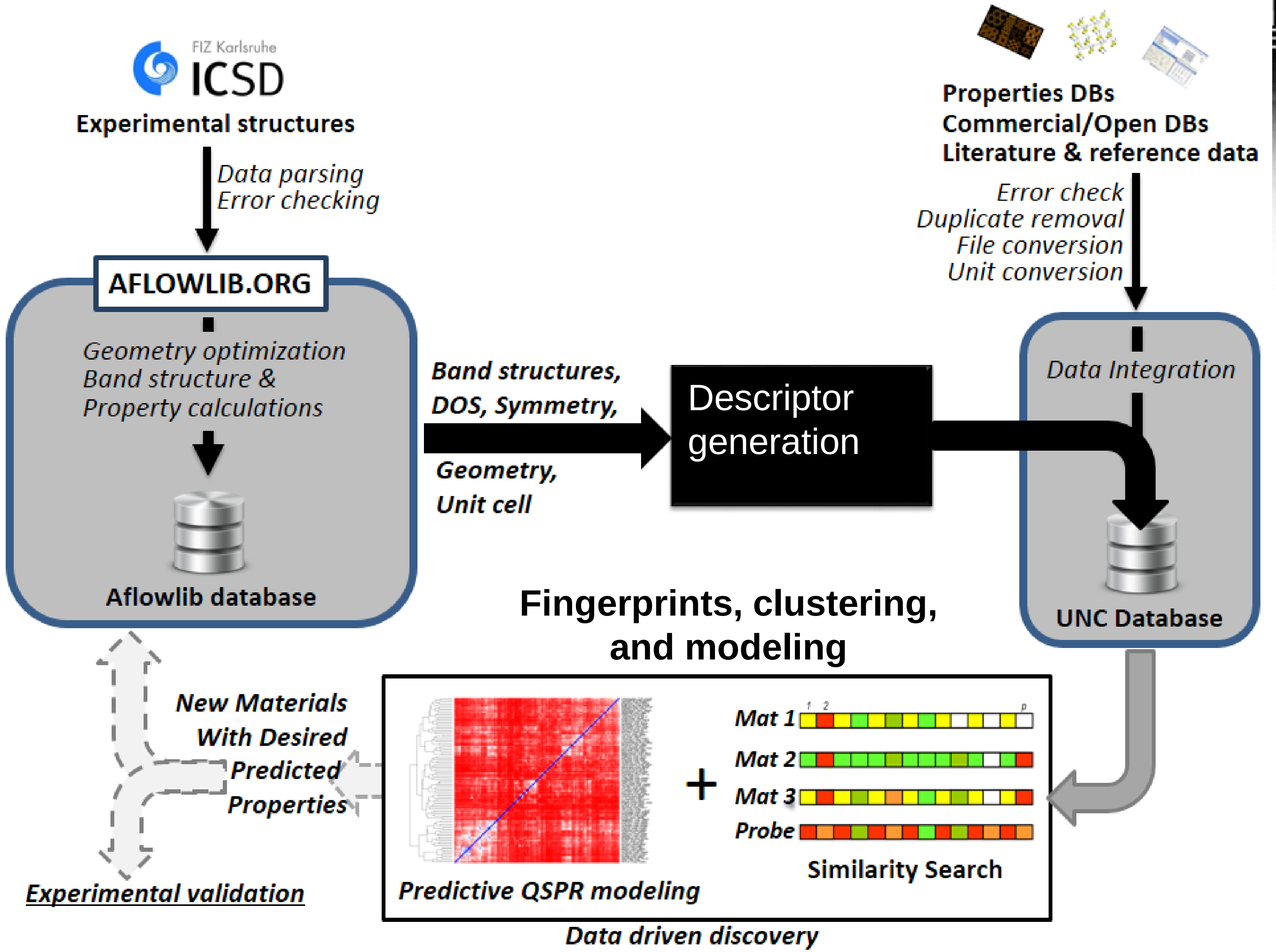
theoretical fingerprint

predictive QSPR modeling

Quantitative Structure Activity Relationship approaches (QSAR)
Quantitative Structure-Property Relationships (QMSPR)

Isayev, Fourches, Muratov, Oses, Rasch, Tropsha, Curtarolo, Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. Chem. Mater., **2015**, 27: 735–743

Material Informatics/MQSAR Workflow



FIZ Karlsruhe ICSD

Experimental structures

Data parsing
Error checking

AFLOWLIB.ORG

Geometry optimization
Band structure &
Property calculations

Aflowlib database

Band structures,
DOS, Symmetry,

Geometry,
Unit cell

Descriptor generation

Properties DBs
Commercial/Open DBs
Literature & reference data

Error check
Duplicate removal
File conversion
Unit conversion

Data Integration

UNC Database

Fingerprints, clustering, and modeling

New Materials
With Desired
Predicted
Properties

Experimental validation

Predictive QSPR modeling

+

Mat 1
Mat 2
Mat 3
Probe

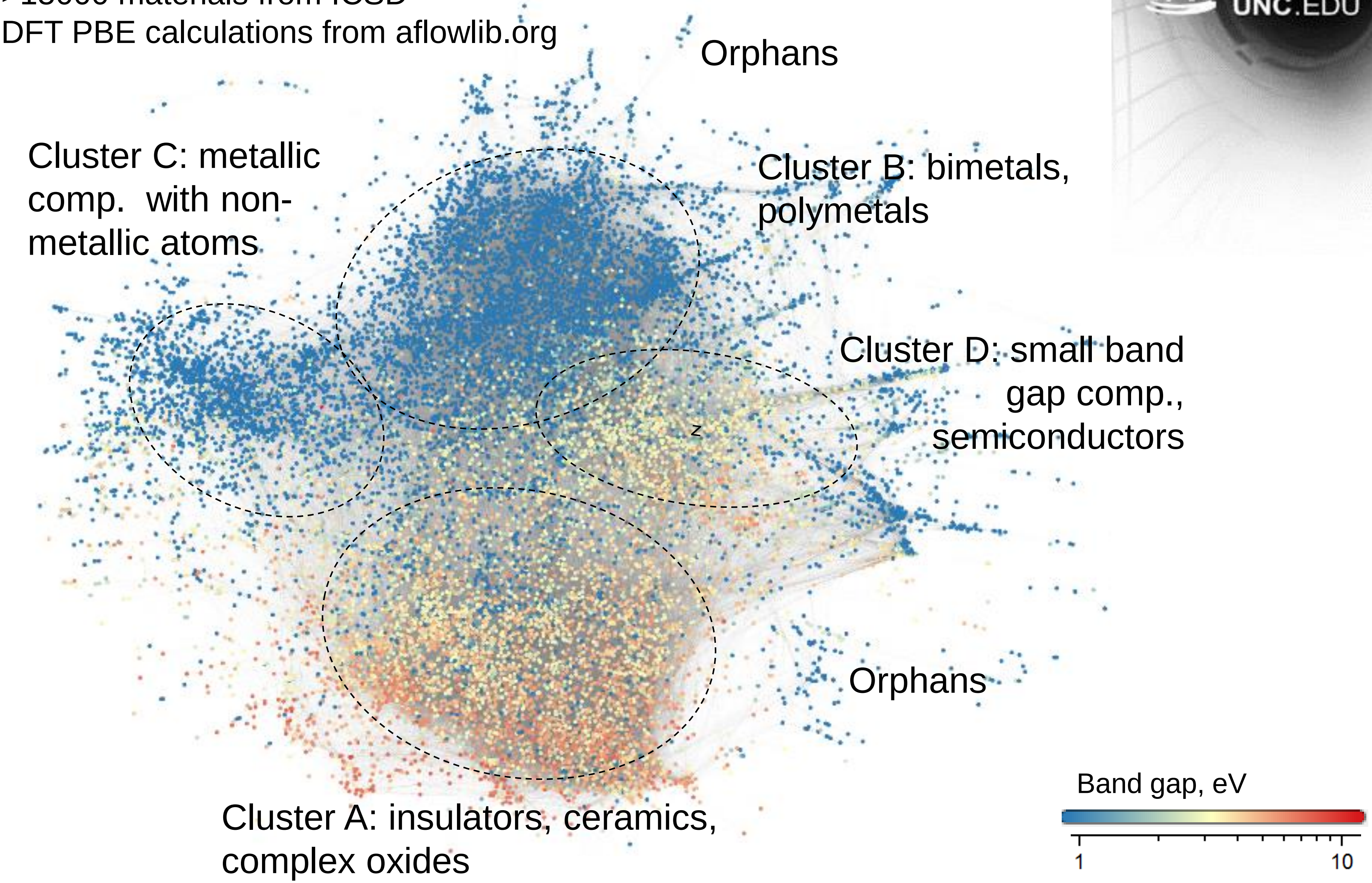
Similarity Search

Data driven discovery

Isayev, Fourches, Muratov, Oses, Rasch, Tropsha, Curtarolo, Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. Chem. Mater., **2015**, 27: 735–743

Material Map (B-Fingerprints)

>15000 materials from ICSD
DFT PBE calculations from aflowlib.org



Orphans

Cluster C: metallic comp. with non-metallic atoms

Cluster B: bimetals, polymetals

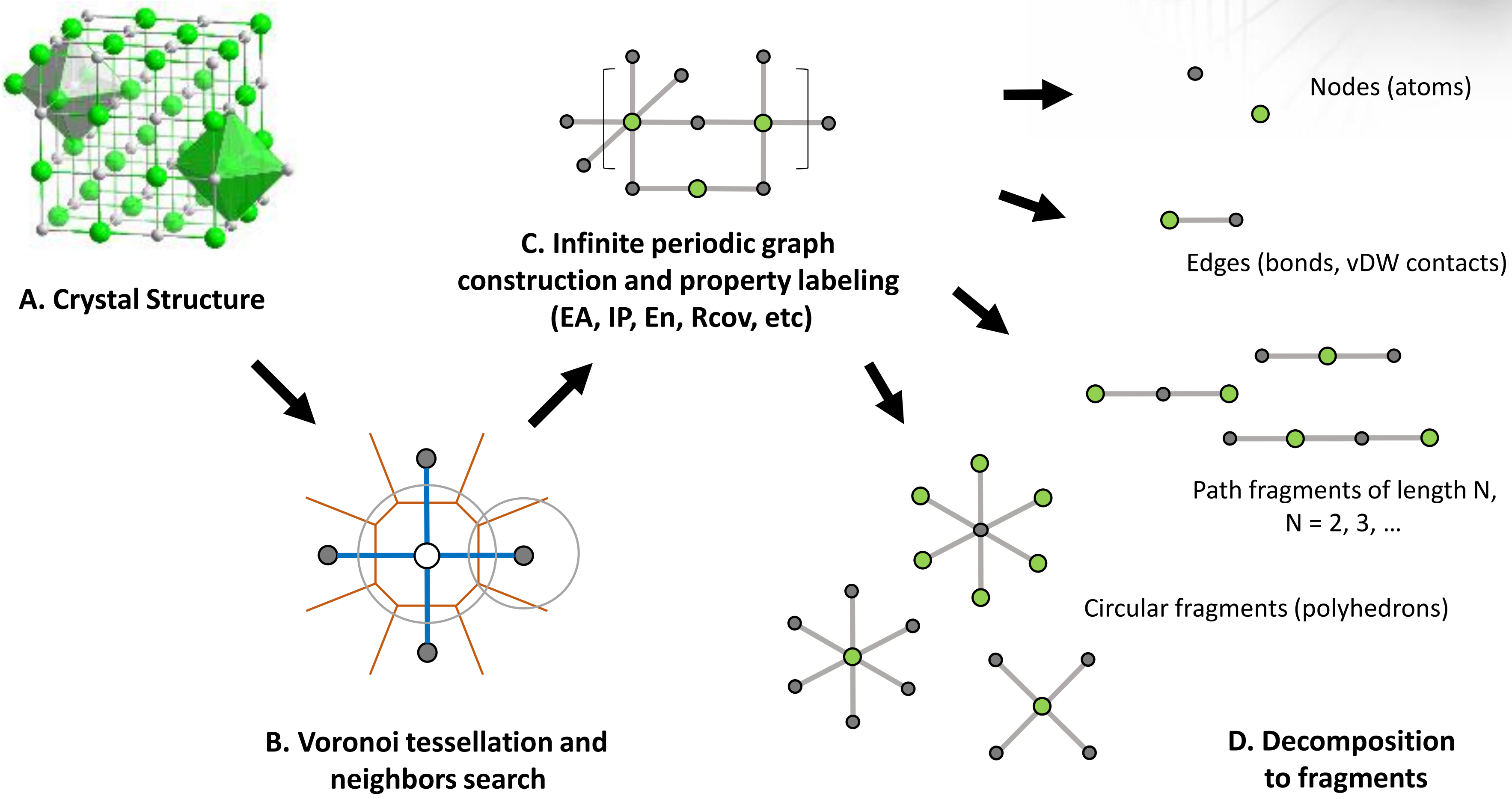
Cluster D: small band gap comp., semiconductors

Orphans

Cluster A: insulators, ceramics, complex oxides

Isayev, Fourches, Muratov, Oses, Rasch, Tropsha, Curtarolo, Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. Chem. Mater., **2015**, 27: 735–743

Systematic representation of materials using fragment descriptors



A. Crystal Structure

B. Voronoi tessellation and neighbors search

C. Infinite periodic graph construction and property labeling (EA, IP, En, Rcov, etc)

Nodes (atoms)

Edges (bonds, vDW contacts)

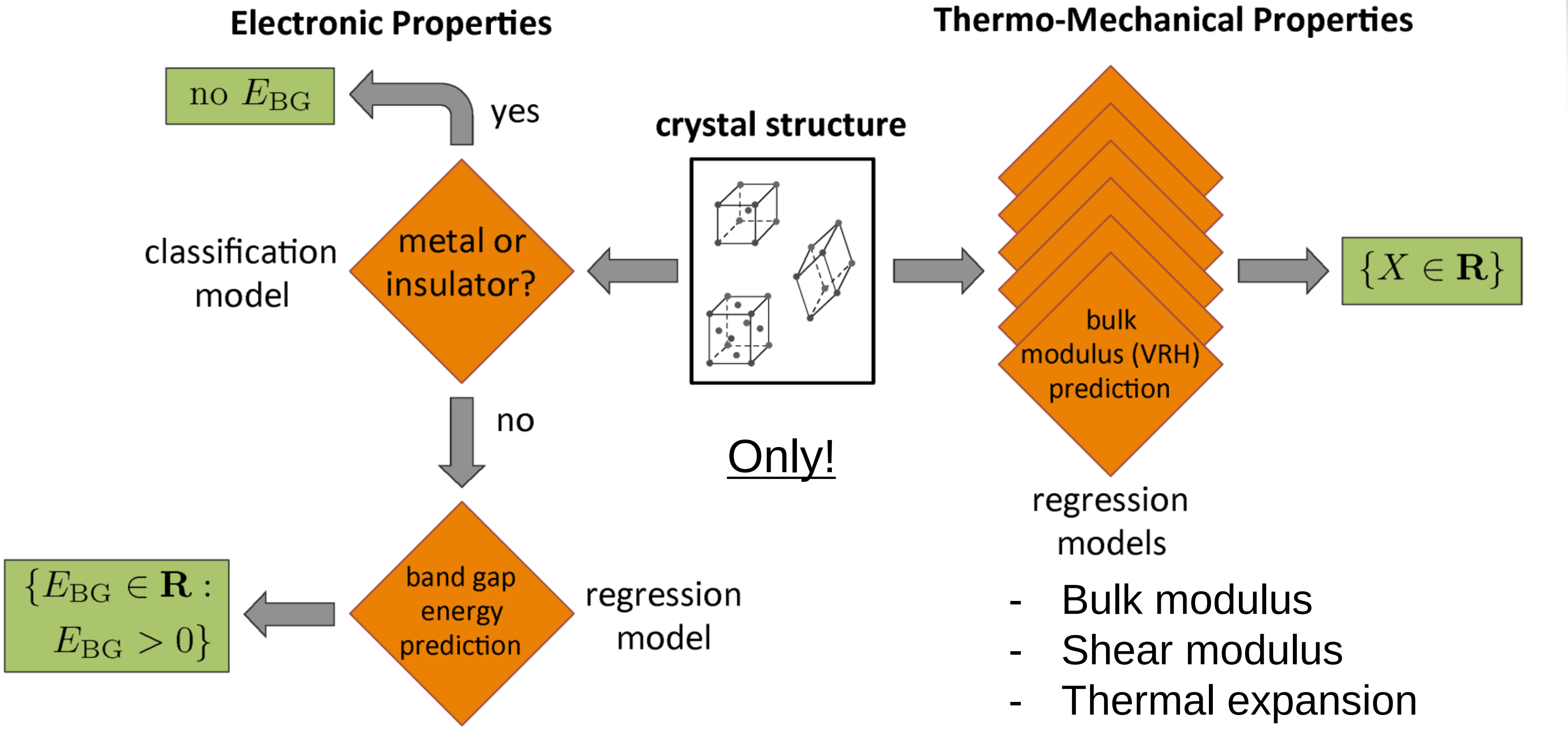
Path fragments of length N, N = 2, 3, …

Circular fragments (polyhedrons)

D. Decomposition to fragments

Isayev et al. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature Comm*, **2017**, 8, 15679.

ML Workflow for Materials Property Prediction



- Bulk modulus
- Shear modulus
- Thermal expansion
- Heat Capacity

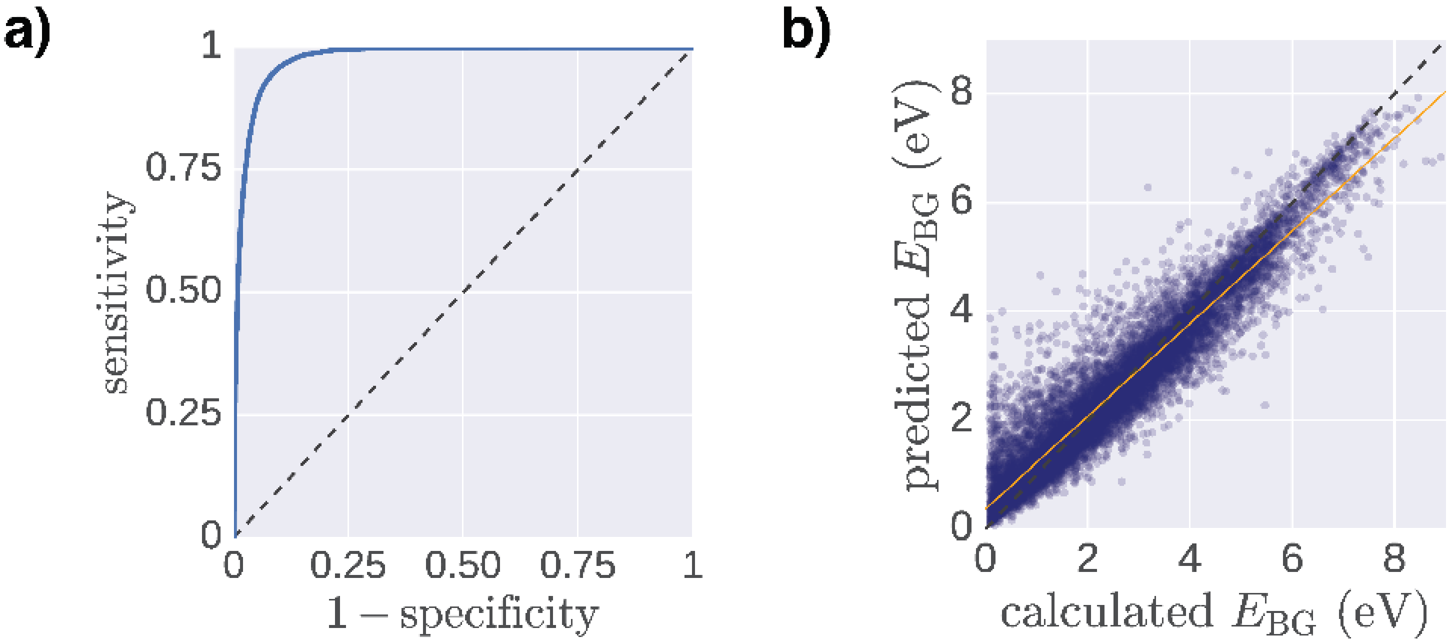
Thermal conductivity, etc.

All models are trained based on DFT-computed properties (VASP s/w from U. Vienna)

Isayev et al. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature Comm*, **2017**, 8, 15679.

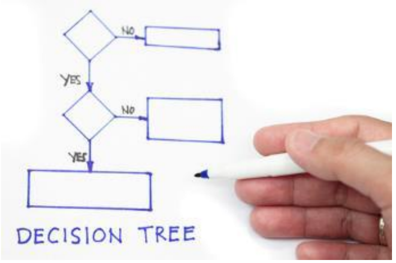
Prediction of Electronic Properties

electronic properties



Classification accuracy 95%
ROC Curve (AUC) 0.98

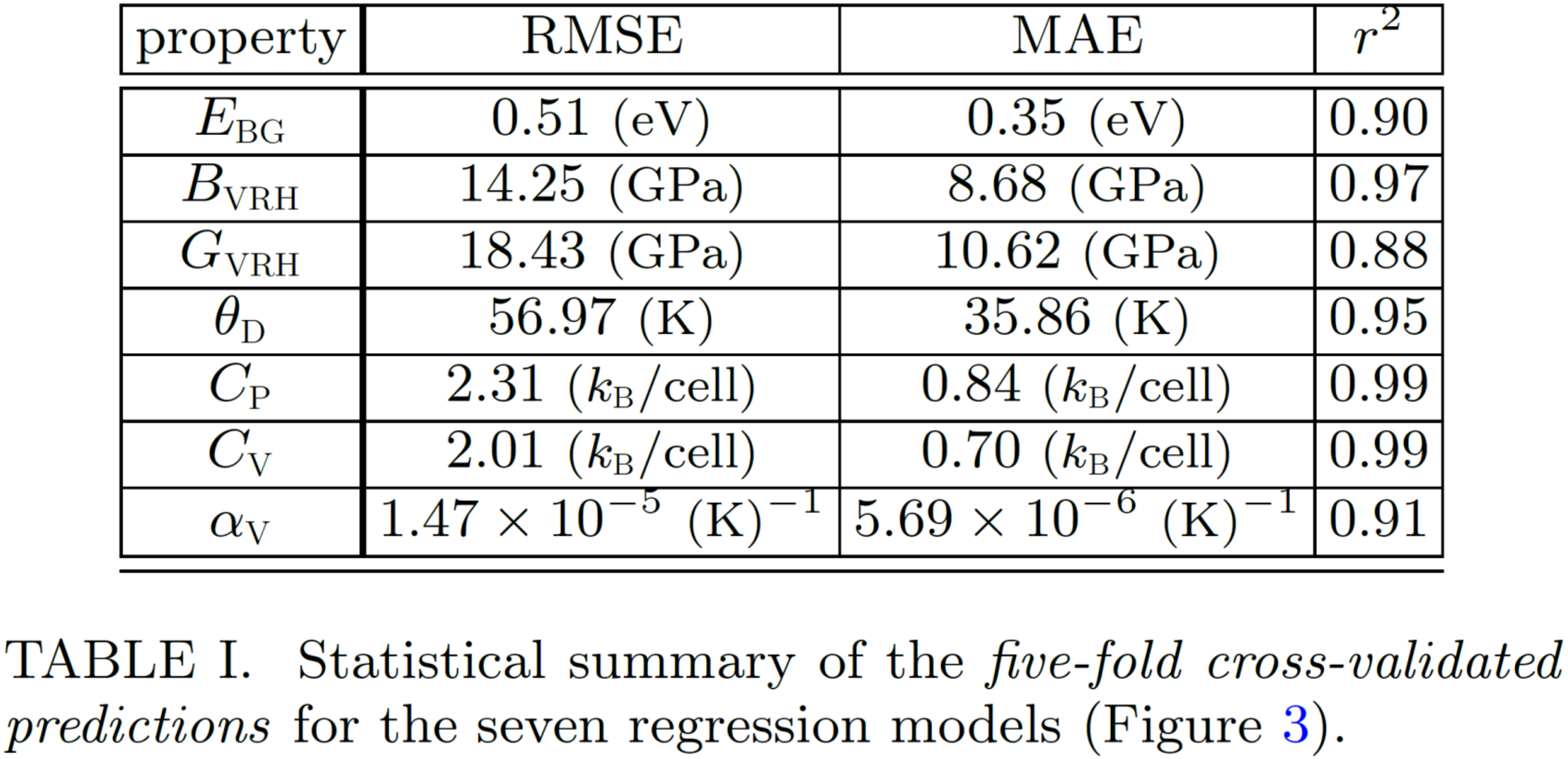
Learning approach for all models:
Gradient Boosting Decision Trees (GBT)



Isayev et al. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature Comm*, **2017**, 8, 15679.

Prediction of Thermomechanical Properties

(E_{BG} - band gap energy; B_{VRH}- bulk modulus; G_{VRH} -shear modulus; t_D - Debye temperature; C_P - heat capacity at constant pressure; C_V - heat capacity at constant volume; a_V -thermal expansion coefficient



property	RMSE	MAE	r^2
E_{BG}	0.51 (eV)	0.35 (eV)	0.90
B_{VRH}	14.25 (GPa)	8.68 (GPa)	0.97
G_{VRH}	18.43 (GPa)	10.62 (GPa)	0.88
θ_{D}	56.97 (K)	35.86 (K)	0.95
C_{P}	2.31 (k_{B}/cell)	0.84 (k_{B}/cell)	0.99
C_{V}	2.01 (k_{B}/cell)	0.70 (k_{B}/cell)	0.99
α_{V}	1.47×10^{-5} $(\mathrm{K})^{-1}$	5.69×10^{-6} $(\mathrm{K})^{-1}$	0.91

TABLE I. Statistical summary of the *five-fold cross-validated predictions* for the seven regression models (Figure 3).

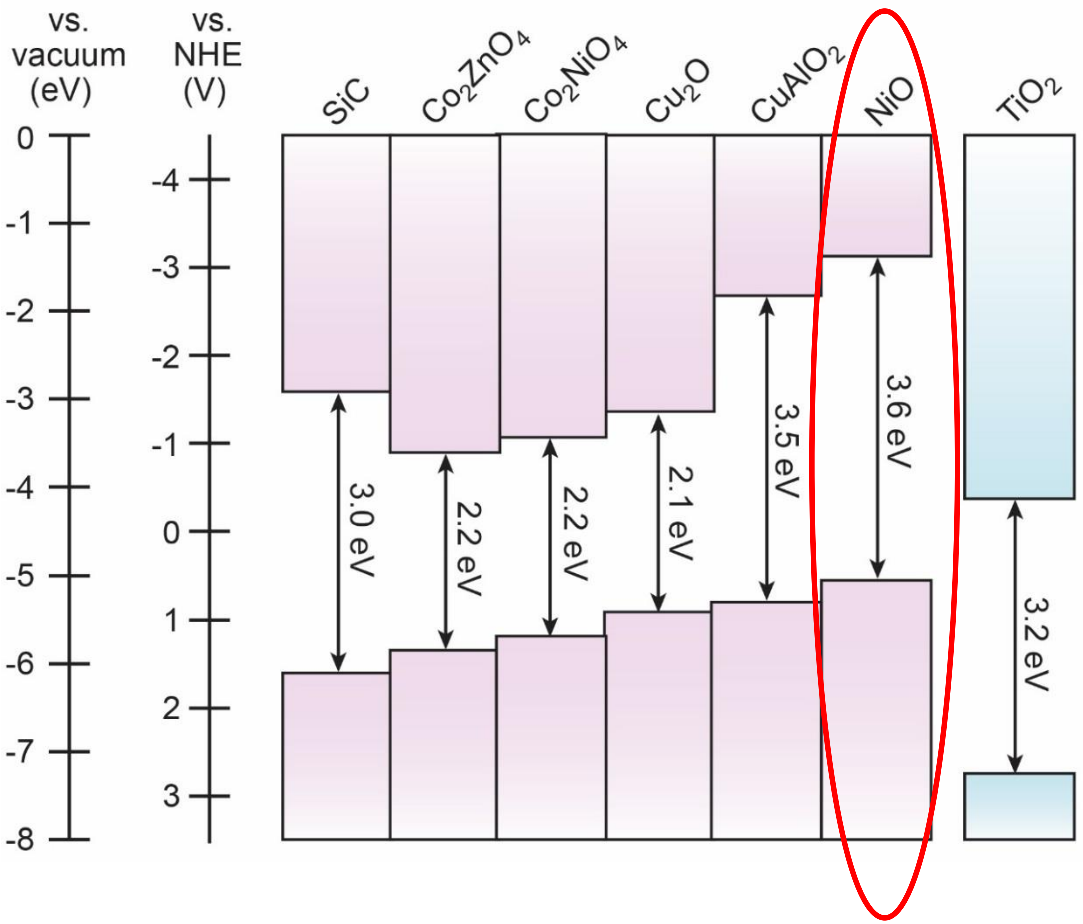
Isayev et al. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nature Comm*, **2017**, 8, 15679.

Summary of Materials Informatics: Methods

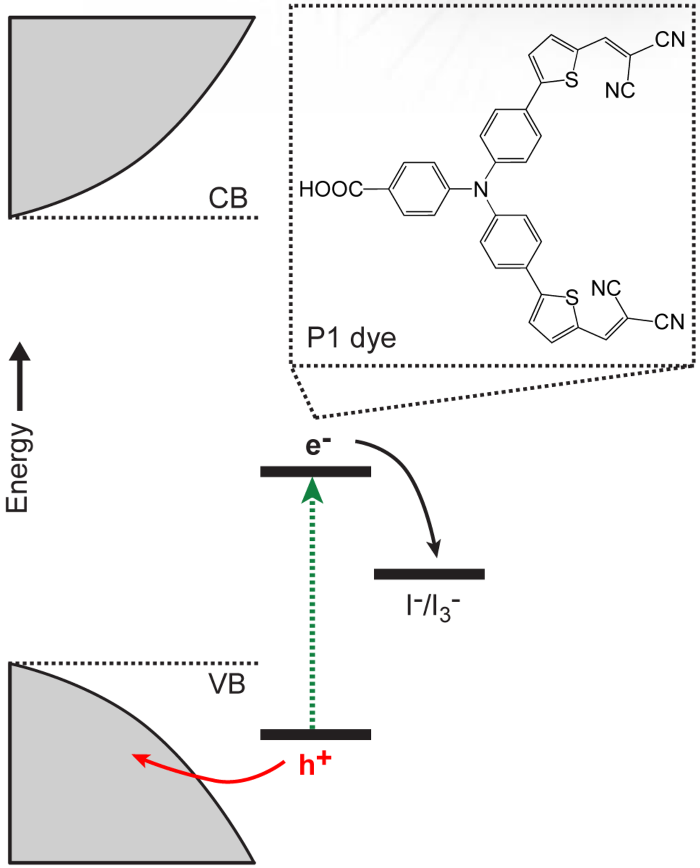
- Fast, accurate general purpose machine learning methods for material's property prediction. Milliseconds on laptop vs. days on HPC cluster
- Universal applicability to different materials: currently covered 85 elements (H – Pu, without noble gases, Tc, Fr, Ra). All types of crystal lattices and symmetries.
  - Most competing approaches are specific to one prototype/family of materials or single property
- Works for other properties: elastic, thermoelectric, etc.
- Possible to gain *some* chemically/physically interpretable insight into "black box" model.
- Possible to derive materials design rules
- User friendly web app and RESTful API (http://aflow.org/aflow-ml/)

Photocathode materials

Evaluated as DSSCs

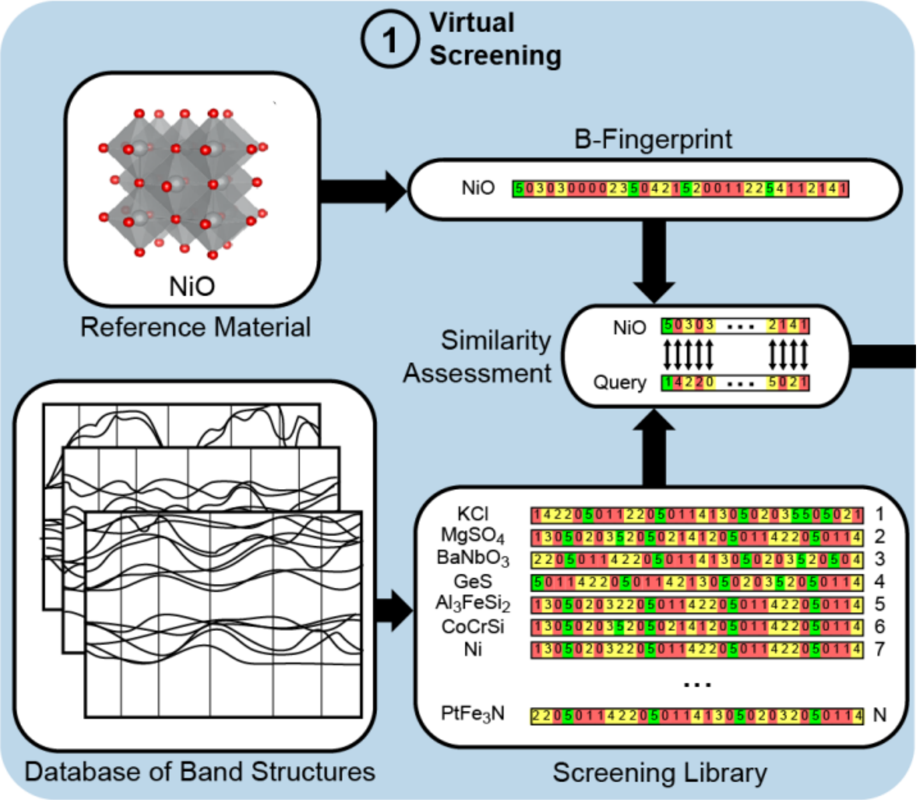


Dye-sensitized solar cells (DSSCs)



Design of alternate photocathodes

A materials informatics approach

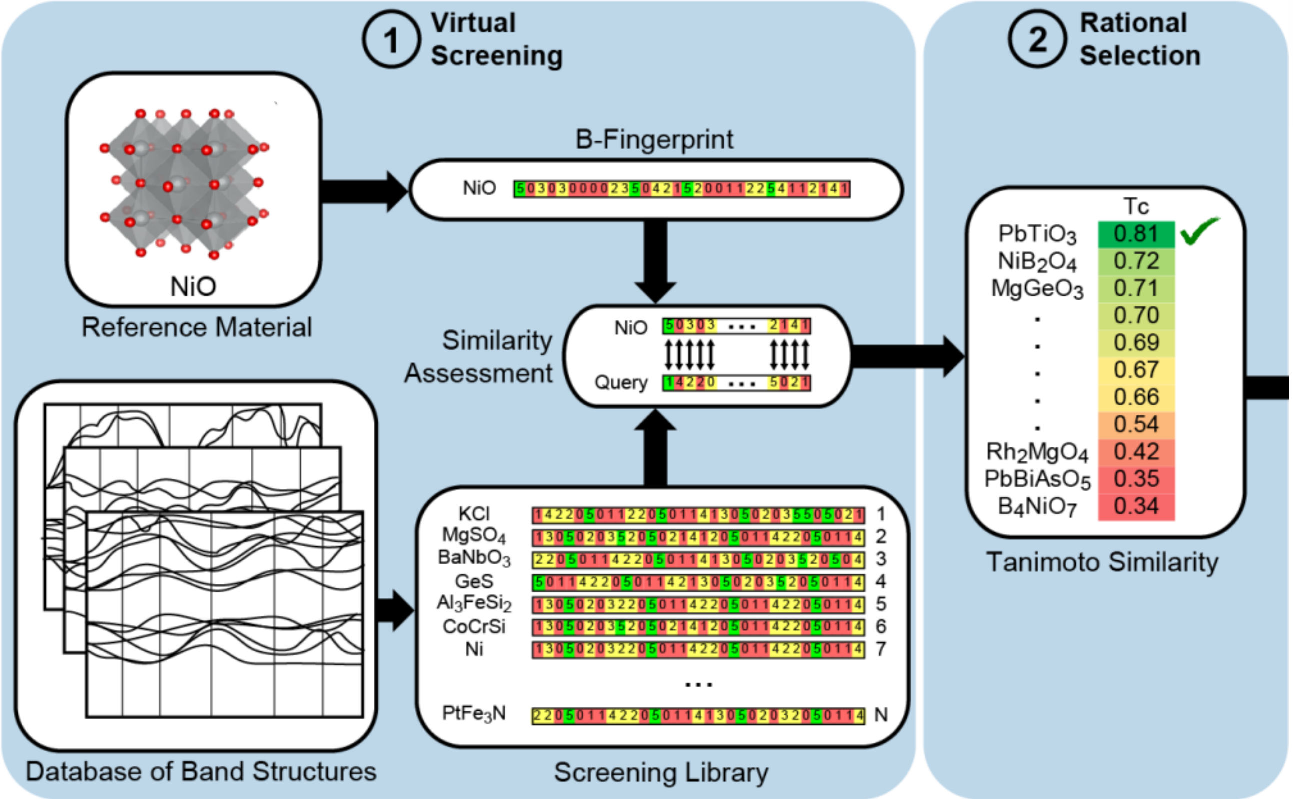

1
Virtual Screening
NiO
Reference Material
B-Fingerprint
NiO 5 0 3 0 3 0 0 0 0 2 3 5 0 4 2 1 5 2 0 0 1 1 2 2 5 4 1 1 2 1 4 1
Similarity Assessment
NiO 5 0 3 0 3 ... 2 1 4 1
Query 1 4 2 2 0 ... 5 0 2 1
Database of Band Structures
KCl 1 4 2 2 0 5 0 1 1 2 2 0 5 0 1 1 4 1 3 0 5 0 2 0 3 5 5 0 5 0 2 1 1
MgSO4 1 3 0 5 0 2 0 3 5 2 0 5 0 2 1 4 1 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 2
BaNbO3 2 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 1 3 0 5 0 2 0 3 5 2 0 5 0 4 3
GeS 5 0 1 1 4 2 2 0 5 0 1 1 4 2 1 3 0 5 0 2 0 3 5 2 0 5 0 1 1 4 4
Al3FeSi2 1 3 0 5 0 2 0 3 2 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 5
CoCrSi 1 3 0 5 0 2 0 3 5 2 0 5 0 2 1 4 1 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 6
Ni 1 3 0 5 0 2 0 3 2 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 7
...
PtFe3N 2 2 0 5 0 1 1 4 2 2 0 5 0 1 1 4 1 3 0 5 0 2 0 3 2 0 5 0 1 1 4 N
Screening Library

(AFLOWLIB)

Moot, Isayev, Tropsha, Cahoon, Materials Discovery, **2016**, 6, 9-16

Design of alternate photocathodes

A materials informatics approach



1 Virtual Screening
2 Rational Selection
NiO
Reference Material
B-Fingerprint
NiO 50303000023504215200112254112141
Similarity Assessment
NiO 50303 ... 2141
Query 14220 ... 5021
Database of Band Structures
KCl 14220501122050114130502035505021 1
$MgSO_4$ 13050203520502141205011422050114 2
$BaNbO_3$ 220501142205011413050203520504 3
GeS 501142205011421305020352050114 4
Al_3FeSi_2 13050203220501142205011422050114 5
CoCrSi 13050203520502141205011422050114 6
Ni 13050203220501142205011422050114 7
...
$PtFe_3N$ 2205011422050114130502032050114 N
Screening Library
Tc
$PbTiO_3$ 0.81 ✓
NiB_2O_4 0.72
$MgGeO_3$ 0.71
0.70
0.69
0.67
0.66
0.54
Rh_2MgO_4 0.42
$PbBiAsO_5$ 0.35
B_4NiO_7 0.34
Tanimoto Similarity

(AFLOWLIB)

$PbTiO_3$ was identified as very similar to NiO

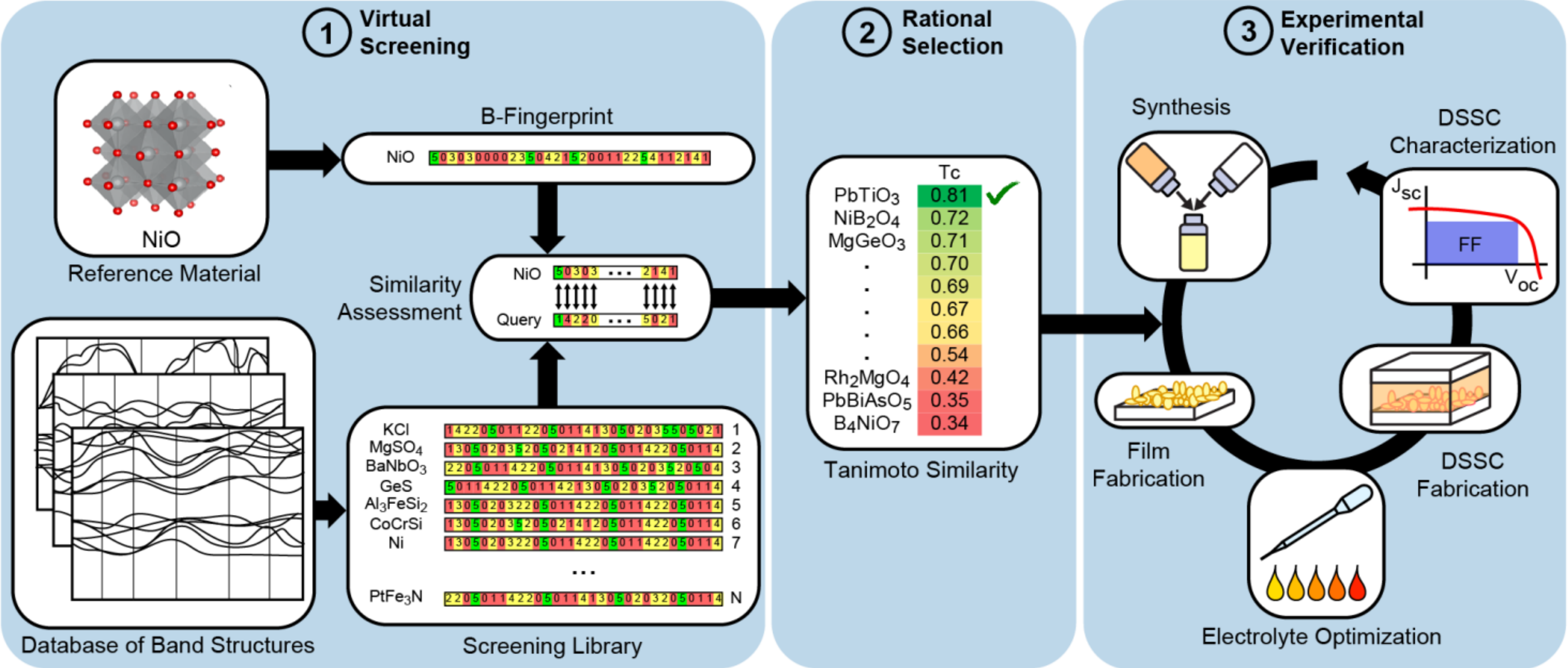
AND

It is has a dielectric constant >100

Moot, Isayev, Tropsha, Cahoon, Materials Discovery, **2016**, 6, 9-16

Design of alternate photocathodes

A materials informatics approach



(AFLOWLIB)

$PbTiO_3$ was identified as very similar to NiO

AND

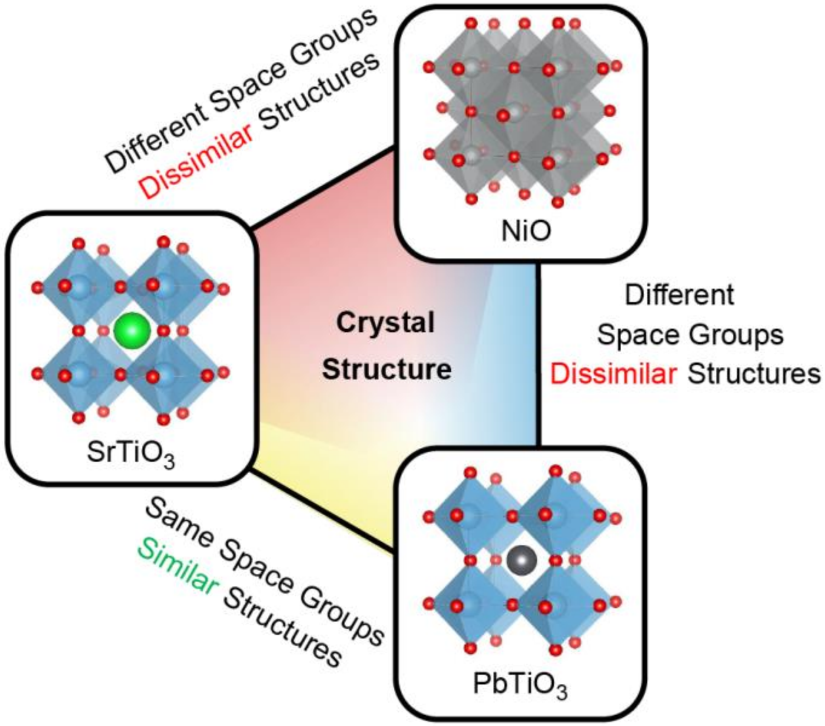
It is has a dielectric constant >100

Moot, Isayev, Tropsha, Cahoon, Materials Discovery, **2016**, 6, 9-16

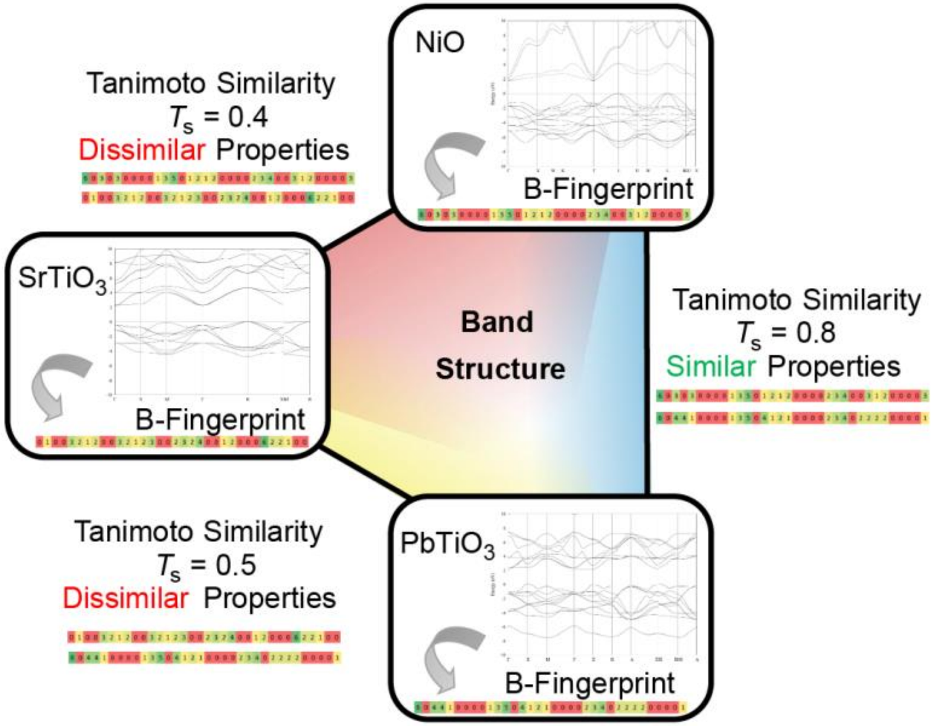
Materials informatics

Identifying top hit: $PbTiO_3$

Structure



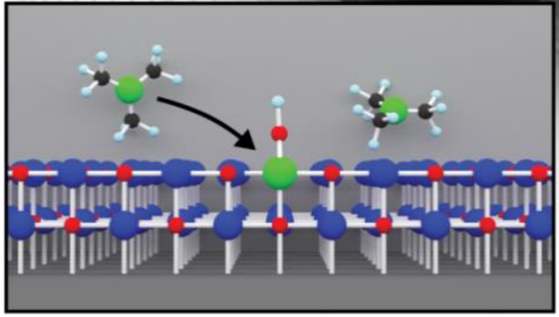
Properties



$PbTiO_3$ was identified as very similar to NiO in terms of electronic properties despite different crystal structures

Moot, Isayev, Tropsha, Cahoon, Materials Discovery, **2016**, 6, 9-16

Summary of Materials Informatics: Supporting Experimental Discovery



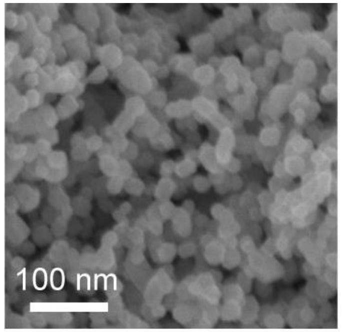
PbTiO3 is identified as a new photocathode material.

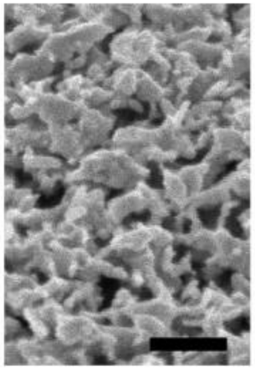
Successful experimental validation

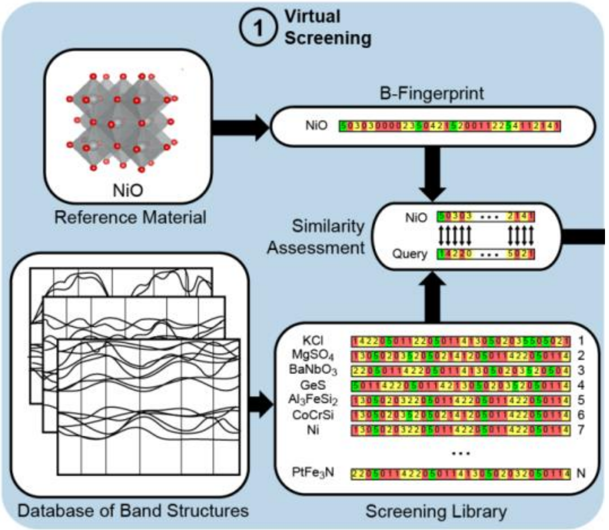
Record fill factors of >50

First fully aqueous DSSC device

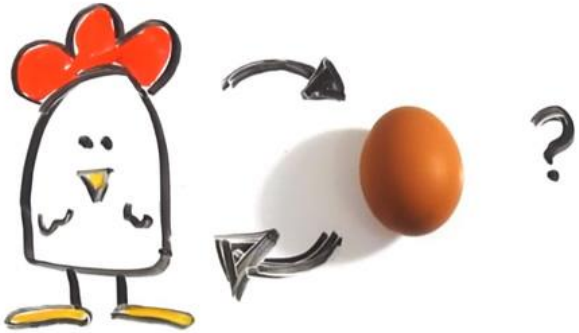
Currently, device performance is low; possible improvement by designing a new dye

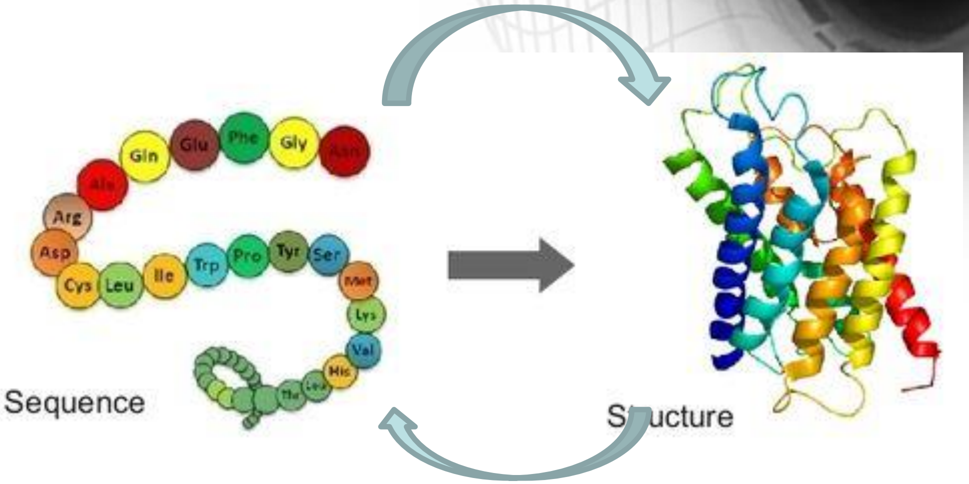






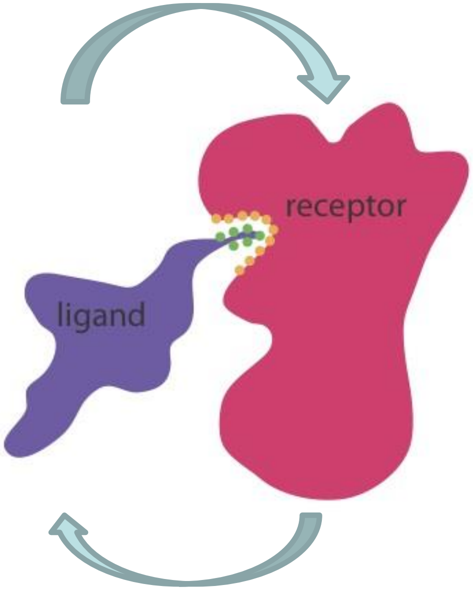
The eternal philosophical question: Which came first?





?

R or R^2 ?



?

Strasbourg Summer School in Chemoinformatics - 2018

Chem info's S^3

?

The eternal question: Which came first?

In the beginning was the Word…

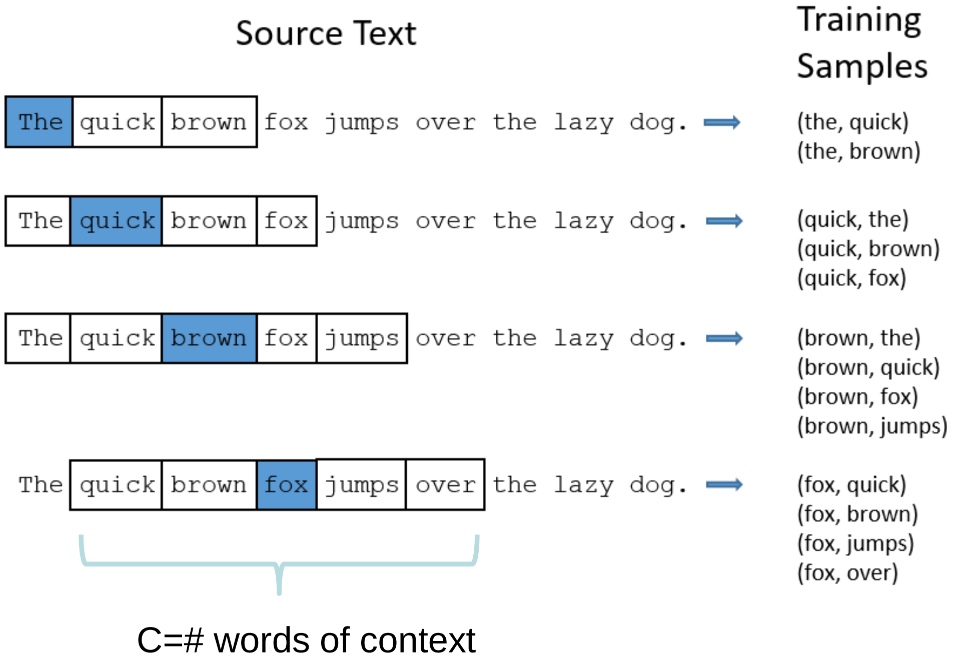
And the Word was… **embedded**

(freely adopted from the Gospel of John)

“You should know a word by the company it keeps”
J.R.Firth 1957

British linguist; formulated the notion of the “context-dependent nature of meaning”

Learning semantic context with Word2Vec



Source Text

Training Samples

The quick brown fox jumps over the lazy dog. → (the, quick) (the, brown)

The quick brown fox jumps over the lazy dog. → (quick, the) (quick, brown) (quick, fox)

The quick brown fox jumps over the lazy dog. → (brown, the) (brown, quick) (brown, fox) (brown, jumps)

The quick brown fox jumps over the lazy dog. → (fox, quick) (fox, brown) (fox, jumps) (fox, over)

C=# words of context

Can be used to learn:

CBOW:

- *Pr(word_k|words_context)*

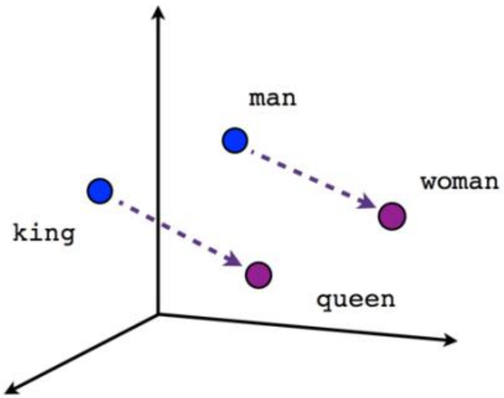
Skip-Gram:

- *Pr(words_context|word_k)*

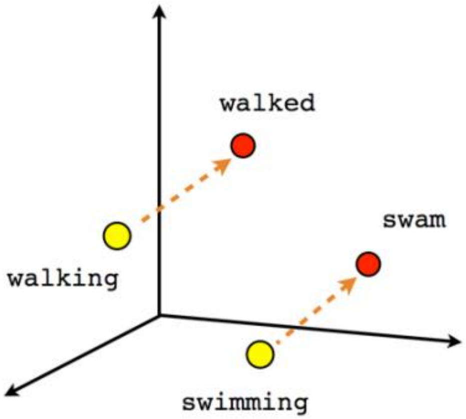
- Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

Word2Vec Images courtesy of Chris McCormick: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

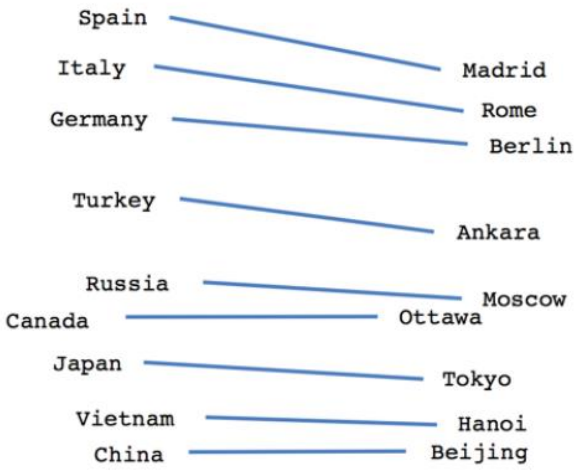
Word embedding and similarity in the semantic space



Male-Female



Verb tense



Country-Capital

SMILES are words that uniquely describe sentence-molecules!

Aspirin, also known as O=C(C)Oc1ccccc1C(=O)O, is a medication used to treat pain, fever, and inflammation.

CC(=C)C1(C=CC(=O)O1)O is a mycotoxin that is produced by Aspergillus flavus and *Penicillium roqueforti* mold.

COMPOUNDS	O=C(C)Oc1ccccc1C(=O)O	Active	1	ACTIVITY
	CCOc1cc(C)ccc1OCC=CF	Inactive	0	
	COc1ccccc1OCCO	Inactive	0	
	CC(N)Sc1ccc(Cl)nc1	Inactive	0	
	COC(=O)NCc1ccccc1Cl	Active	1	

ReLeaSE* design principles: learning and exploiting structural linguistics of SMILES notation

- SMILES notations reflect rules of Chemistry
- SMILES notation embeds linguistic rules
- Neural nets could learn both of the above types of rules
- This knowledge can be transformed into the generation of new SMILES corresponding to novel chemically feasible molecules (generative model)
- One can build QSAR models based solely on SMILES notation (predictive model)
- QSAR models can be used as a reward function for reinforcement learning to bias the design of novel libraries

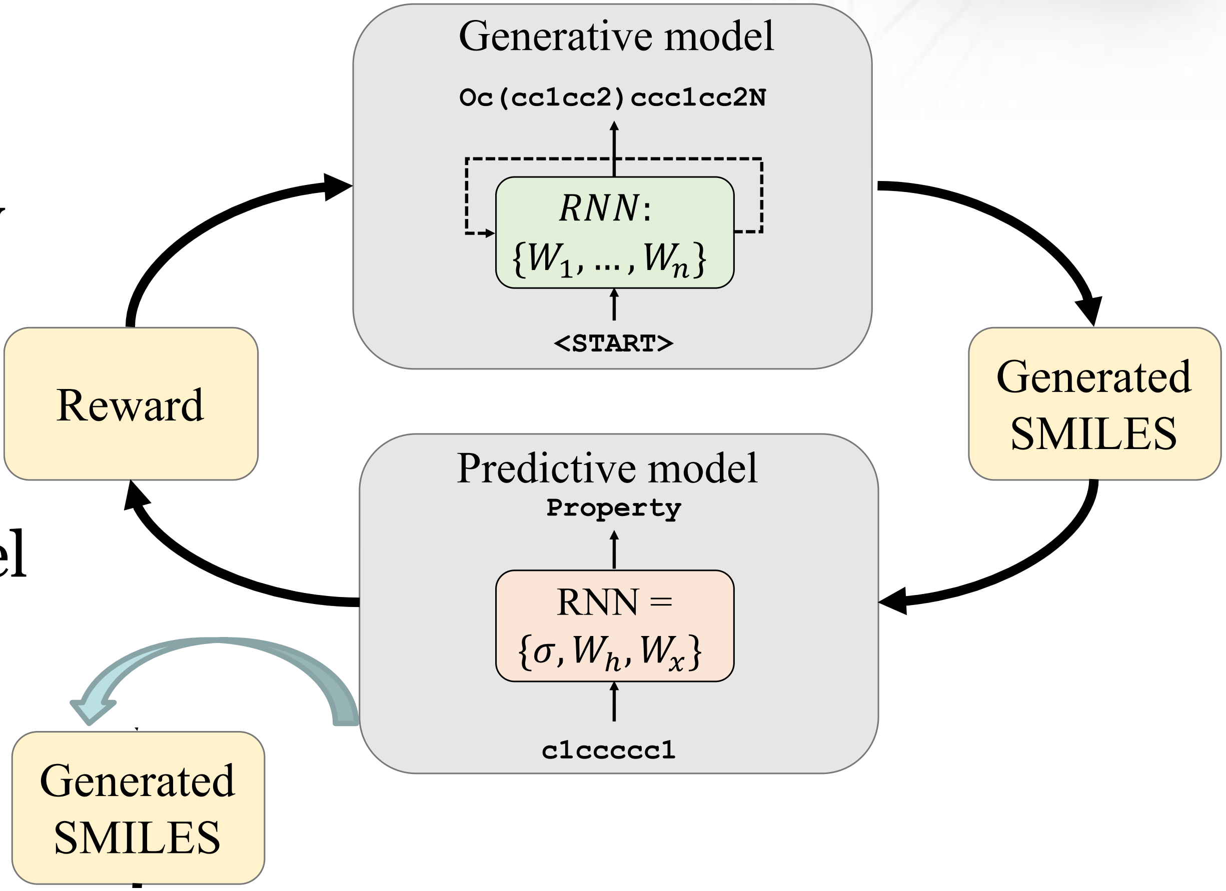
*Popova, Mariya, Olexandr Isayev, and Alexander Tropsha. "Deep reinforcement learning for de-novo drug design." *arXiv preprint arXiv:1711.10907* (2017). Science Advances, in press (2018)

Design of the ReLeaSE* method

(Reinforcement Learning for Structural Evolution)

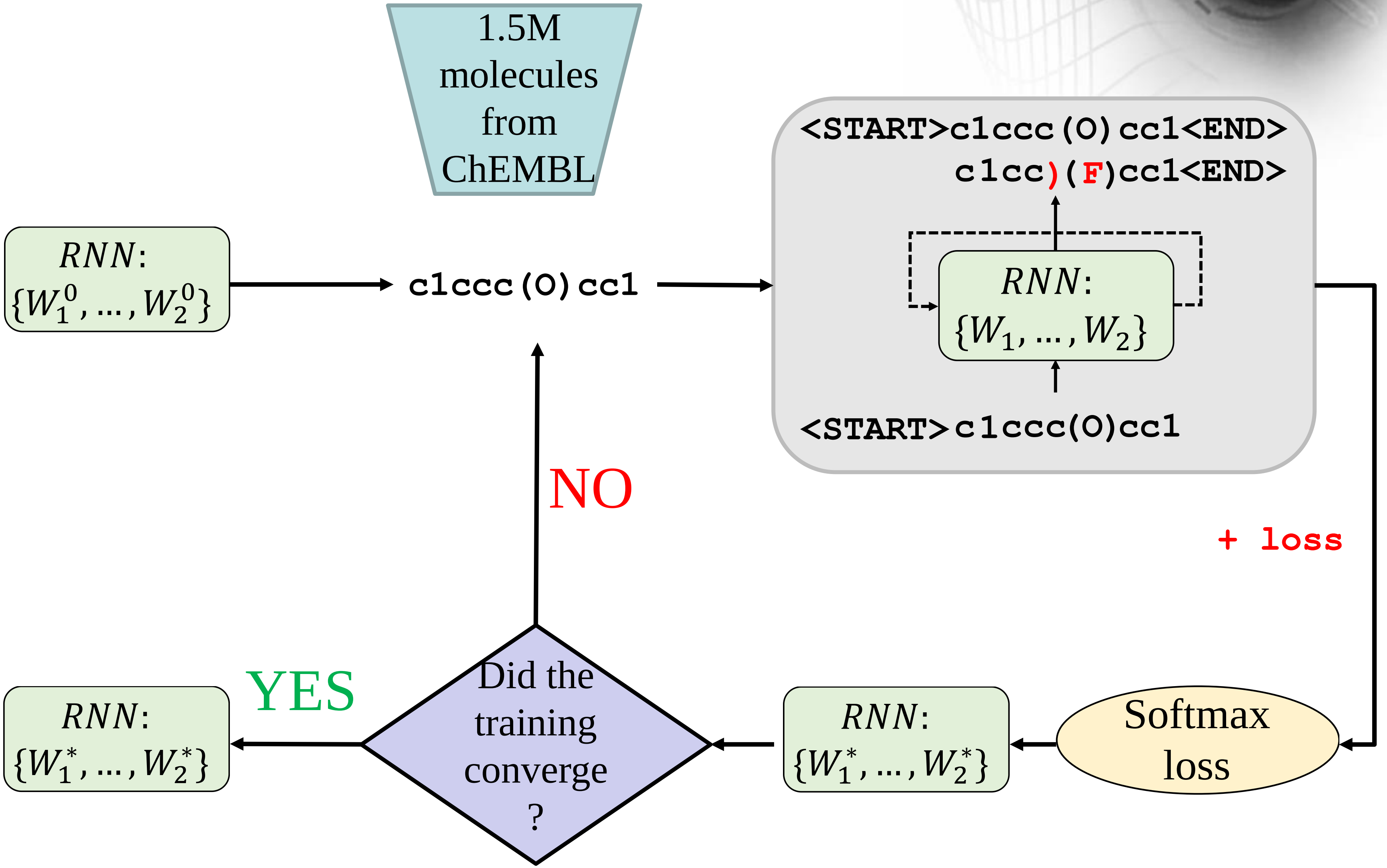
Elements of the thought cycle (molecules->models-molecules):

- Generate chemically feasible SMILES
- Develop SMILES-based QSAR model
- Employ QSAR model to bias library generation
- Produce new SMILES



*Popova, Mariya, Olexandr Isayev, and Alexander Tropsha. "Deep reinforcement learning for de-novo drug design." *arXiv preprint arXiv:1711.10907* (2017); Science Advances (in press).

Generative model: training mode



1.5M molecules from ChEMBL

RNN: $\{W_1^0, \ldots, W_2^0\}$

c1ccc(O)cc1

<START>c1ccc(O)cc1<END>

c1cc)(F)cc1<END>

RNN: $\{W_1, \ldots, W_2\}$

<START>c1ccc(O)cc1

+ loss

Softmax loss

RNN: $\{W_1^*, \ldots, W_2^*\}$

Did the training converge ?

NO

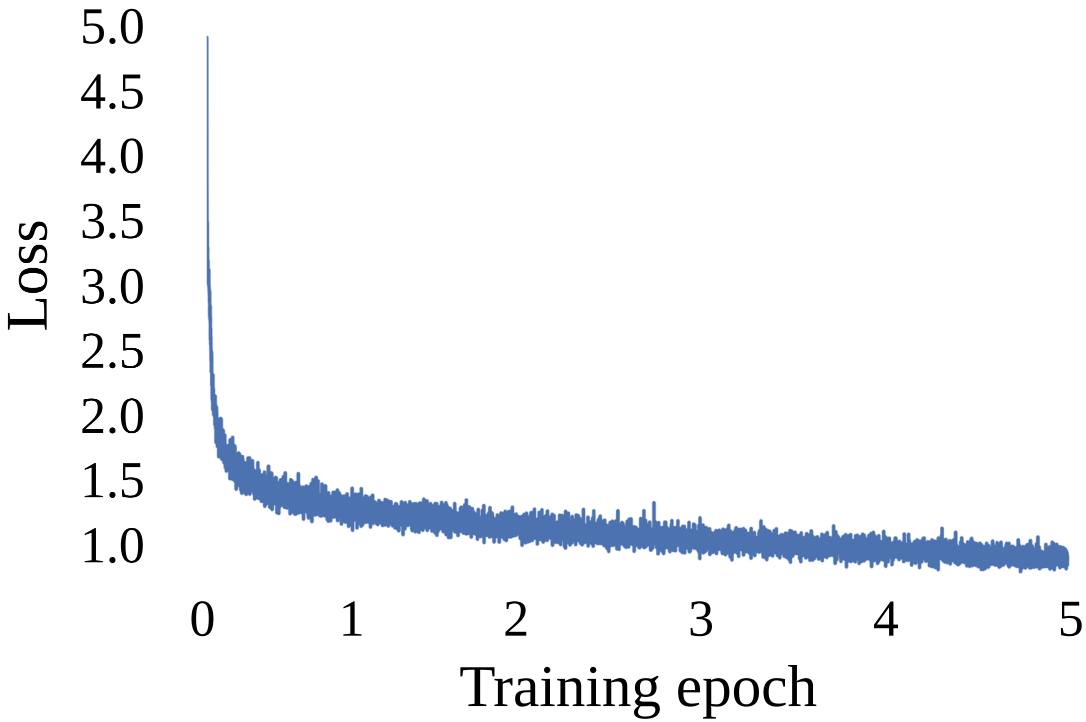
YES

RNN: $\{W_1^*, \ldots, W_2^*\}$

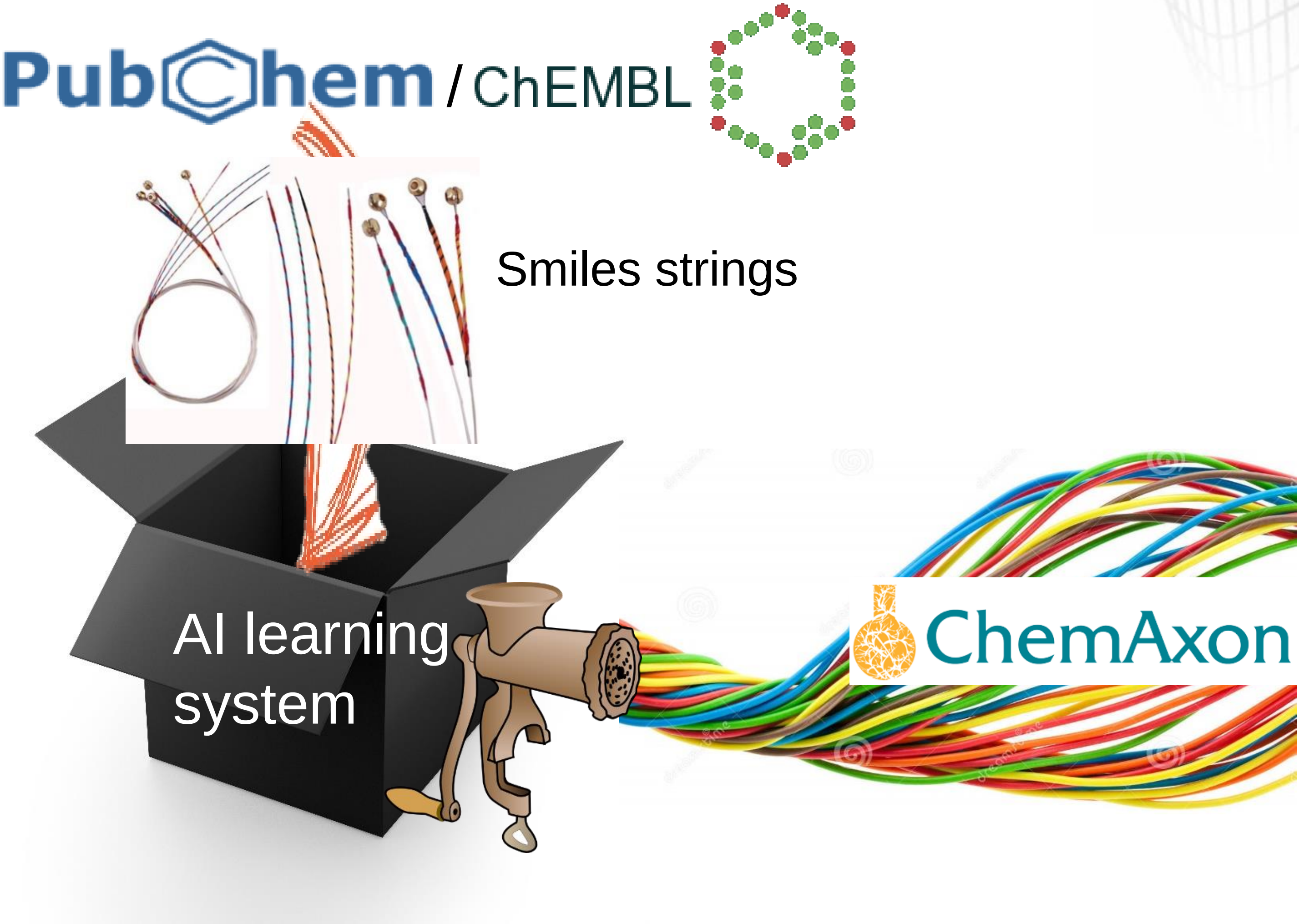
Generative model: training mode

- Training continues until convergence

- Every SMILES from ChEMBL is used as training example ~ 3-5 times



Are we making legitimate Smiles?



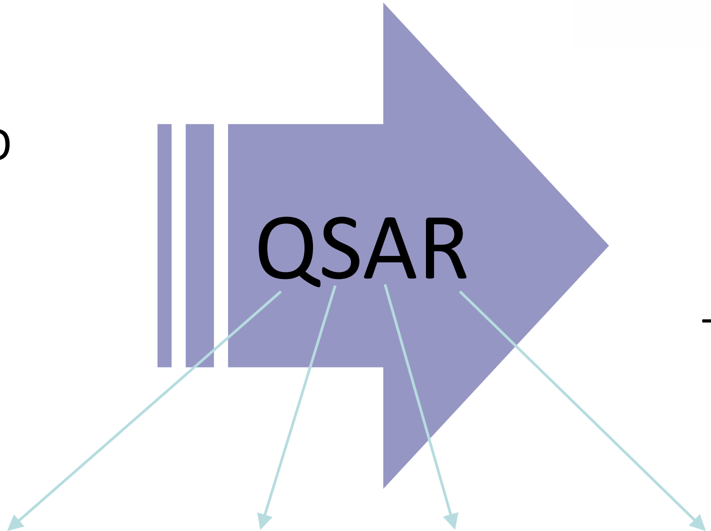
Smiles strings

AI learning system

95% Valid Chemically-feasible molecules

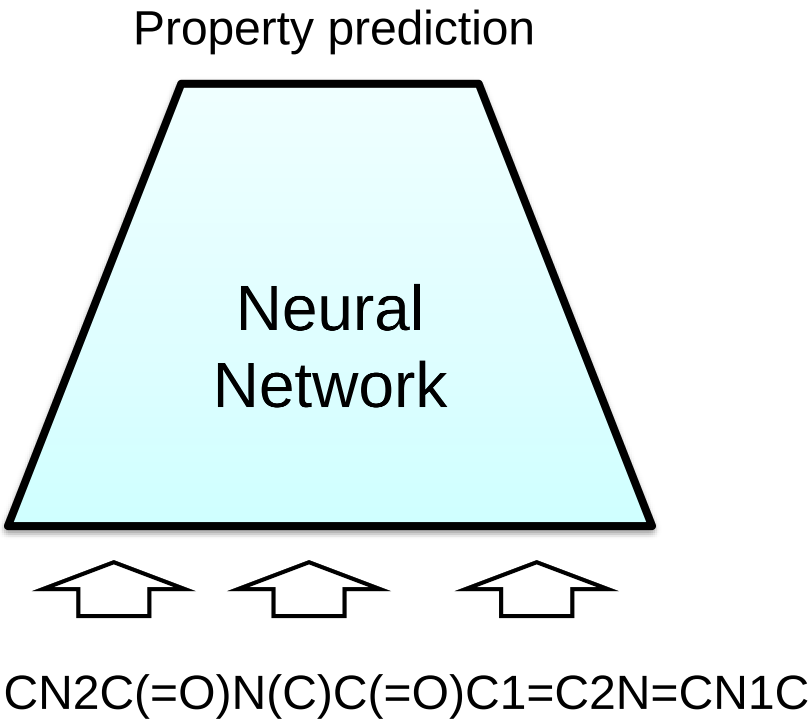
Smile-ification of QSAR!

COMPOUNDS	ACTIVITY
O=C(C)Oc1ccccc1C(=O)O	0.531
CCOc1cc(C)ccc1OCC=CF	1.299
COc1ccccc1OCCO	0.946
CC(N)Sc1ccc(Cl)nc1	-0.218
COC(=O)NCc1ccccc1Cl	0.017



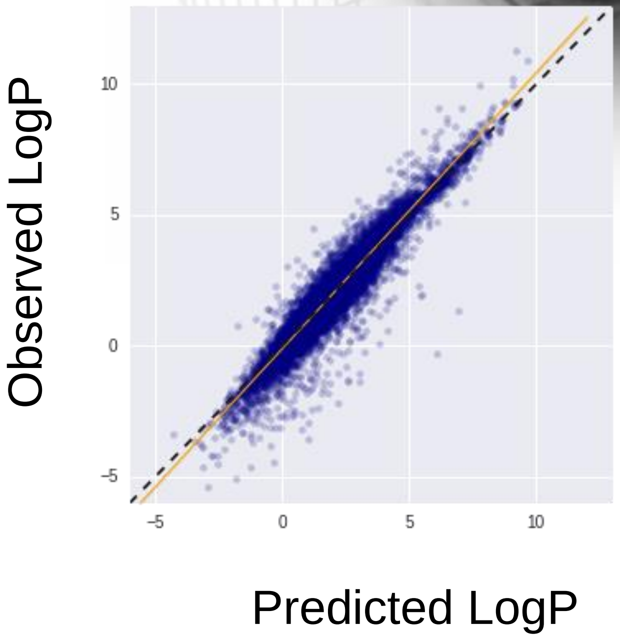
Quantitative Smiles – Activity Relationships

QSAR modeling using Smiles strings only*


Property prediction

Neural Network

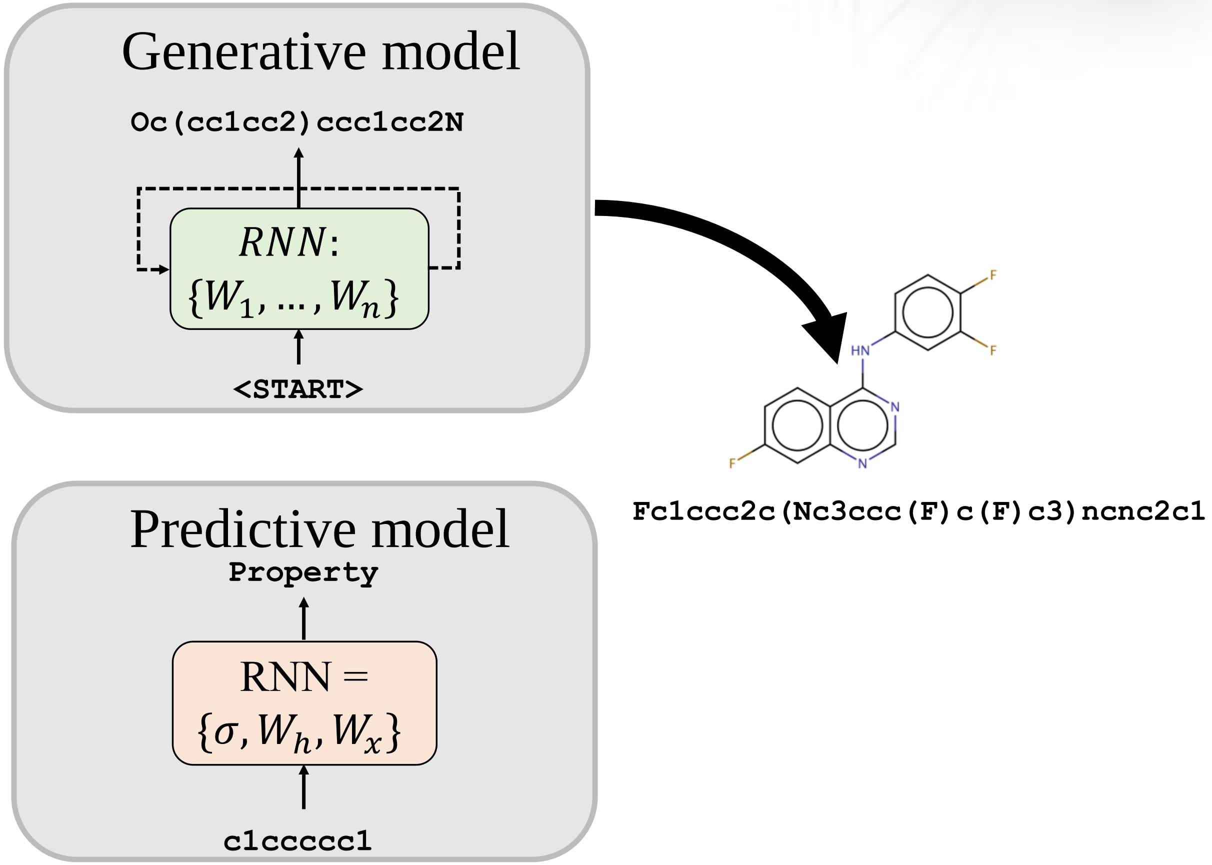
CN2C(=O)N(C)C(=O)C1=C2N=CN1C



	5CV RF model with DRAGON7 Descriptors	5CV NN model with SMILES directly
RMSE:	0.57	0.53
MAE:	0.37	0.35
R^2_{ext}:	0.90	0.91

*LogP data for ~16K molecules from PHYSPROP (srcinc.com), Toxcast Dashboard (https://comptox.epa.gov/dashboard), and others.

Reinforcement learning for chemical design



Generative model

`Oc(cc1cc2)ccc1cc2N`

RNN: $\{W_1, \ldots, W_n\}$

`<START>`

`Fc1ccc2c(Nc3ccc(F)c(F)c3)ncnc2c1`

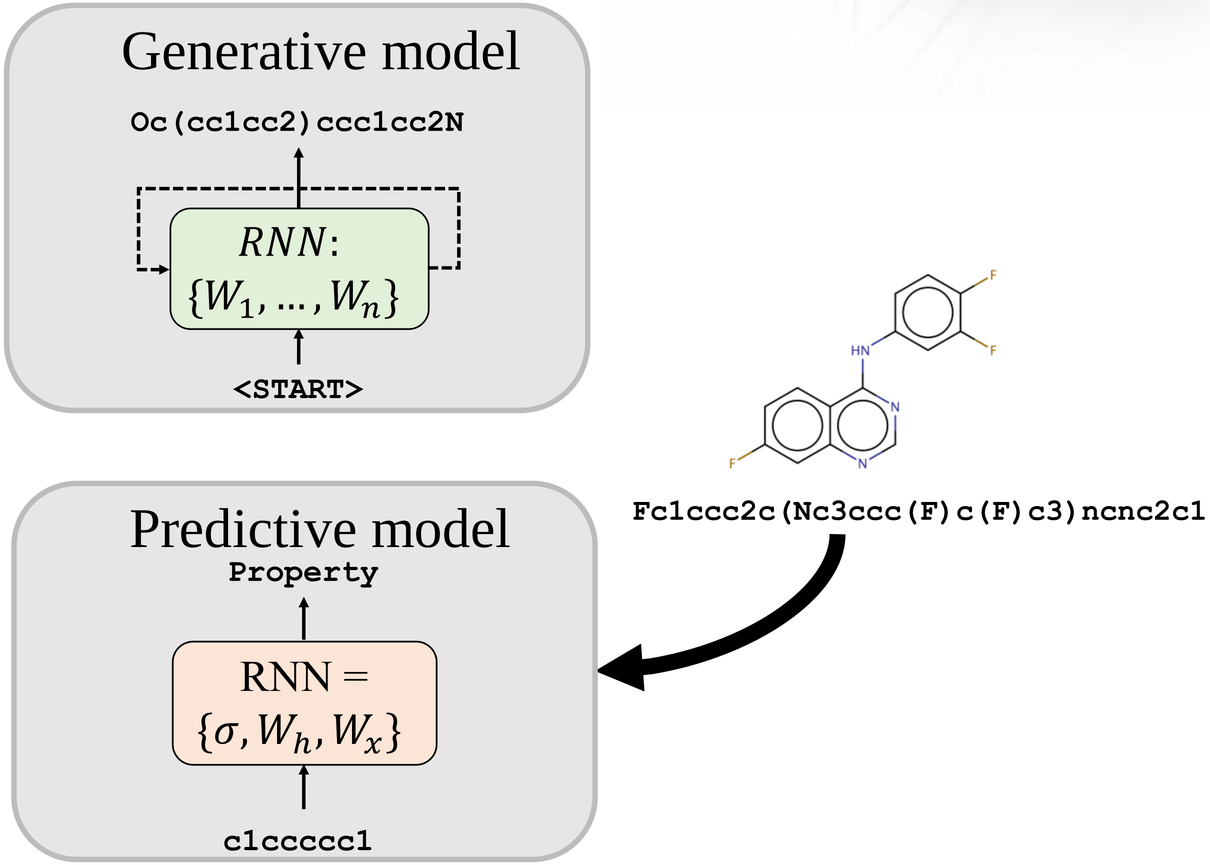
Predictive model

`Property`

RNN = $\{\sigma, W_h, W_x\}$

`c1ccccc1`

Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN:

$\{W_1, \ldots, W_n\}$

<START>

Fc1ccc2c(Nc3ccc(F)c(F)c3)ncnc2c1

Predictive model

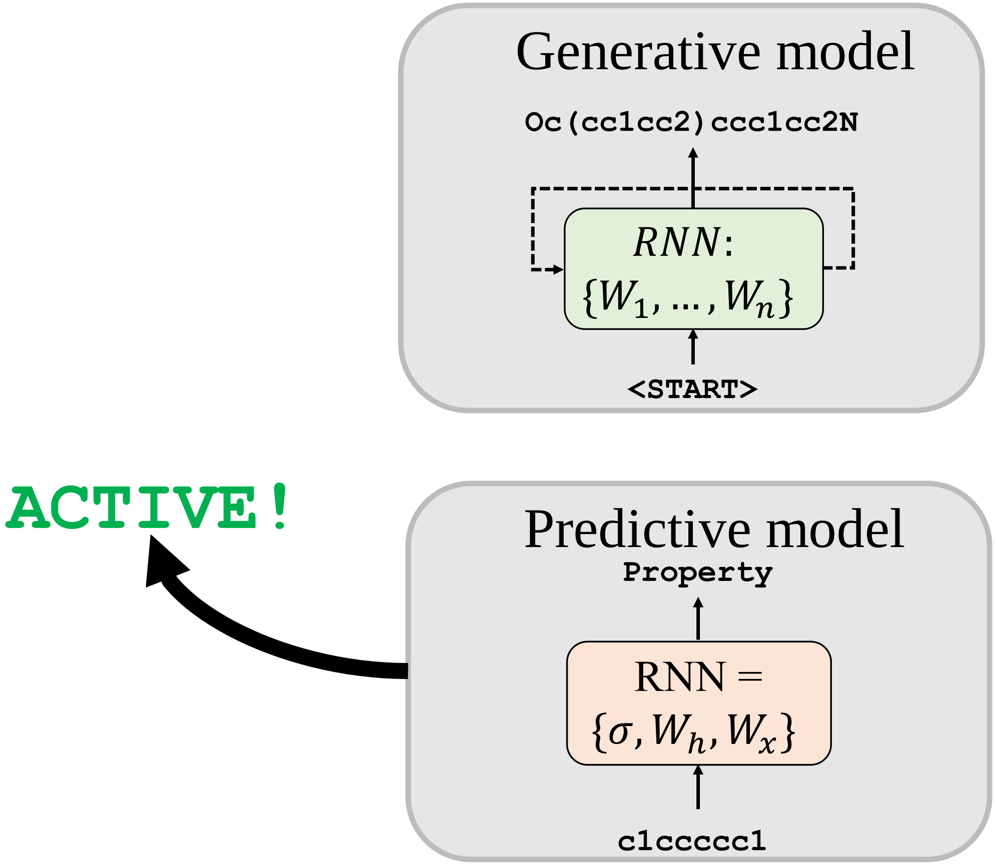
Property

RNN =

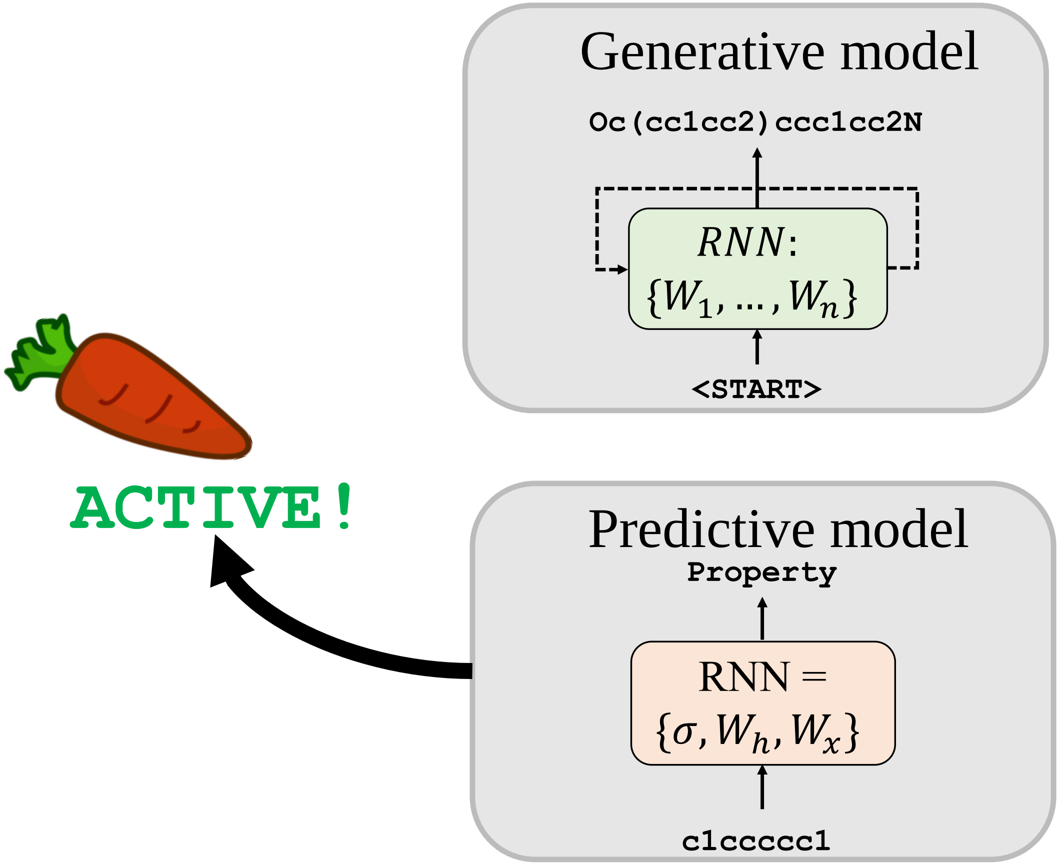
$\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design



Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN: $\{W_1, \dots, W_n\}$

<START>

ACTIVE!

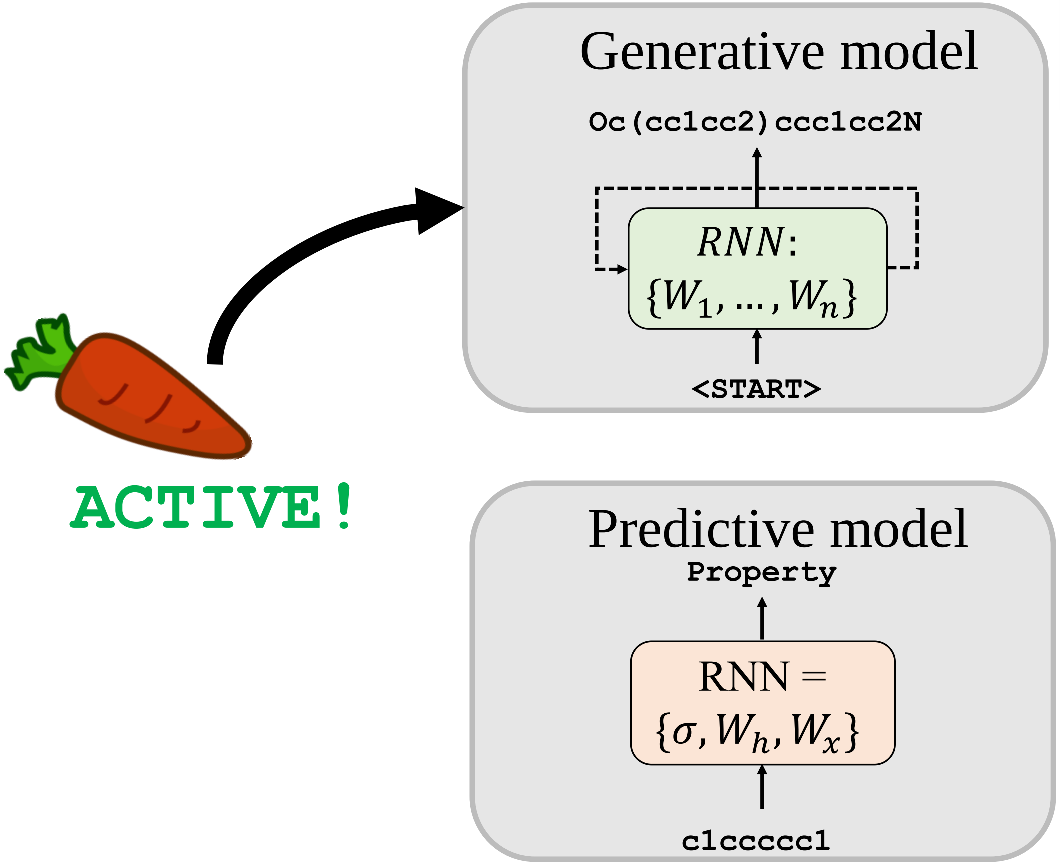
Predictive model

Property

RNN = $\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN: $\{W_1, \dots, W_n\}$

<START>

ACTIVE!

Predictive model

Property

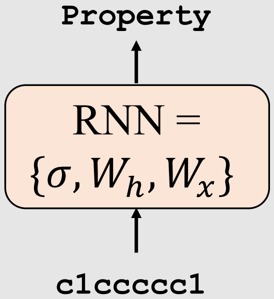
RNN = $\{\sigma, W_h, W_x\}$

c1ccccc1

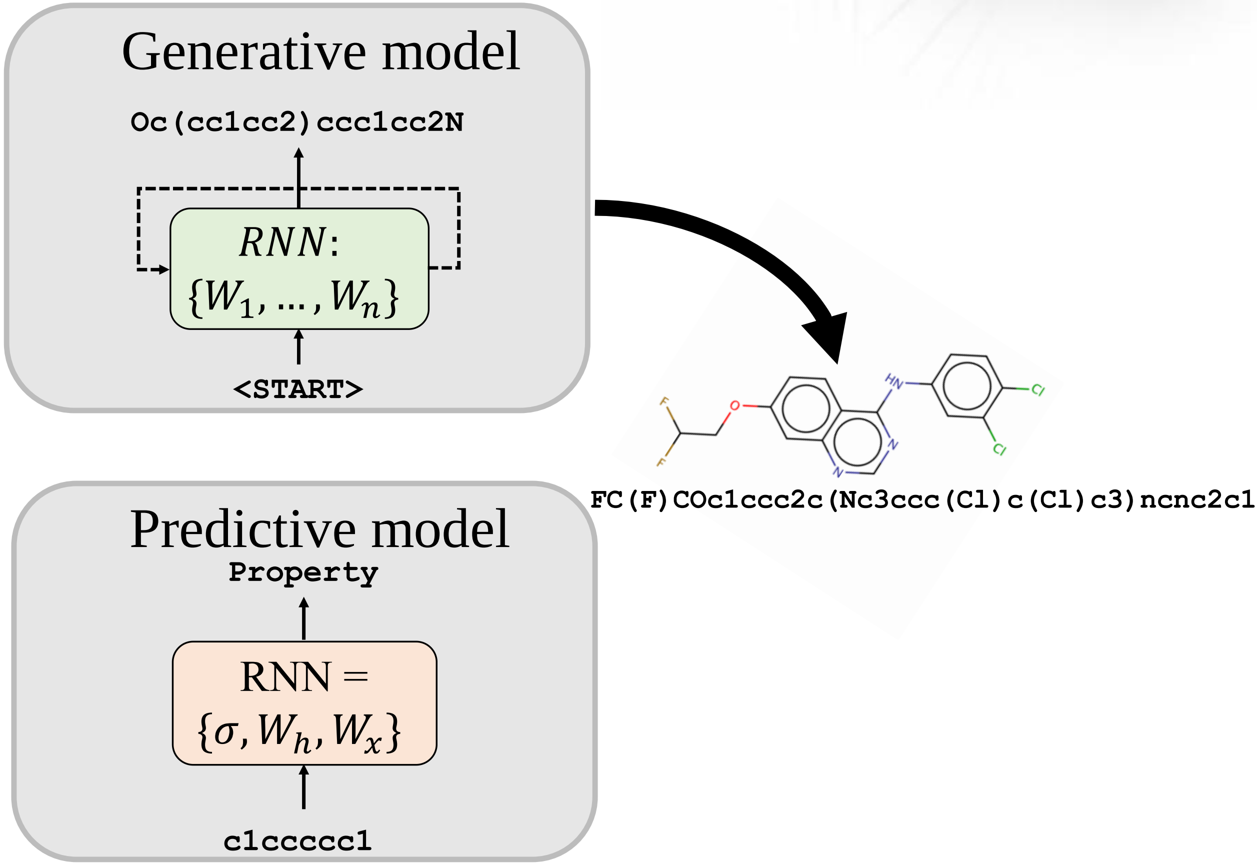
Reinforcement learning for chemical design

Generative model

Predictive model



Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN:

$\{W_1, \dots, W_n\}$

<START>

FC(F)COc1ccc2c(Nc3ccc(Cl)c(Cl)c3)ncnc2c1

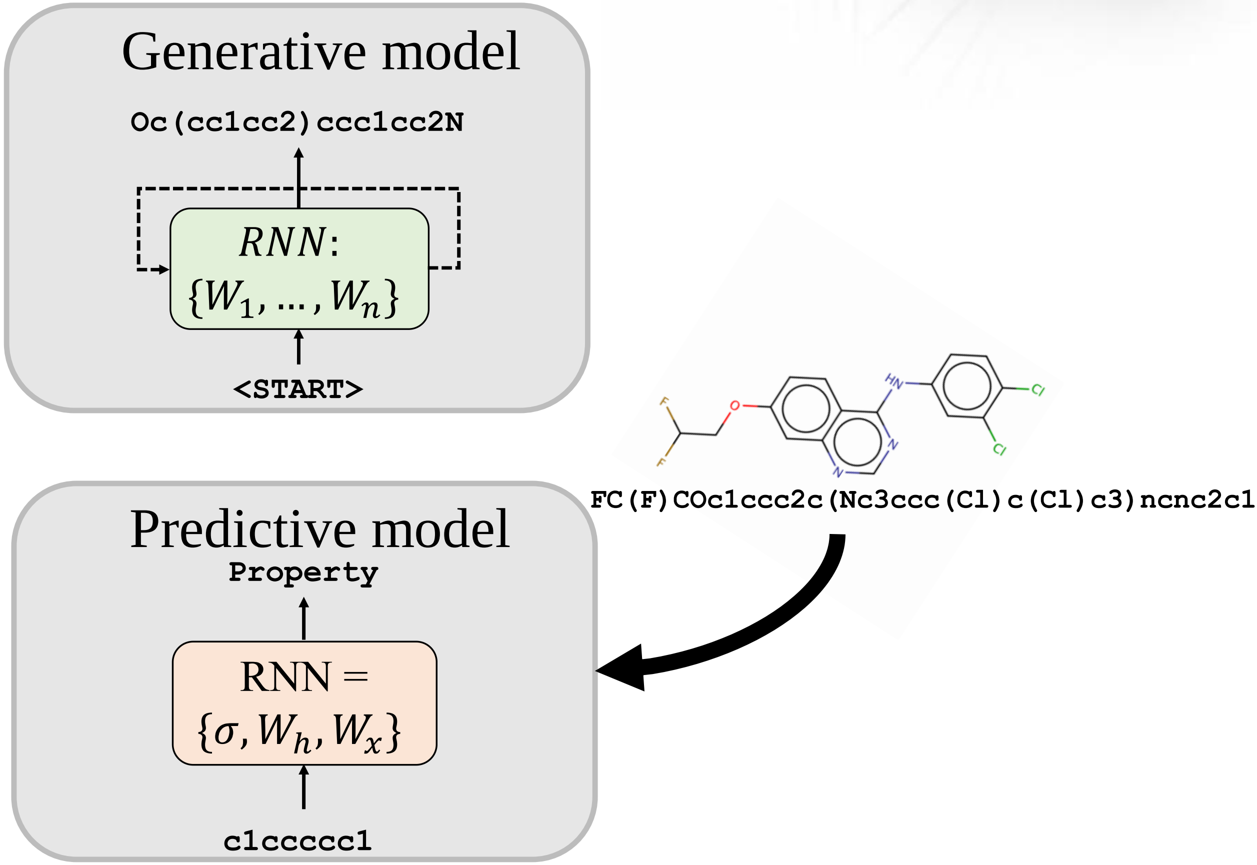
Predictive model

Property

RNN = $\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN: $\{W_1, \dots, W_n\}$

<START>

FC(F)COc1ccc2c(Nc3ccc(Cl)c(Cl)c3)ncnc2c1

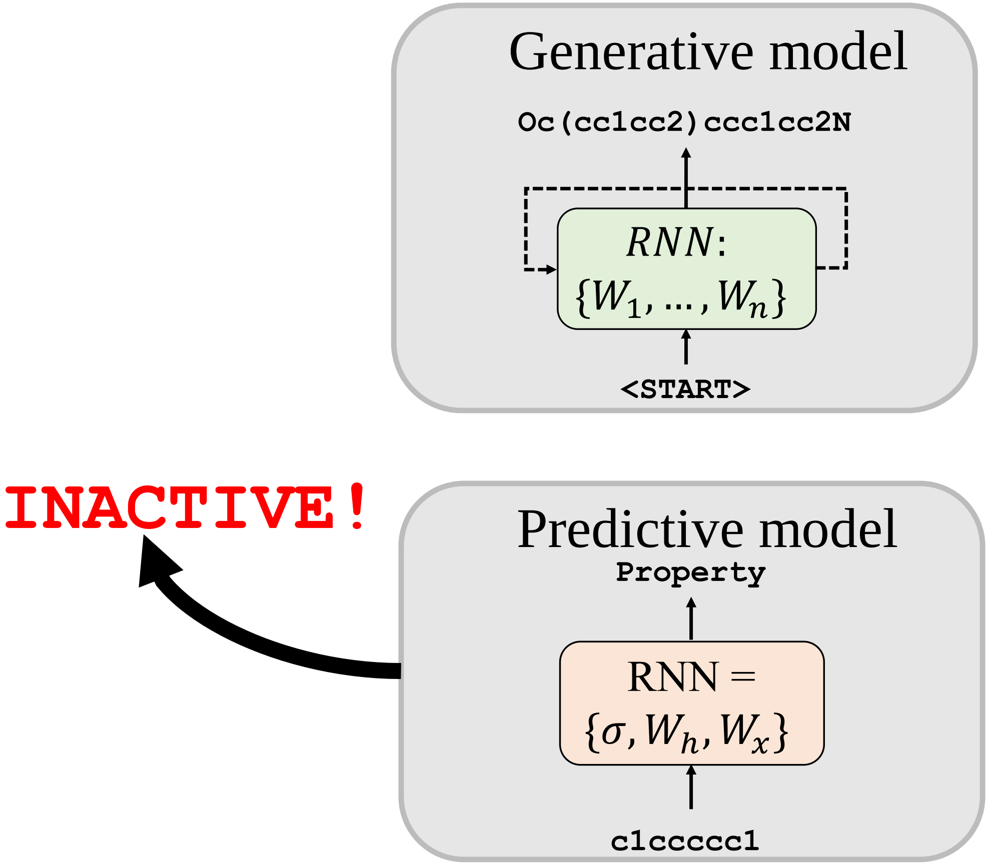
Predictive model

Property

RNN = $\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

$RNN:$

$\{W_1, \ldots, W_n\}$

<START>

INACTIVE!

Predictive model

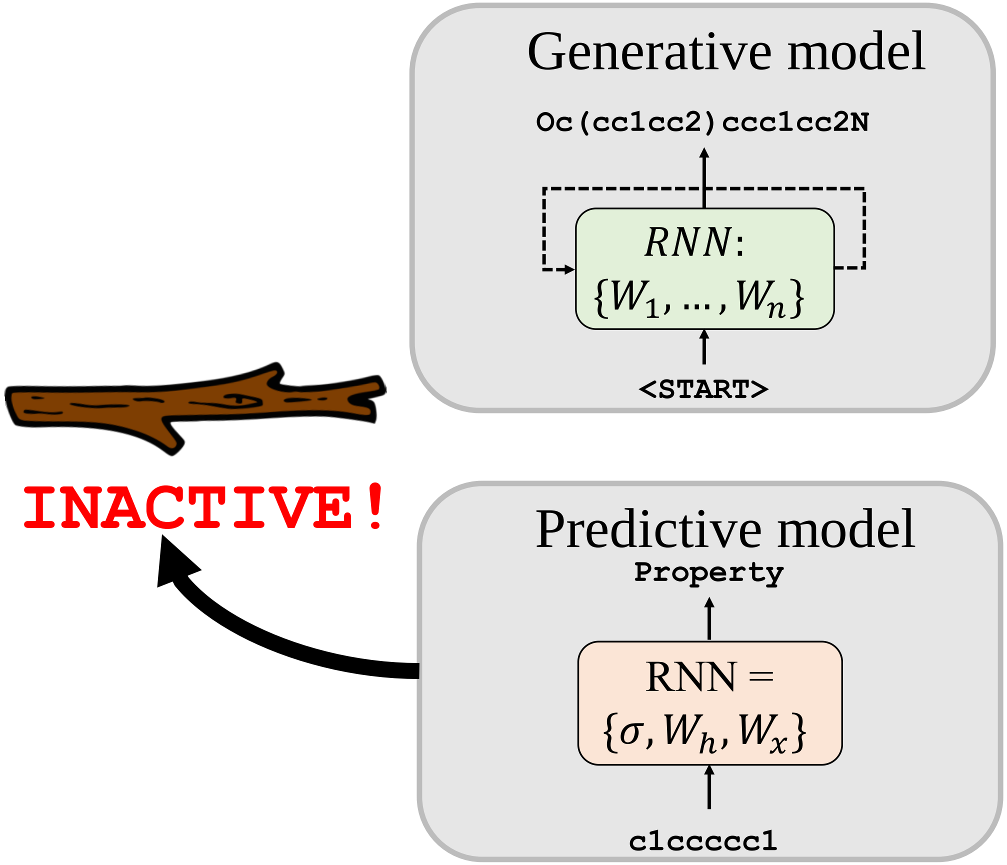
Property

$\text{RNN} =$

$\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN: $\{W_1, \ldots, W_n\}$

<START>

INACTIVE!

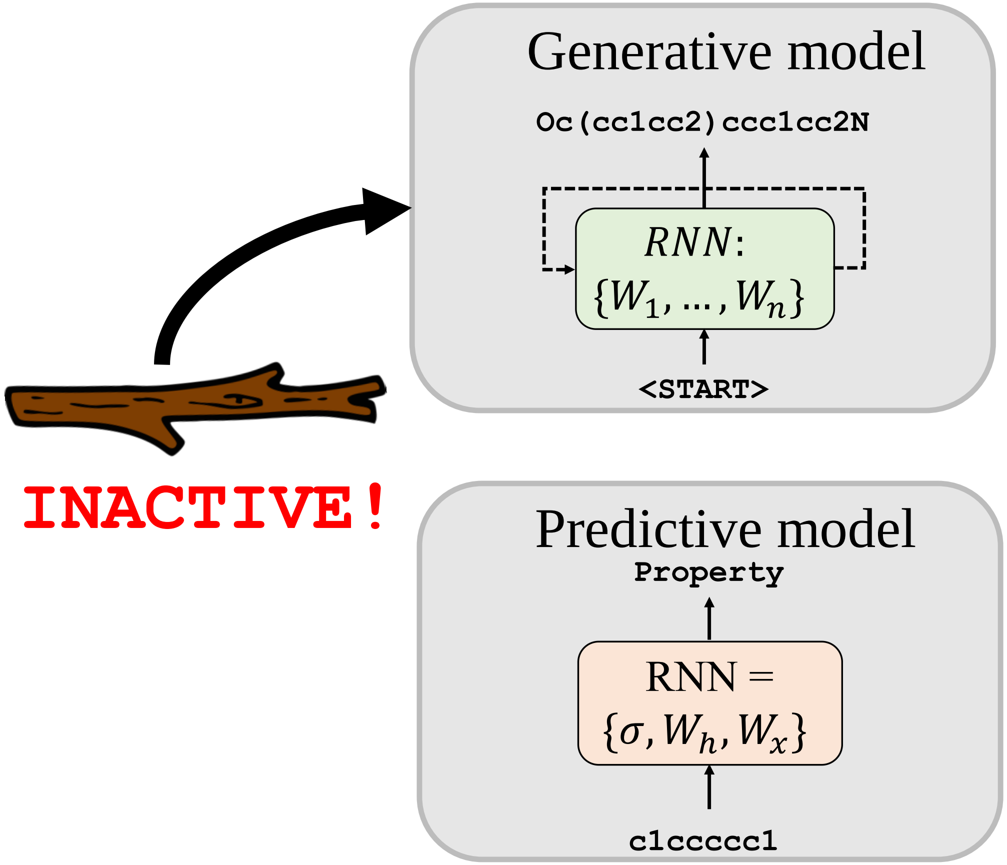
Predictive model

Property

RNN = $\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design



Generative model

Oc(cc1cc2)ccc1cc2N

RNN:

$\{W_1, \dots, W_n\}$

<START>

INACTIVE!

Predictive model

Property

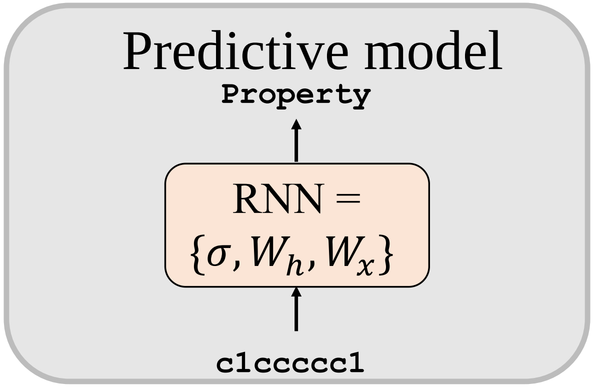
RNN =

$\{\sigma, W_h, W_x\}$

c1ccccc1

Reinforcement learning for chemical design

Generative model

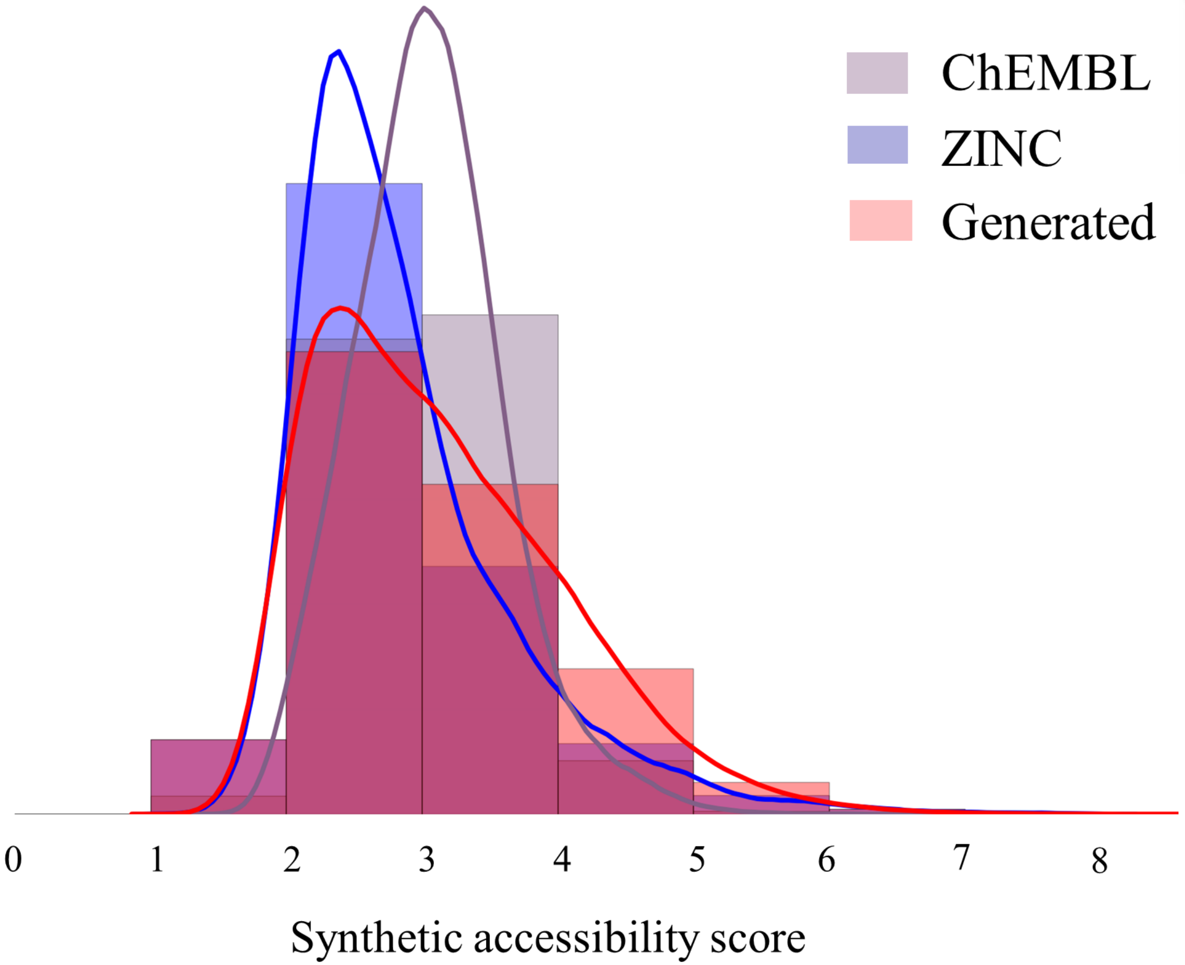


Predictive model

Technical details

- Models were trained on Nvidia Titan X and Titan V GPUs
- Training the generative model on ChEMBL took ~ 25 days
- Training of predictive models took ~ 2 hours
- Biasing the generative model with reinforcement learning for one property ~ 1 day
- Generative model produces 1000 compounds per minute

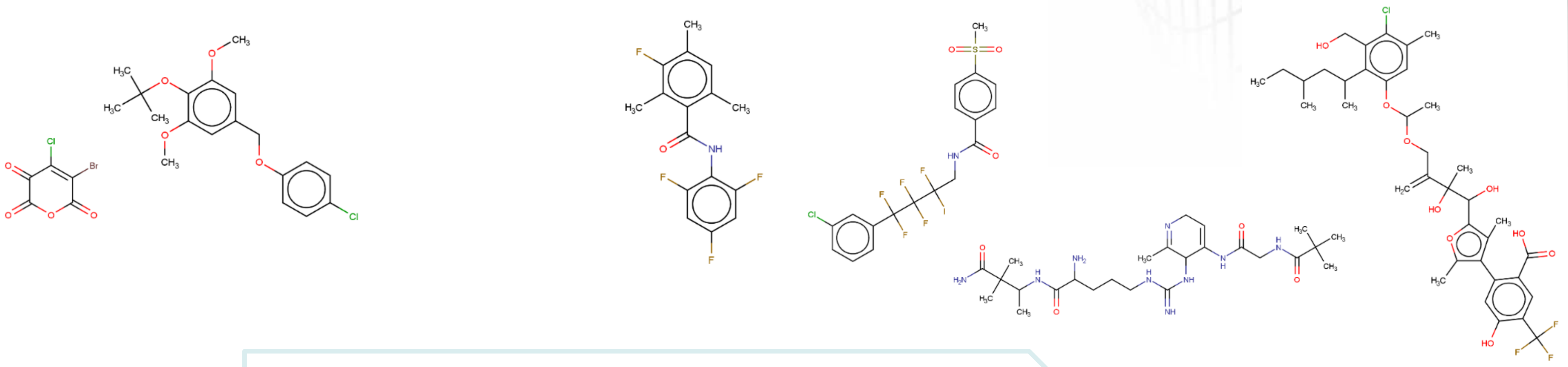
Results: Synthetic accessibility score* of the designed libraries



*Ertl, Peter, and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions." *Journal of cheminformatics* 1.1 (2009): 8.

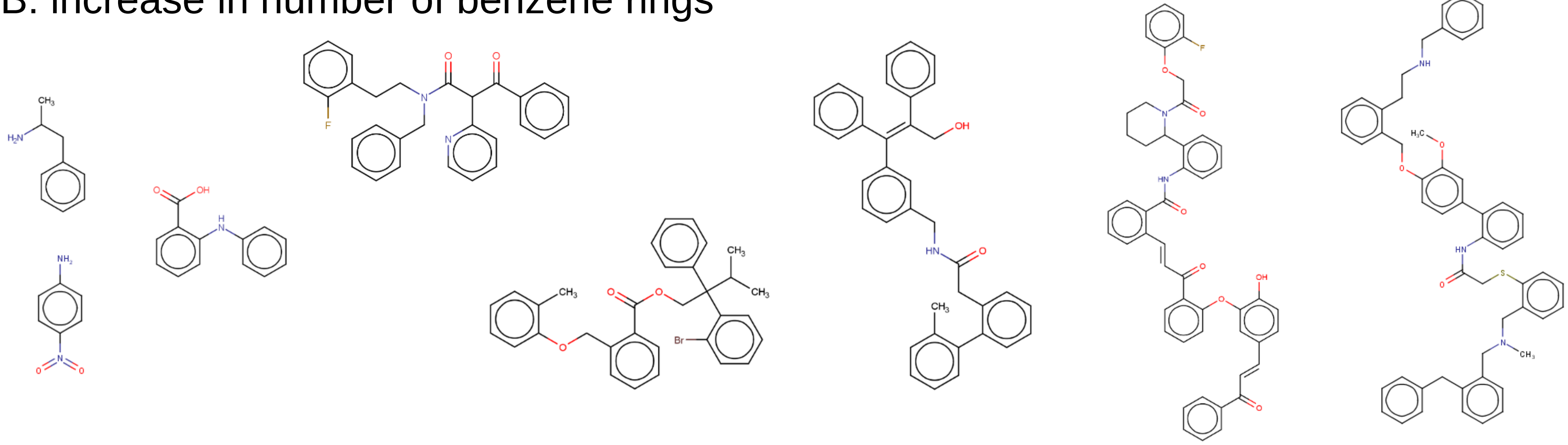
PoC: Structural Bias

A: increase in number of substituents

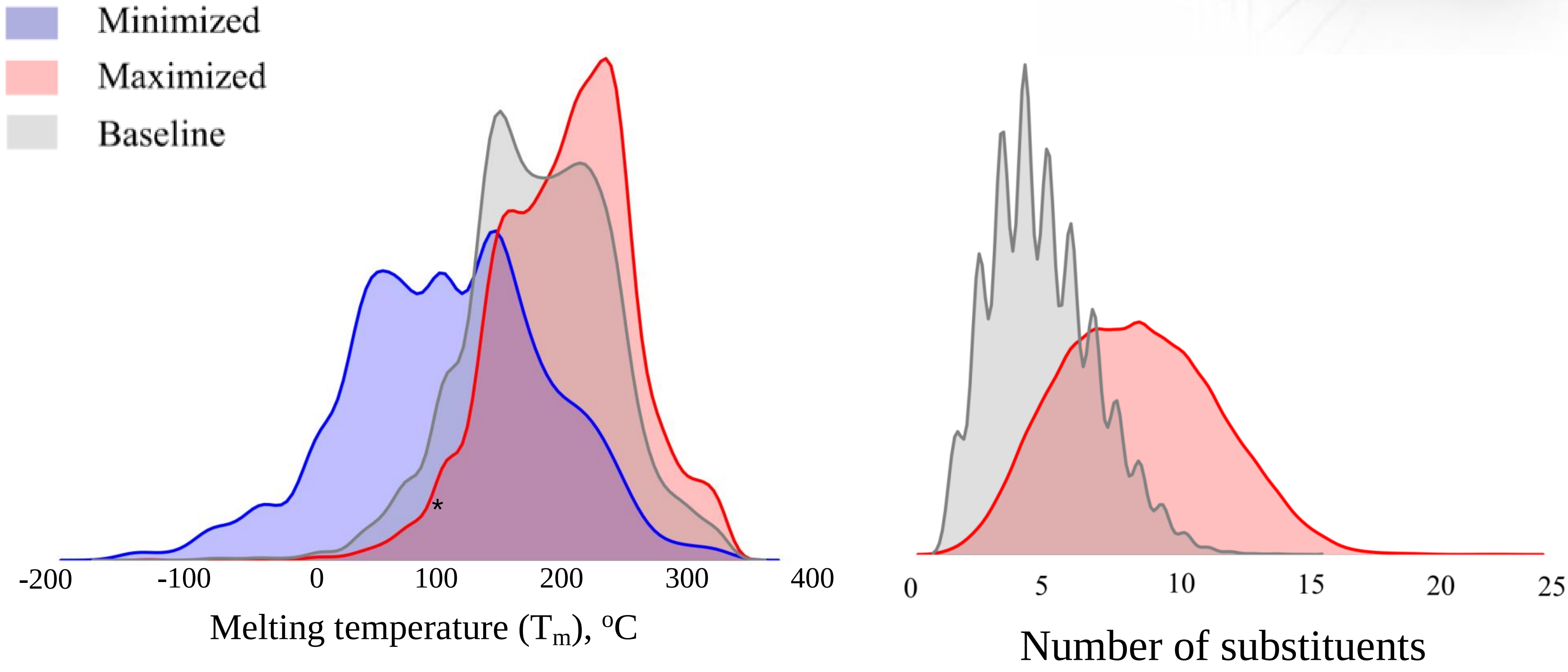


Reward increase

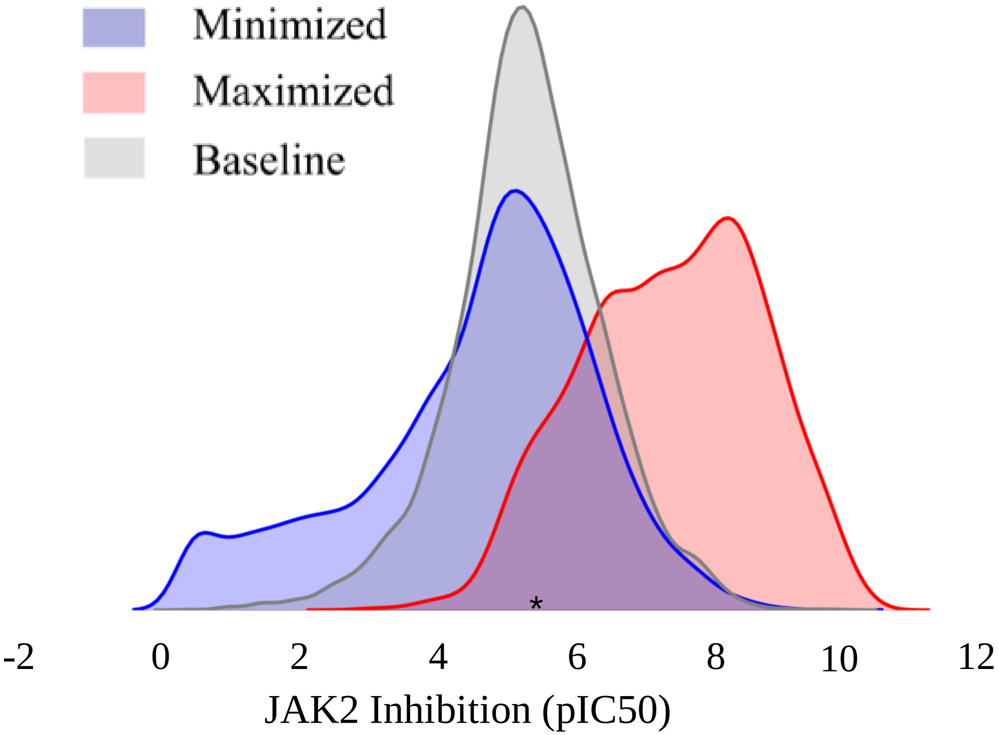
B: increase in number of benzene rings

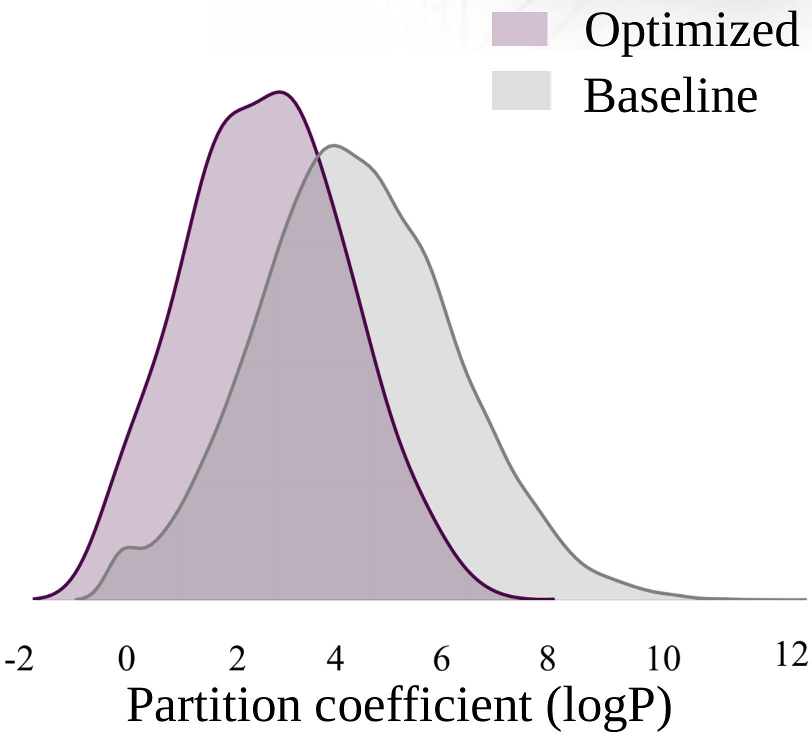


Results: Biasing target properties in the designed libraries

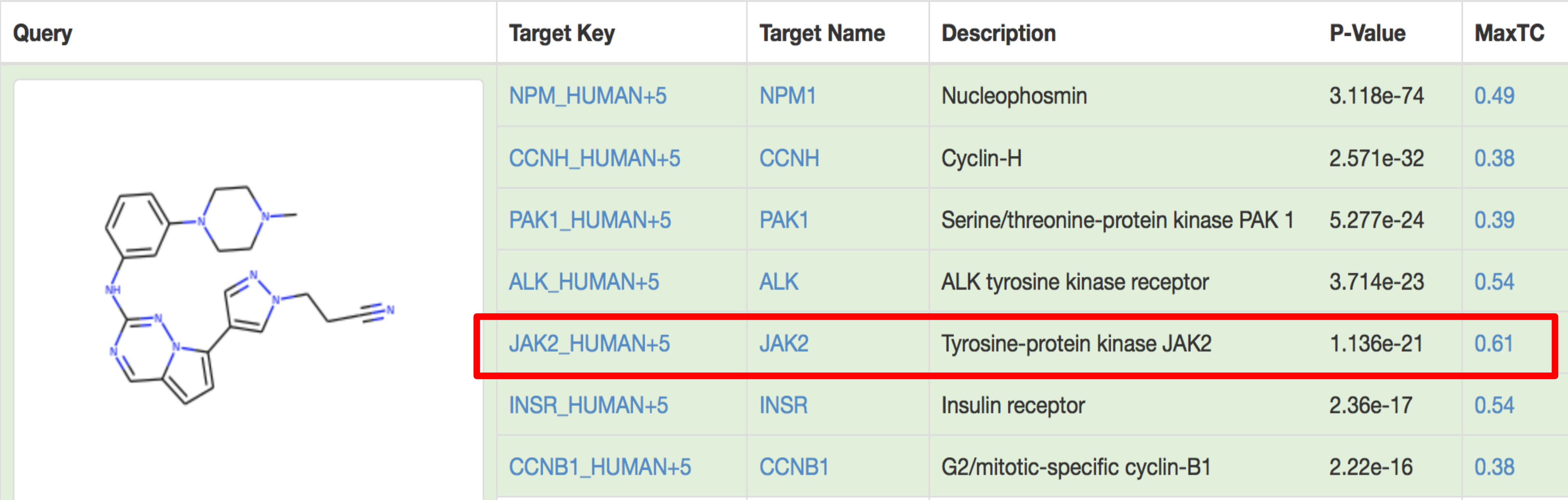


Results: Biasing target properties in the designed libraries





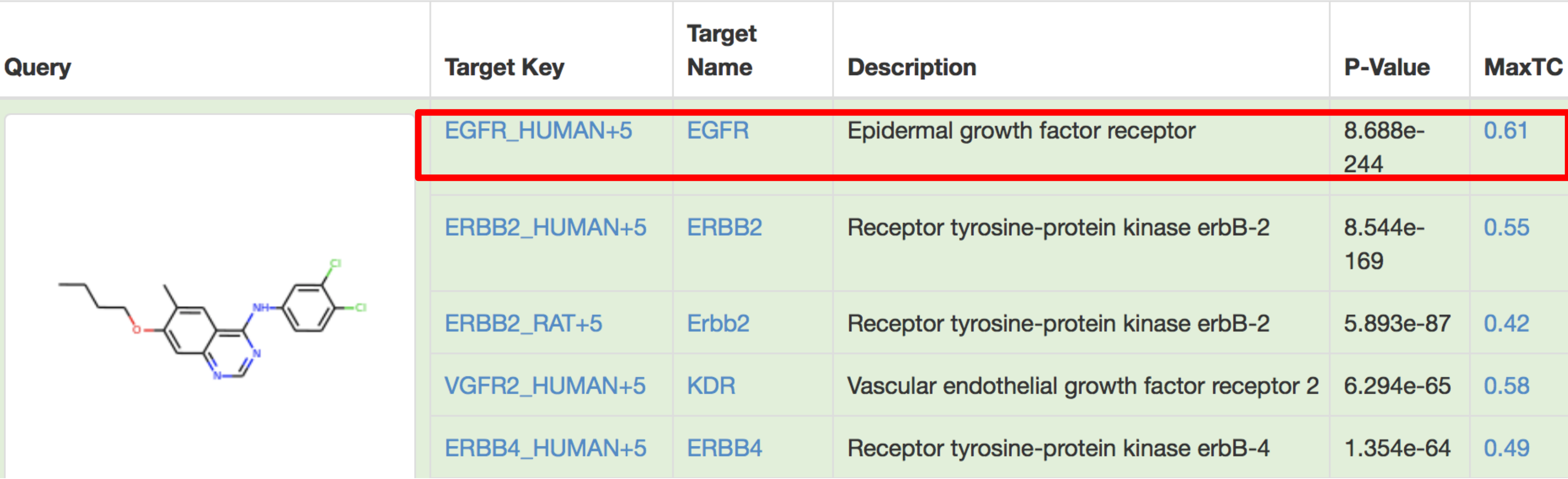
Target predictions for generated compounds using SEA*



Query	Target Key	Target Name	Description	P-Value	MaxTC
	NPM_HUMAN+5	NPM1	Nucleophosmin	3.118e-74	0.49
	CCNH_HUMAN+5	CCNH	Cyclin-H	2.571e-32	0.38
	PAK1_HUMAN+5	PAK1	Serine/threonine-protein kinase PAK 1	5.277e-24	0.39
	ALK_HUMAN+5	ALK	ALK tyrosine kinase receptor	3.714e-23	0.54
	JAK2_HUMAN+5	JAK2	Tyrosine-protein kinase JAK2	1.136e-21	0.61
	INSR_HUMAN+5	INSR	Insulin receptor	2.36e-17	0.54
	CCNB1_HUMAN+5	CCNB1	G2/mitotic-specific cyclin-B1	2.22e-16	0.38

*Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotech* **25** (2), 197-206 (2007).

Target predictions for generated compounds using SEA*

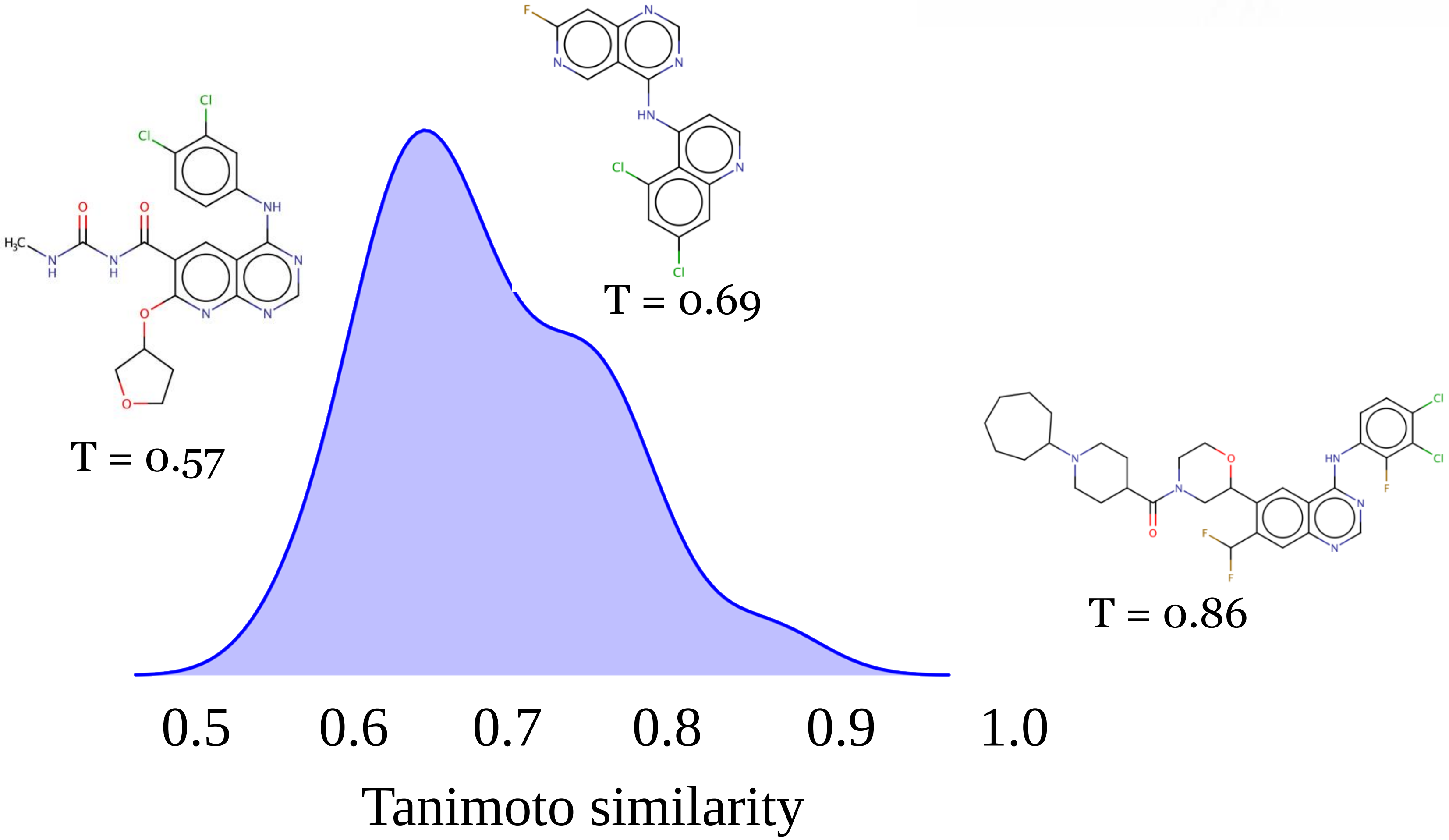


Query	Target Key	Target Name	Description	P-Value	MaxTC
	EGFR_HUMAN+5	EGFR	Epidermal growth factor receptor	8.688e-244	0.61
	ERBB2_HUMAN+5	ERBB2	Receptor tyrosine-protein kinase erbB-2	8.544e-169	0.55
	ERBB2_RAT+5	Erbb2	Receptor tyrosine-protein kinase erbB-2	5.893e-87	0.42
	VGFR2_HUMAN+5	KDR	Vascular endothelial growth factor receptor 2	6.294e-65	0.58
	ERBB4_HUMAN+5	ERBB4	Receptor tyrosine-protein kinase erbB-4	1.354e-64	0.49

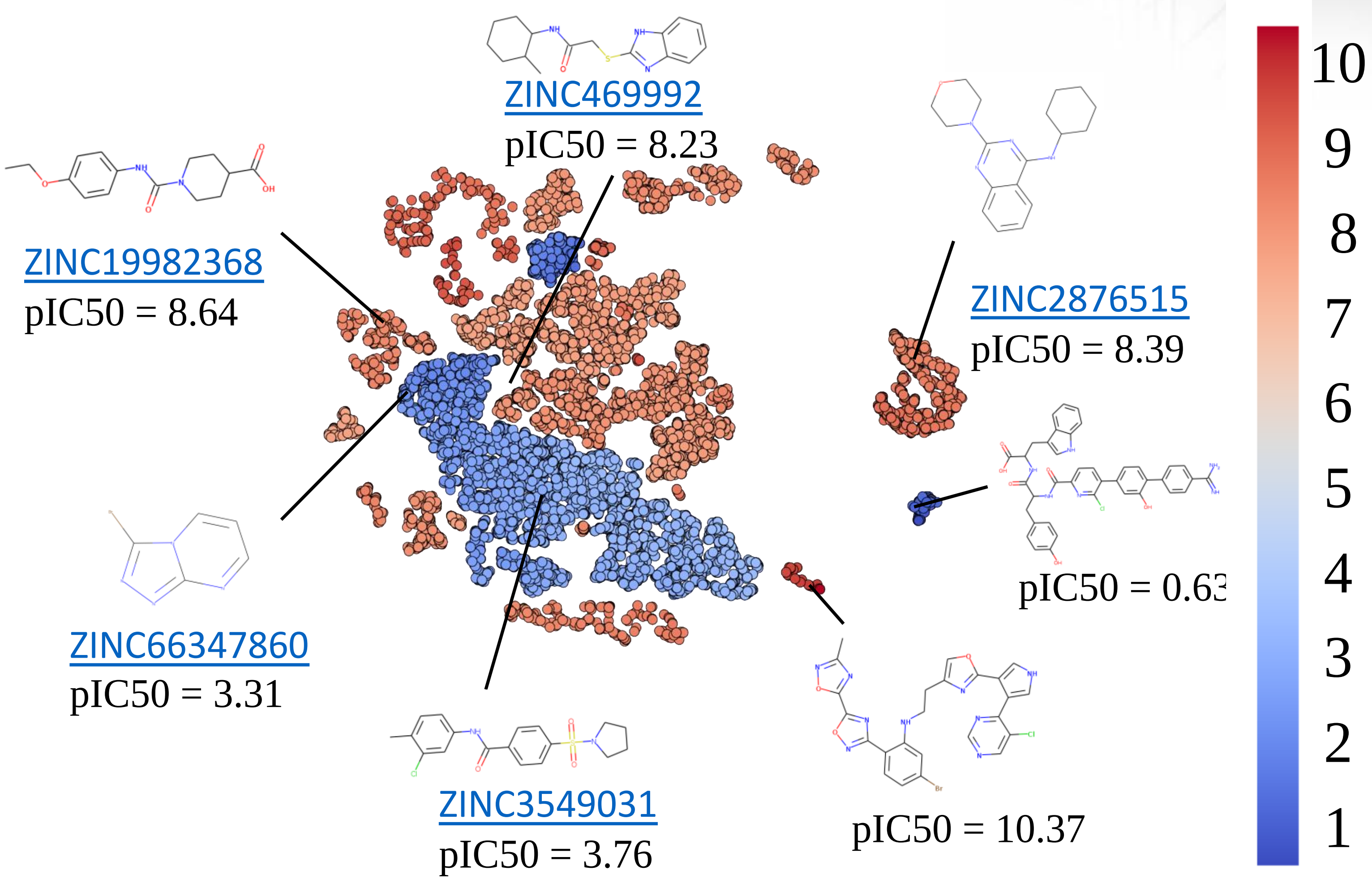
*Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotech* **25** (2), 197-206 (2007).

Results: analysis of similarity

Distribution of Tanimoto similarity to the nearest neighbor in training dataset for compounds predicted to be active for EGFR by consensus of QSAR models:



Model visualization for putative JAK2 inhibitors (projection using t-SNE)



ZINC469992
pIC50 = 8.23

ZINC19982368
pIC50 = 8.64

ZINC2876515
pIC50 = 8.39

pIC50 = 0.63

ZINC66347860
pIC50 = 3.31

ZINC3549031
pIC50 = 3.76

pIC50 = 10.37

Summary

- AI methods coupled with SMILES representation (only!) afford biased library generation
- The system naturally embeds reinforcement learning to produce novel structures with the desired property
- The system can be tuned to bias libraries towards specific property ranges

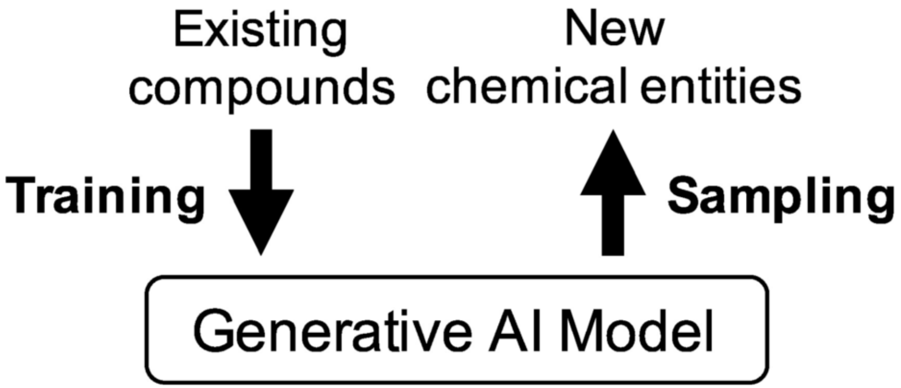
- Next phase is experimental validation of hits

Summary of recent AI-based studies on chemical library design

Molecular representations	Generative models	Method of biasing generated compounds
• Fingerprints • SMILES • Graphs	• Autoencoders • Generative adversarial models • Recurrent neural networks • Convolutional neural networks	• None • Latent space optimization • Fine-tuning on small subset of molecules with the desired property • Reinforcement Learning

An example of experimental validation of AI-based models*

- First training on large dataset
- Then fine-tuning on small subset of active compounds
- “These observations corroborate the ability of the generative AI model to **produce novel chemical entities within the training data domain**”.



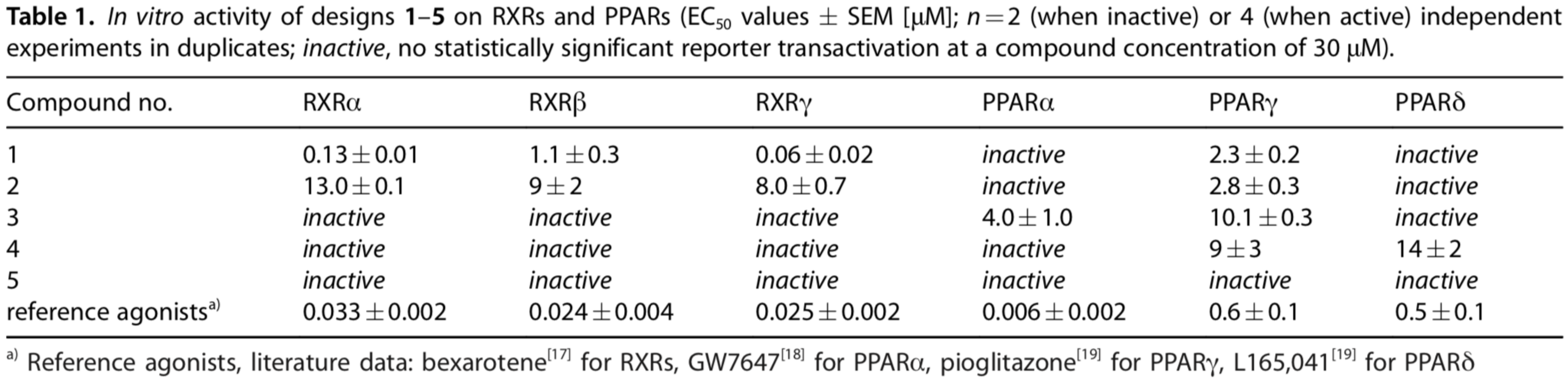


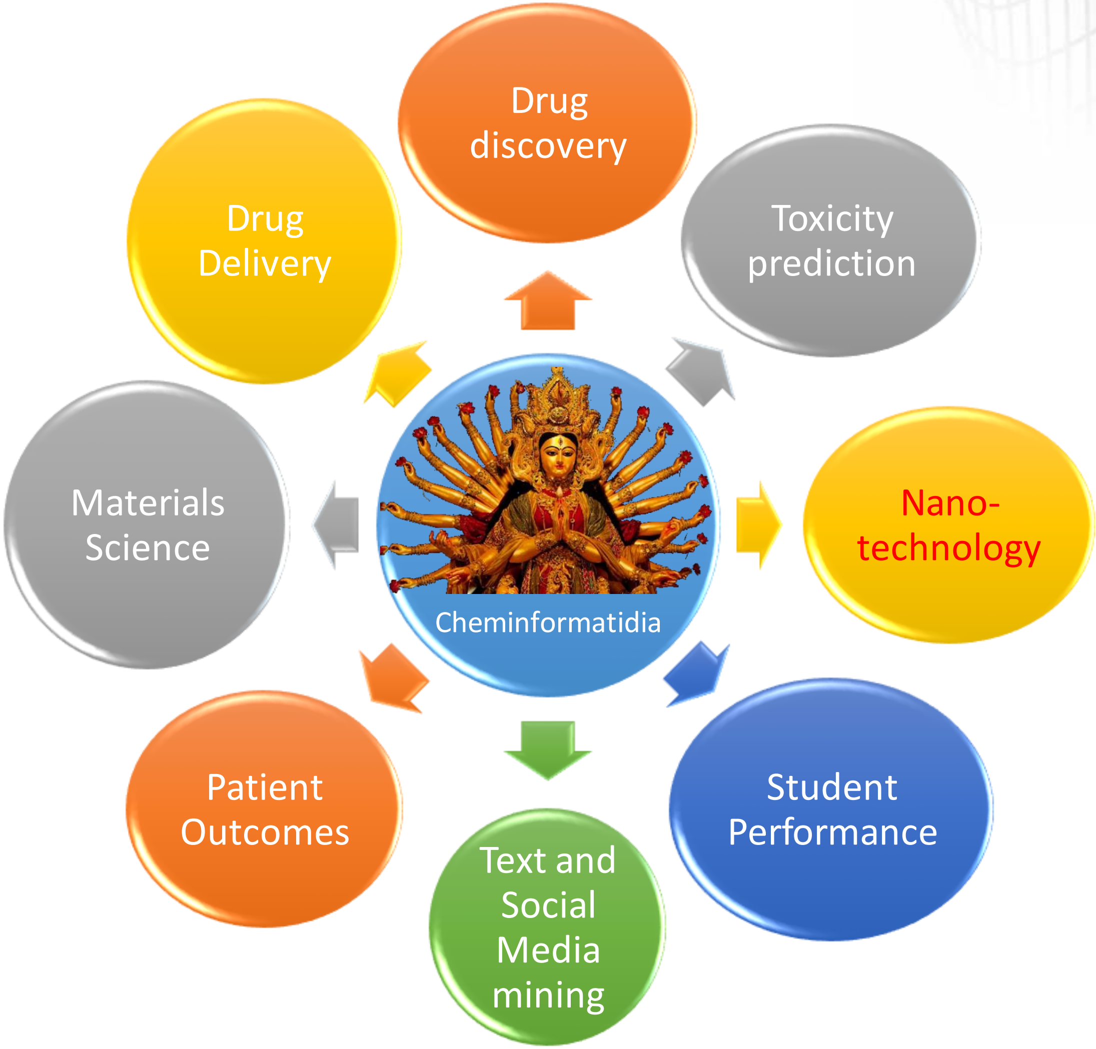
Table 1. *In vitro* activity of designs **1–5** on RXRs and PPARs (EC_{50} values ± SEM [µM]; $n = 2$ (when inactive) or 4 (when active) independent experiments in duplicates; *inactive*, no statistically significant reporter transactivation at a compound concentration of 30 µM).

Compound no.	RXRα	RXRβ	RXRγ	PPARα	PPARγ	PPARδ
1	0.13 ± 0.01	1.1 ± 0.3	0.06 ± 0.02	*inactive*	2.3 ± 0.2	*inactive*
2	13.0 ± 0.1	9 ± 2	8.0 ± 0.7	*inactive*	2.8 ± 0.3	*inactive*
3	*inactive*	*inactive*	*inactive*	4.0 ± 1.0	10.1 ± 0.3	*inactive*
4	*inactive*	*inactive*	*inactive*	*inactive*	9 ± 3	14 ± 2
5	*inactive*	*inactive*	*inactive*	*inactive*	*inactive*	*inactive*
reference agonists[a)]	0.033 ± 0.002	0.024 ± 0.004	0.025 ± 0.002	0.006 ± 0.002	0.6 ± 0.1	0.5 ± 0.1

[a)] Reference agonists, literature data: bexarotene[17] for RXRs, GW7647[18] for PPARα, pioglitazone[19] for PPARγ, L165,041[19] for PPARδ

* D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.

Many virtues of Cheminformatics



Drug discovery

Toxicity prediction

Nano-technology

Student Performance

Text and Social Media mining

Patient Outcomes

Materials Science

Drug Delivery

Cheminformatidia

Acknowledgements

Principal Investigator

Alexander Tropsha

Research Professors

Alexander Golbraikh
Olexander Isayev
Eugene Muratov

Postdoctoral Fellows

Vinicius Alves
Stephen Capuzzi
Joyce Borba

Graduate students

Sherif Faraq
Kyle Bowers
Maria Popova
Andrew Thieme

- **Duke University**
  - Stefano Curtarolo
  - Corey Oses
- **UNC Chemistry**
  - Jim Cahoon
  - Taylor Moot
  - Aaron Taggart

MAJOR FUNDING

NIH

- 1U01CA207160
- R01-GM114015
- 5U54CA198999
- 1OT3TR002020

ONR

- N00014-16-1-2311