Cambridge Healthtech Institute's 14th Annual
***GPCR-Based Drug Discovery — Discovery on Target* Conference**

September 19, 2019 | Boston, MA

https://www.discoveryontarget.com/GPCR-drug-discovery

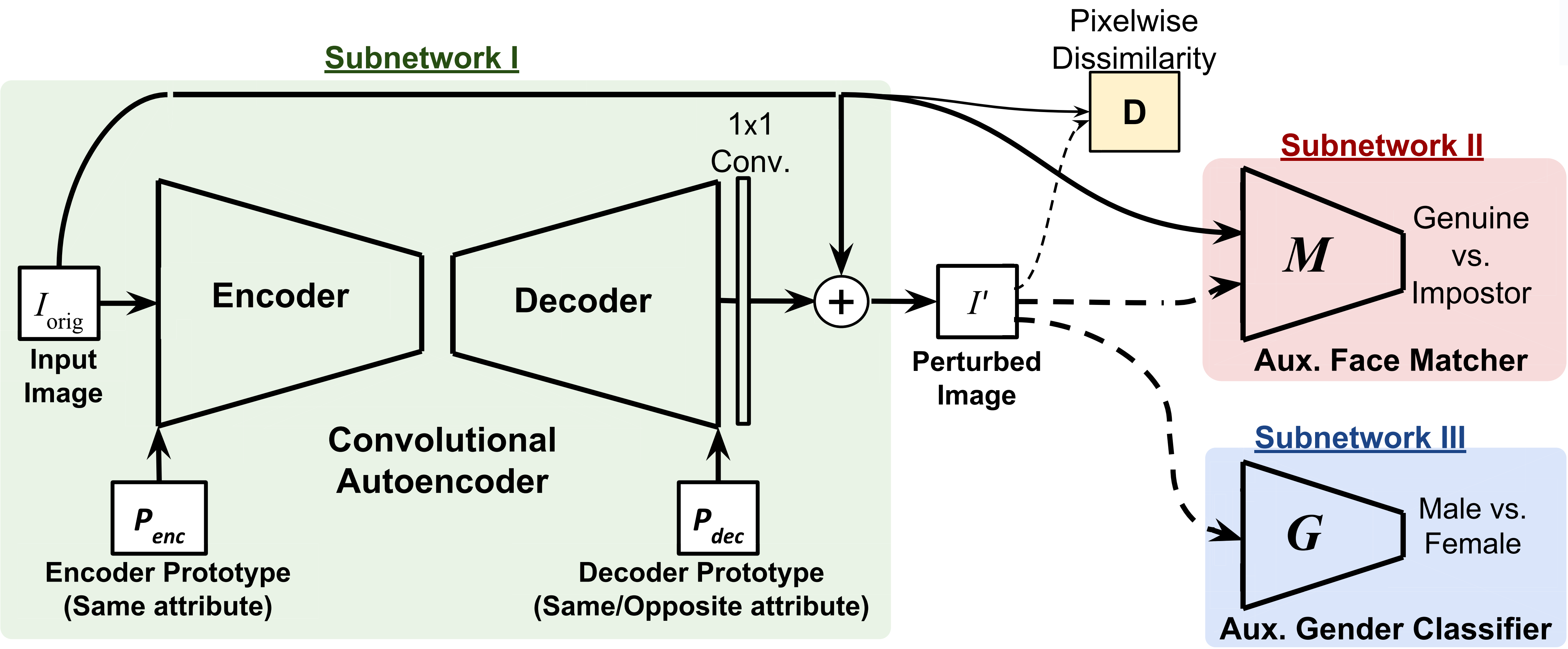
Machine-Learning & AI-based Approaches for GPCR Bioactive Ligand Discovery

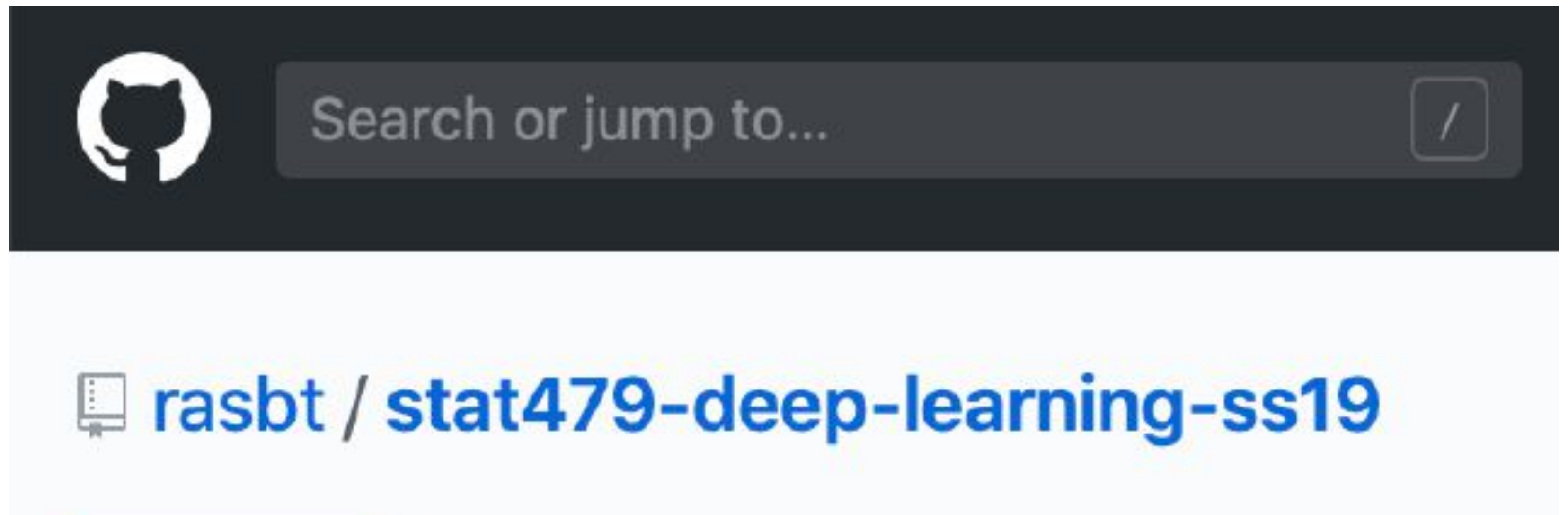
Sebastian Raschka, Ph.D.
Assistant Professor
Department of Statistics

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

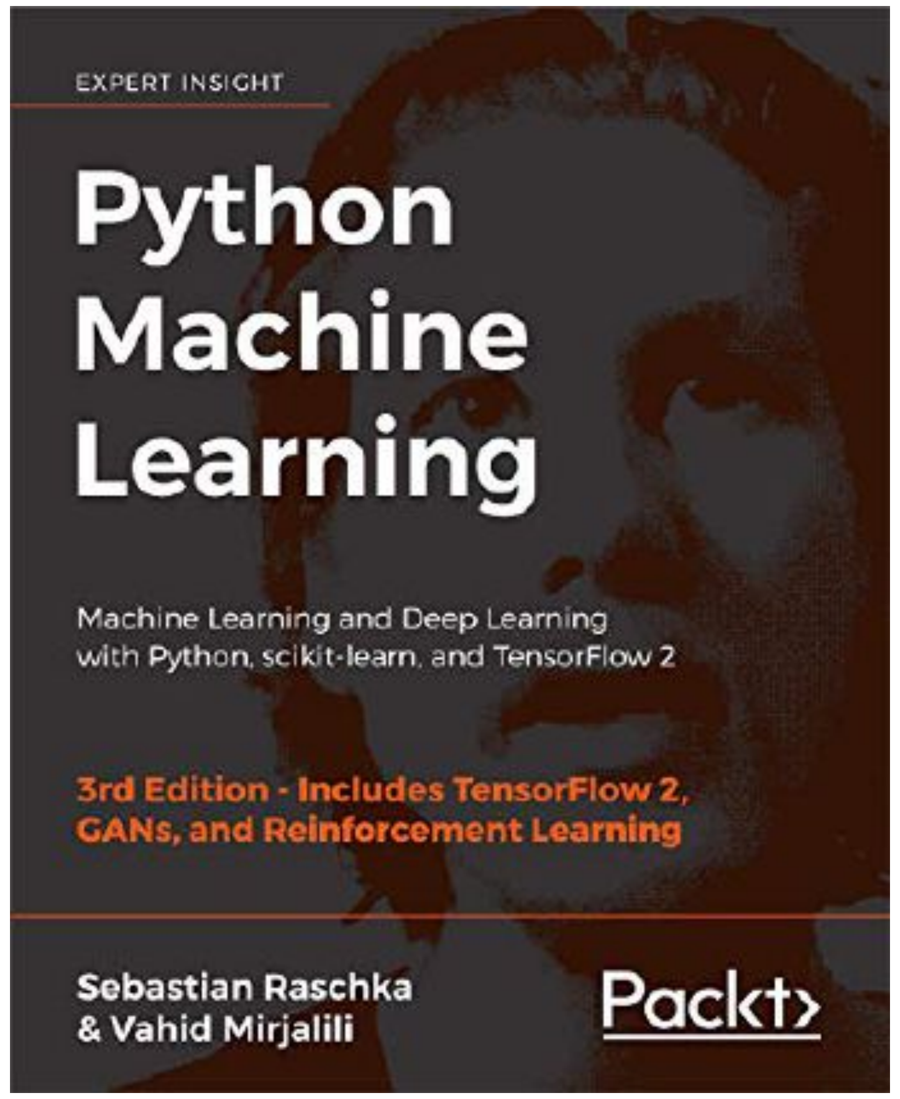
sraschka@wisc.edu
http://stat.wisc.edu/~sraschka/

My background and interests




Search or jump to...

rasbt / stat479-deep-learning-ss19


EXPERT INSIGHT

Python Machine Learning

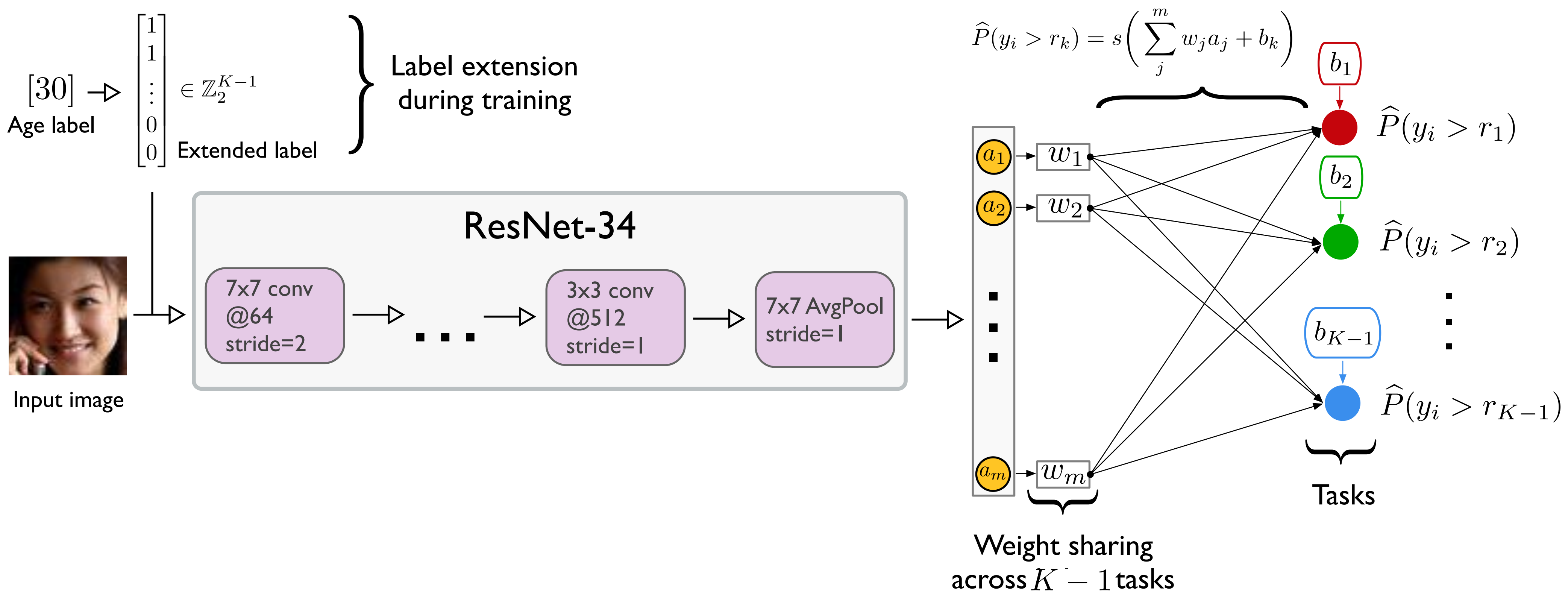
Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2

3rd Edition - Includes TensorFlow 2, GANs, and Reinforcement Learning

Sebastian Raschka & Vahid Mirjalili

Packt>

BioPandas

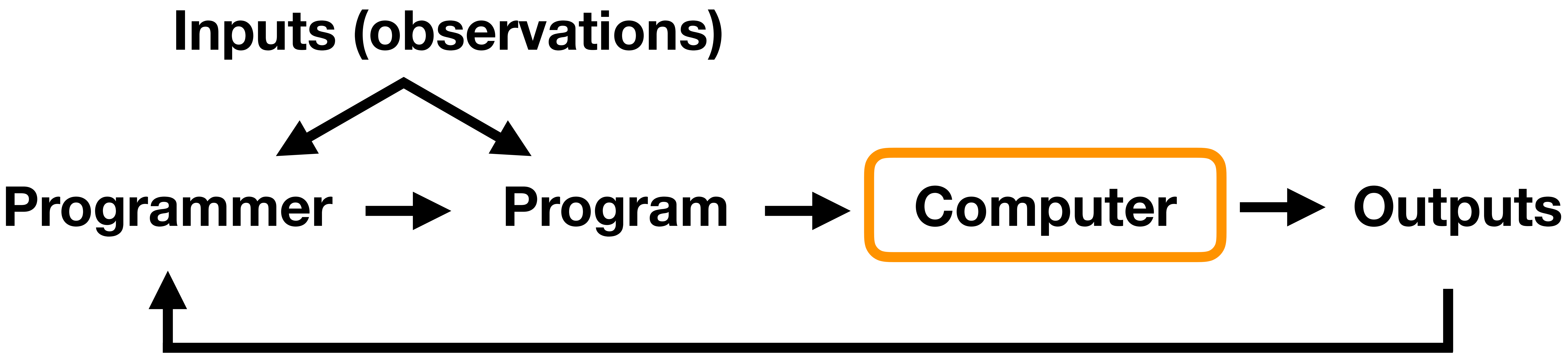


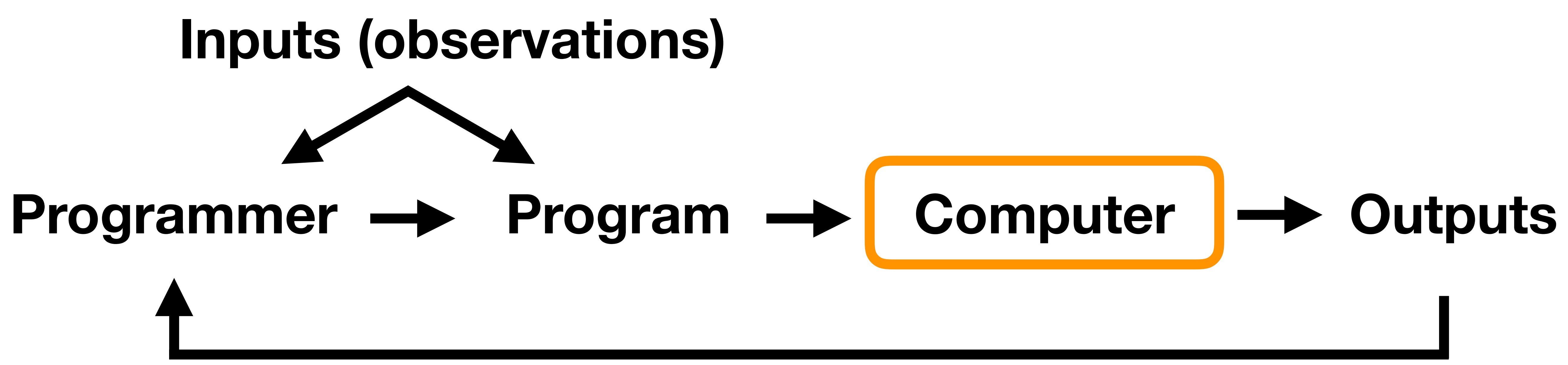
Site Interlock

HBIND

screenlamp

The Traditional Programming Paradigm



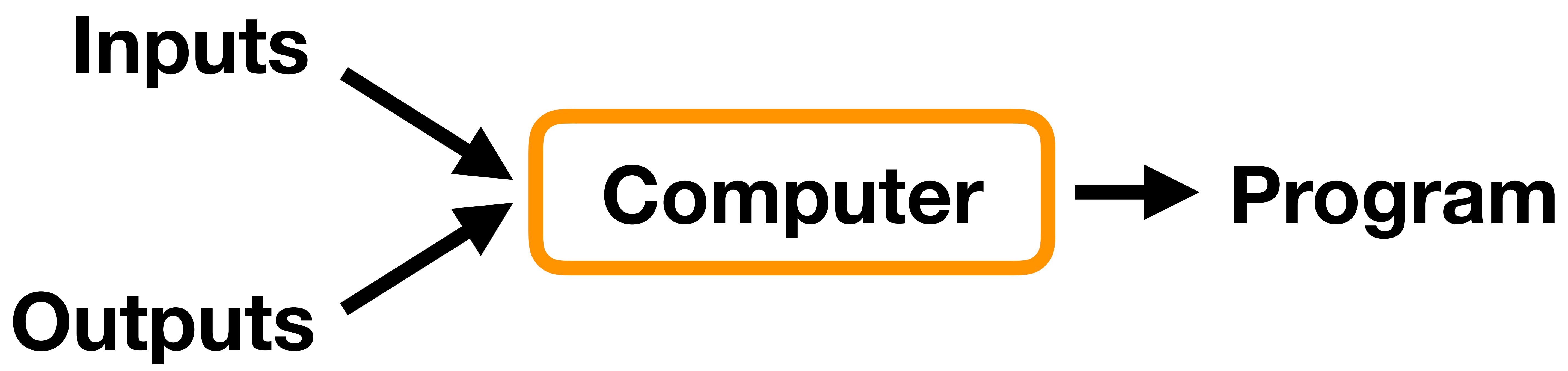


Inputs (observations)

Programmer → Program → Computer → Outputs

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed

— *Arthur Samuel (1959)*

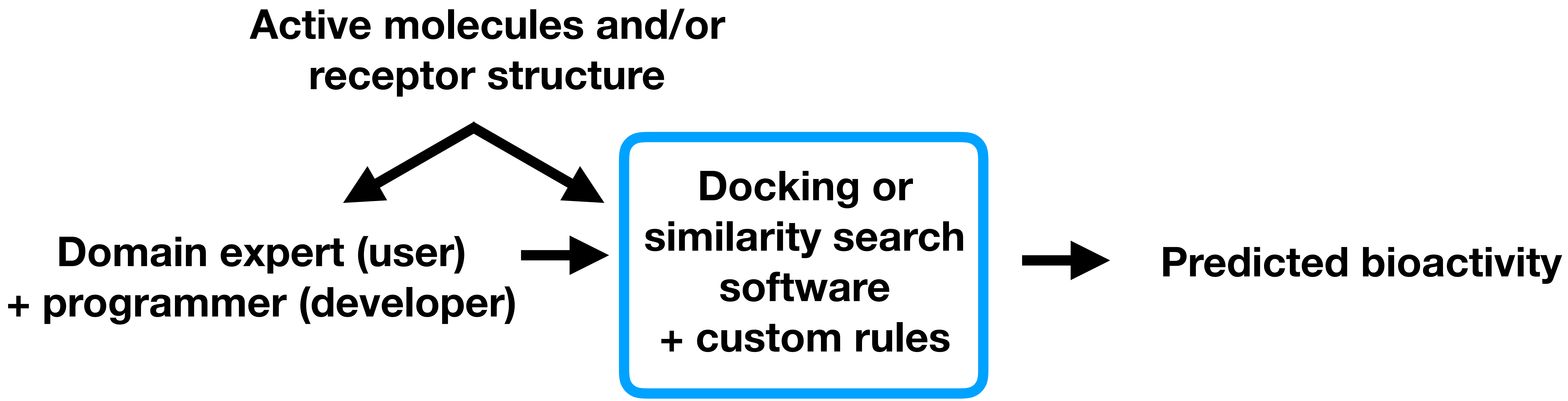


Inputs

Outputs

Computer → Program

The Traditional Ligand Discovery Paradigm



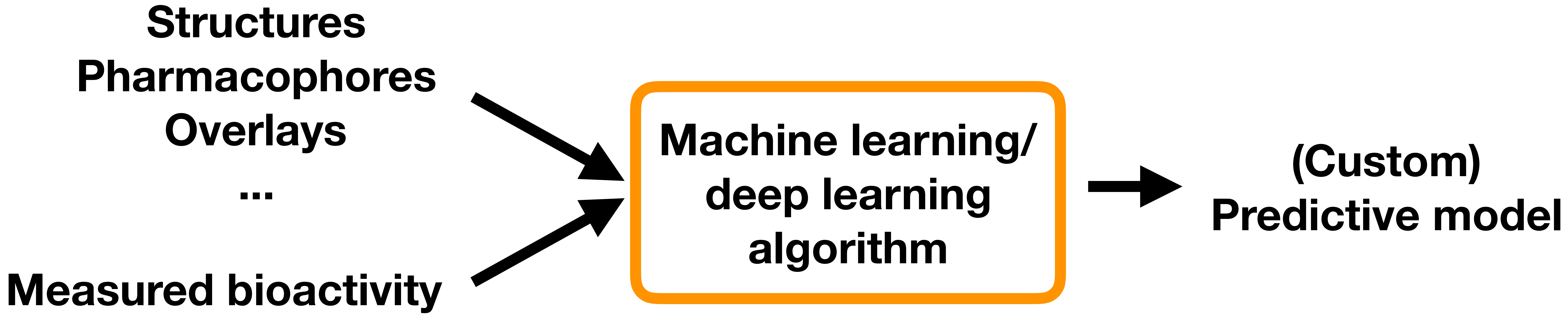
Active molecules and/or receptor structure

Domain expert (user)
+ programmer (developer)

Docking or similarity search software
+ custom rules

Predicted bioactivity

Machine Learning-augmented Ligand Discovery Paradigm



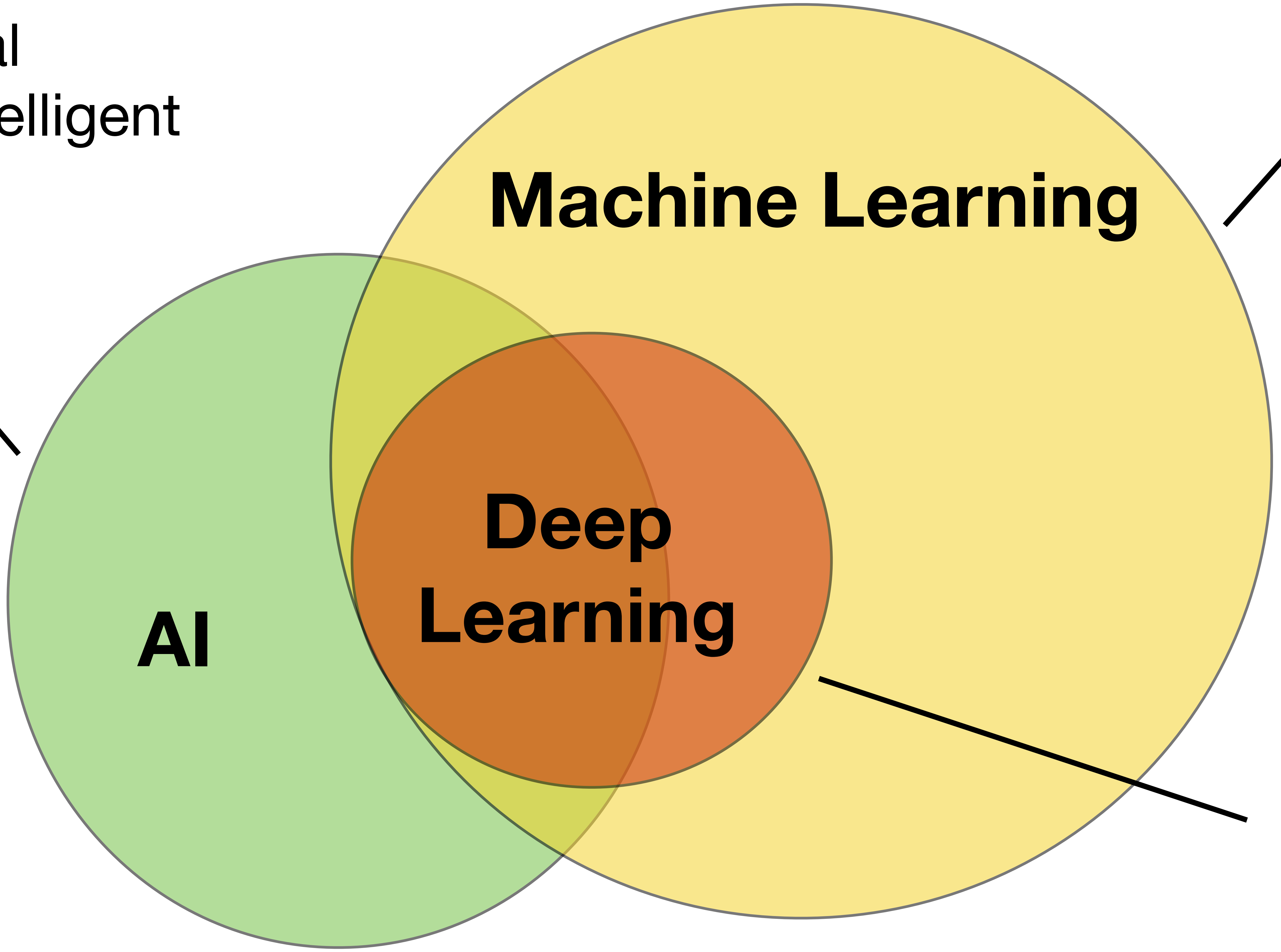
Structures
Pharmacophores
Overlays
...

Measured bioactivity

Machine learning/ deep learning algorithm

(Custom)
Predictive model

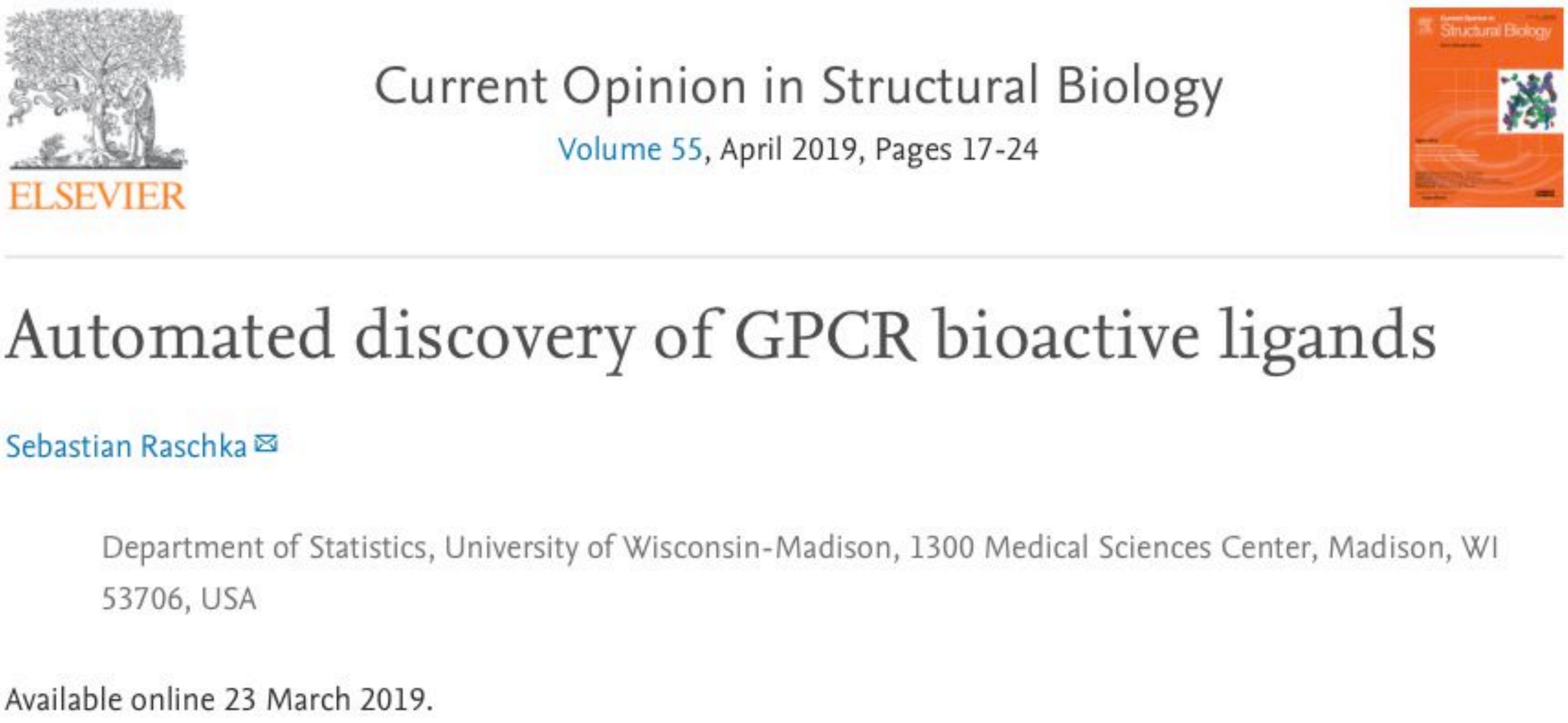
The Connection Between Fields



= a non-biological system that is intelligent through rules

= algorithms that learn models/representations/ rules automatically from data/examples

= algorithms that parameterize multi-layer neural networks that then learn representations of data with multiple layers of abstraction



ELSEVIER

Current Opinion in Structural Biology

Volume 55, April 2019, Pages 17-24

Automated discovery of GPCR bioactive ligands

Sebastian Raschka

Department of Statistics, University of Wisconsin-Madison, 1300 Medical Sciences Center, Madison, WI 53706, USA

Available online 23 March 2019.

https://doi.org/10.1016/j.sbi.2019.02.011

1 Training
Test Set Molecules
New Molecules
Feature Extraction
Molecules
Machine Learning Algorithm
Model
3
Inference
Experimental Activity Data
Experimental Activity Data (Test Set)
Predicted Activity Data
Validation 2
Experimental Assays

S. Raschka (2019) *Automated discovery of GPCR bioactive ligands*. Current Opinion in Structural Biology 2019, 55:17–24

https://www.sciencedirect.com/science/article/abs/pii/S0959440X18301362

1 Training
Test Set Molecules
New Molecules
Feature Extraction
Molecules
Machine Learning Algorithm
Model
3
Inference
Experimental Activity Data
Experimental Activity Data (Test Set)
Predicted Activity Data
Iterative process
Validation 2
Experimental Assays

S. Raschka (2019) *Automated discovery of GPCR bioactive ligands*. Current Opinion in Structural Biology 2019, 55:17–24

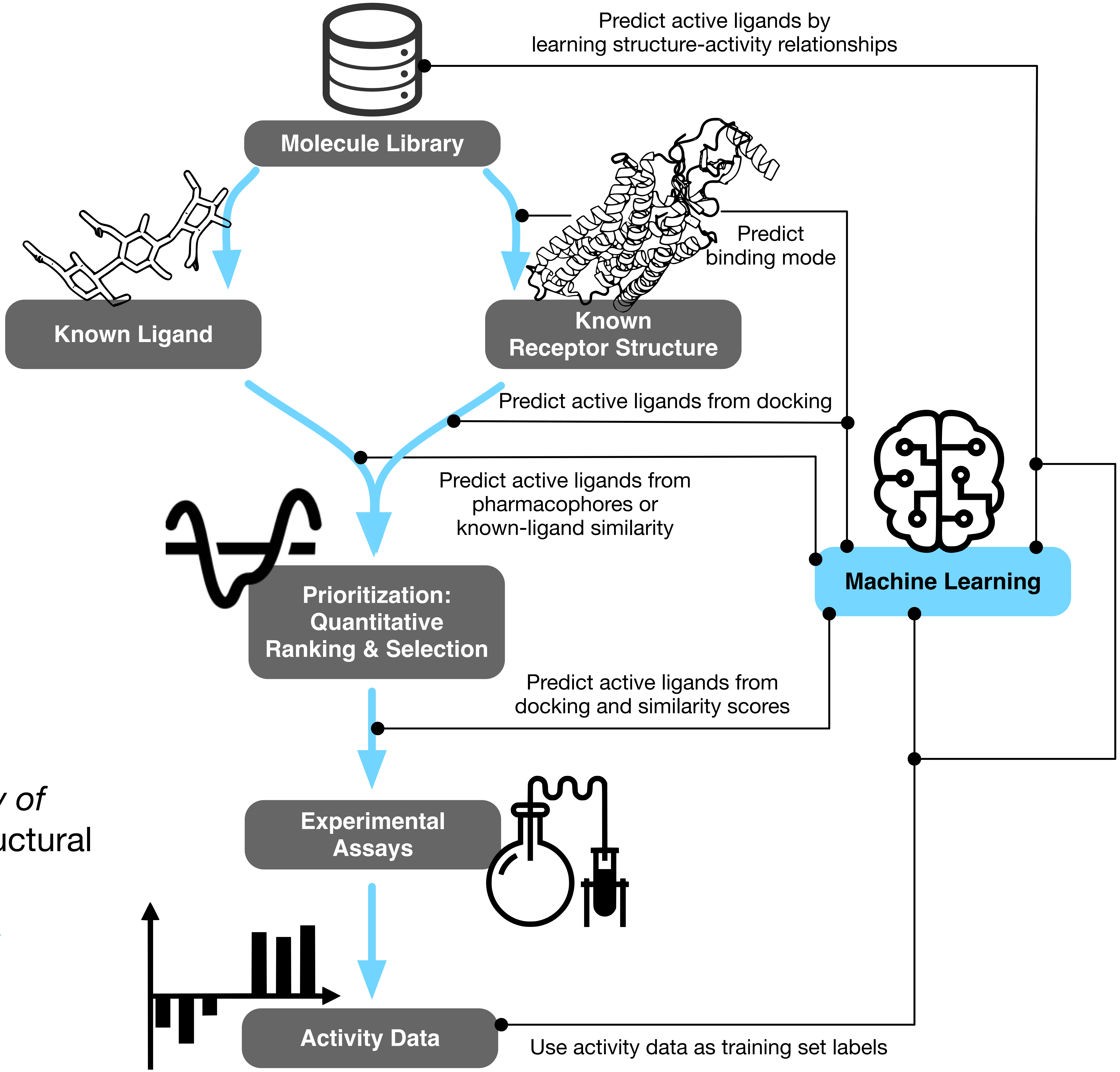
https://www.sciencedirect.com/science/article/abs/pii/S0959440X18301362

- ML particularly attractive as activity data become available after initial rounds of screening and assaying

- Use ML to guide further rounds of screening and experimental testing

Sebastian Raschka (2019) *Automated discovery of GPCR bioactive ligands*. Current Opinion in Structural Biology 2019, 55:17–24

https://www.sciencedirect.com/science/article/abs/pii/S0959440X18301362


Predict active ligands by learning structure-activity relationships
Molecule Library
Predict binding mode
Known Ligand
Known Receptor Structure
Predict active ligands from docking
Predict active ligands from pharmacophores or known-ligand similarity
Machine Learning
Prioritization: Quantitative Ranking & Selection
Predict active ligands from docking and similarity scores
Experimental Assays
Activity Data
Use activity data as training set labels

Case study 1

GPCR inhibitor discovery for invasive species control

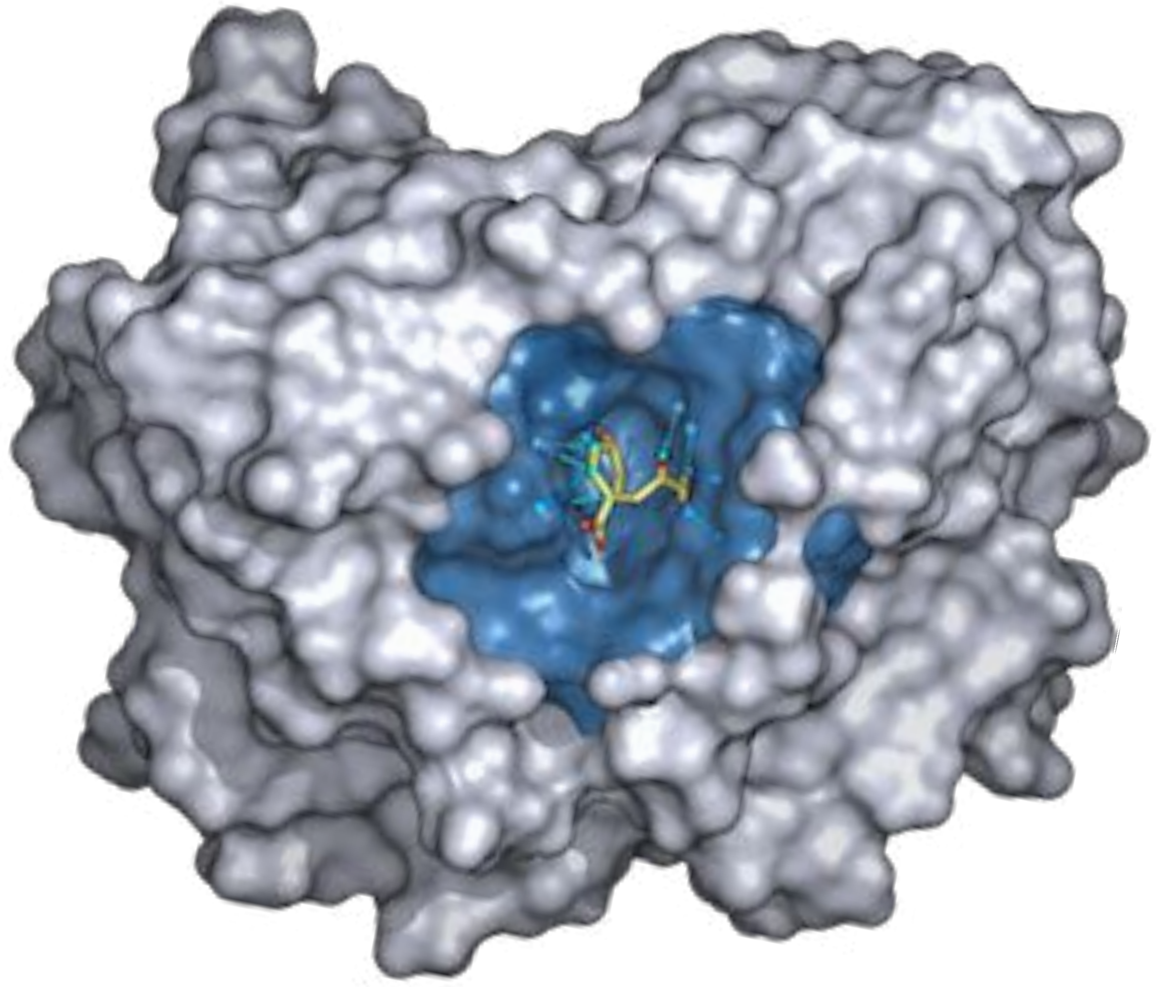
Identifying a pheromone inhibitor in low nanomolar concentration

Discovery of a pheromone receptor inhibitor for invasive species control (sea lamprey) in the Great Lakes

https://en.wikipedia.org/wiki/Sea_lamprey#/media/File:Sea_lamprey_on_brown_trout_flipped.jpg

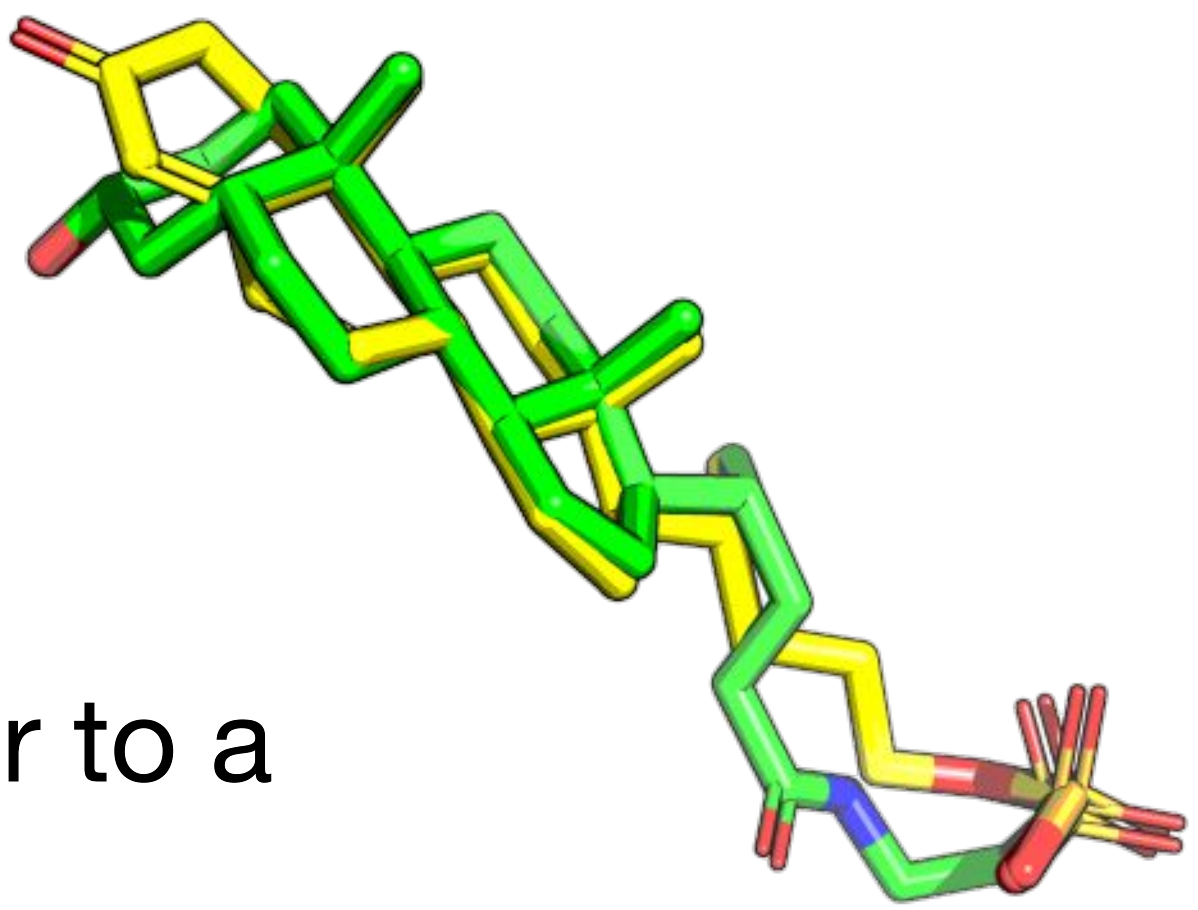
Virtual screening

Receptor structure-based



Small molecule-based

Assuming molecules similar to a
known binder are also
likely to bind the target receptor



Pheromone
(@ 10^{-12} M)

+

Antagonist Discovered
(@ $5x10^{-13}$ M)

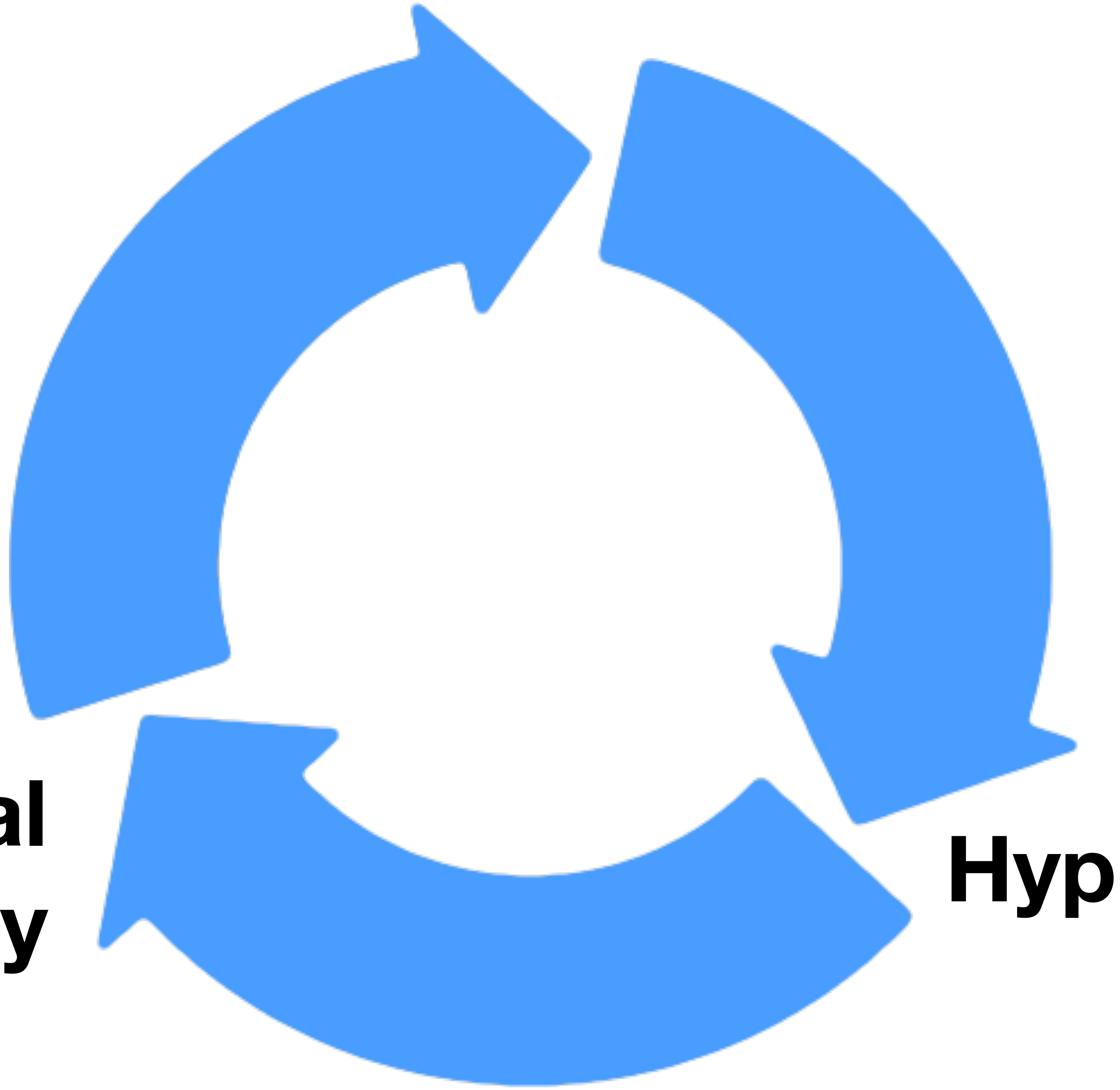
Pheromone
(@ 10^{-12} M)

Hypothesis-based Filtering

Millions of molecules

Small number of (potentially) active molecule

Machine learning



Experimental assay

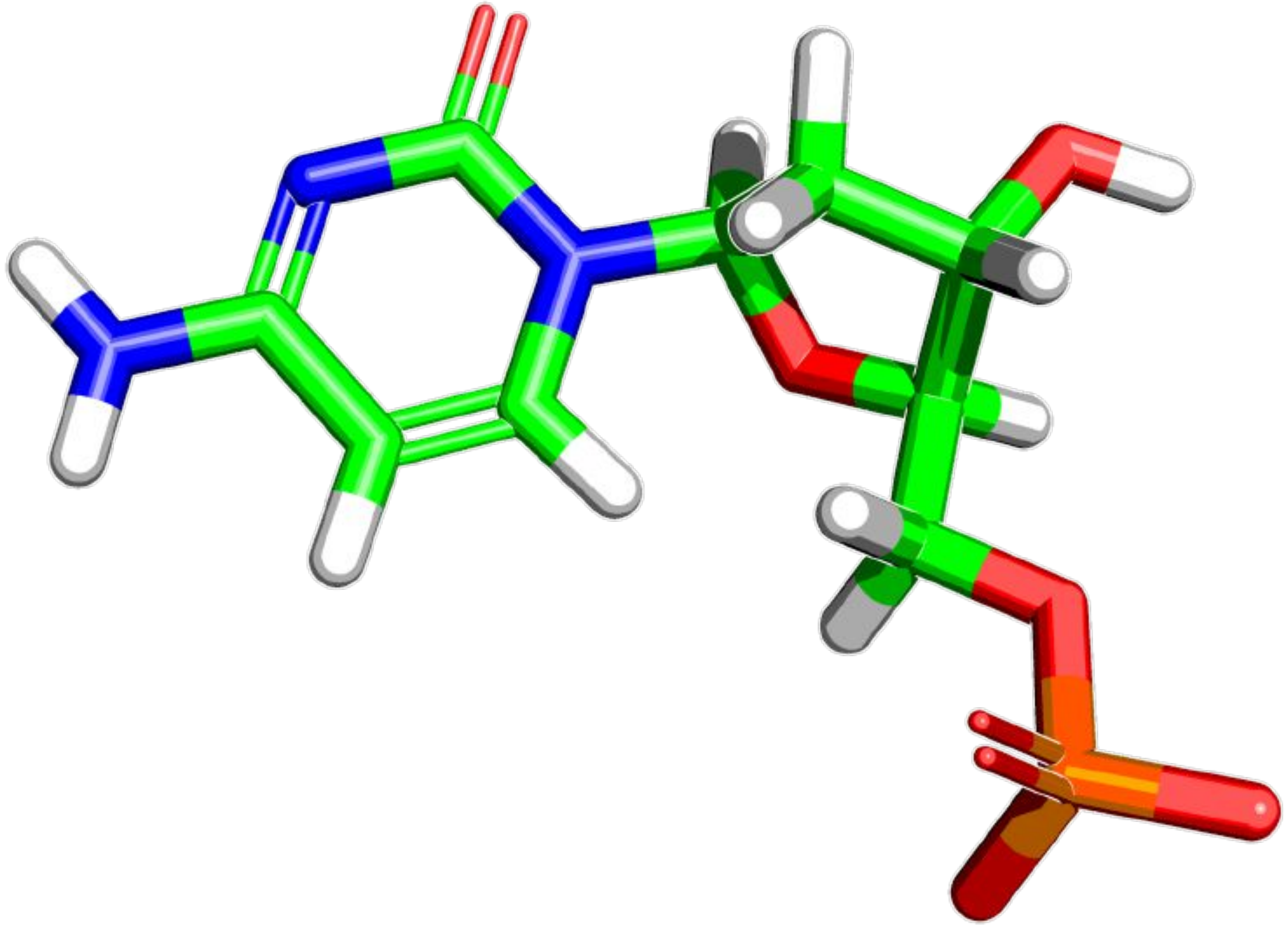
Hypothesis

screenlamp

S. Raschka, N. Liu, S. Gunturu, A.M. Scott, M. Huertas, W. Li, and L.A. Kuhn (2018)
Facilitating the hypothesis-driven prioritization of small molecules in large databases: Screenlamp and its application to GPCR inhibitor discovery.
Journal of Computer-Aided Molecular Design, 32(3), 415-433.

https://psa-lab.github.io/screenlamp

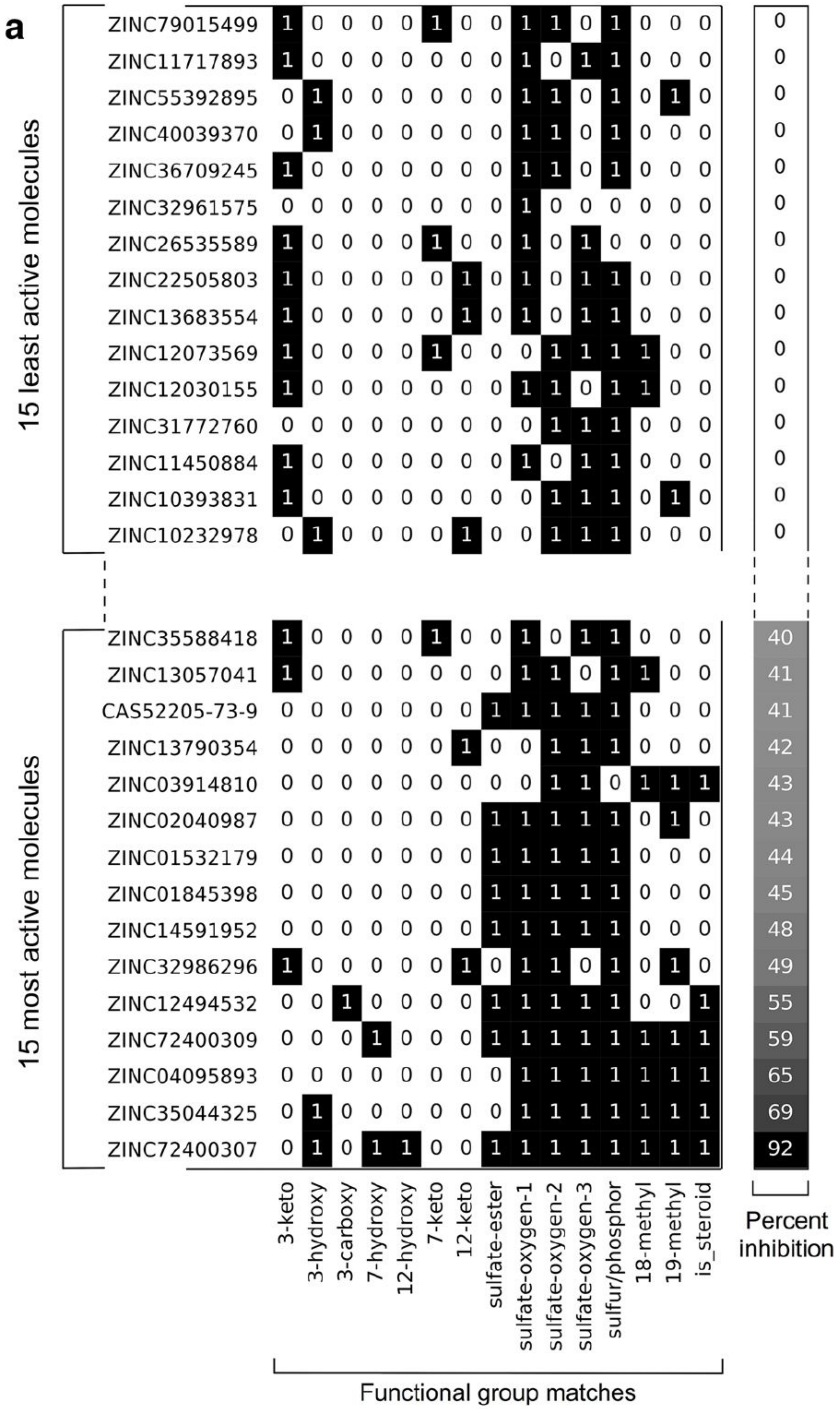
BioPandas



Sebastian Raschka (2017) *BioPandas: Working with molecular structures in Pandas DataFrames*. The Journal of Open Source Software 2.14.

http://rasbt.github.io/biopandas/

Tabulating functional group matches (via screenlamp) from 3D volumetric and electrostatic (via OpenEye ROCS) with a known bioactive molecule

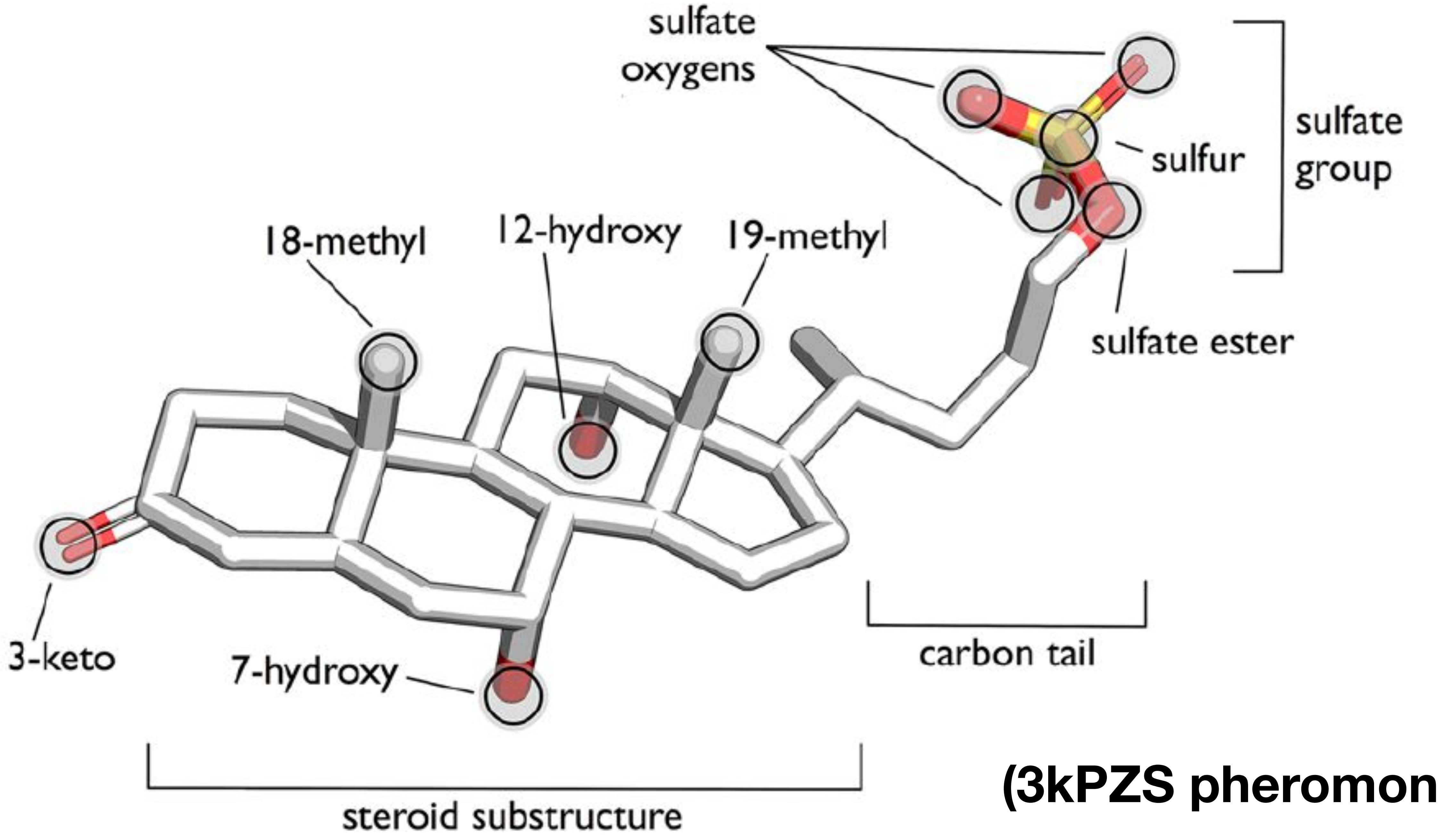


a

		3-keto	3-hydroxy	3-carboxy	7-hydroxy	12-hydroxy	7-keto	12-keto	sulfate-ester	sulfate-oxygen-1	sulfate-oxygen-2	sulfate-oxygen-3	sulfur/phosphor	18-methyl	19-methyl	is_steroid	Percent inhibition
15 least active molecules	ZINC79015499	1	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0
	ZINC11717893	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
	ZINC55392895	0	1	0	0	0	0	0	0	1	1	0	1	0	1	0	0
	ZINC40039370	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0
	ZINC36709245	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
	ZINC32961575	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
	ZINC26535589	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
	ZINC22505803	1	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0
	ZINC13683554	1	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0
	ZINC12073569	1	0	0	0	0	1	0	0	0	1	1	1	1	0	0	0
	ZINC12030155	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0
	ZINC31772760	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0
	ZINC11450884	1	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
	ZINC10393831	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0
	ZINC10232978	0	1	0	0	0	0	1	0	0	1	1	1	0	0	0	0
15 most active molecules	ZINC35588418	1	0	0	0	0	1	0	0	1	0	1	1	0	0	0	40
	ZINC13057041	1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	41
	CAS52205-73-9	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	41
	ZINC13790354	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	42
	ZINC03914810	0	0	0	0	0	0	0	0	0	1	1	0	1	1	1	43
	ZINC02040987	0	0	0	0	0	0	0	1	1	1	1	1	0	1	0	43
	ZINC01532179	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	44
	ZINC01845398	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	45
	ZINC14591952	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	48
	ZINC32986296	1	0	0	0	0	0	1	0	1	1	0	1	0	1	0	49
	ZINC12494532	0	0	1	0	0	0	0	1	1	1	1	1	0	0	1	55
	ZINC72400309	0	0	0	1	0	0	0	1	1	1	1	1	1	1	1	59
	ZINC04095893	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	65
	ZINC35044325	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	69
	ZINC72400307	0	1	0	1	1	0	0	1	1	1	1	1	1	1	1	92

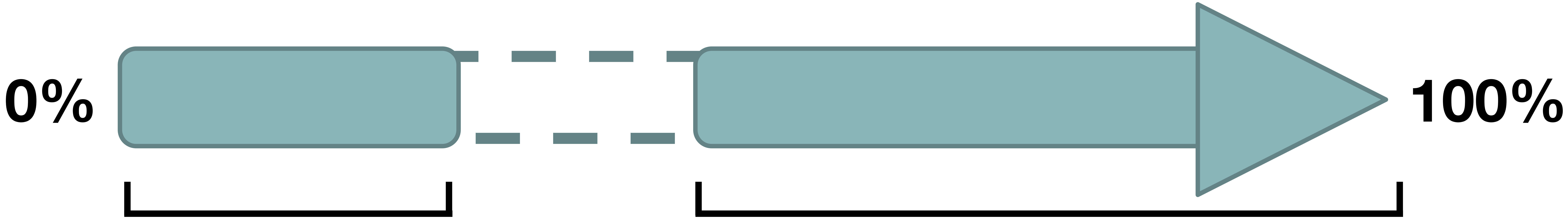
Functional group matches

Assay data



(3kPZS pheromone)

Thresholding Assay Data



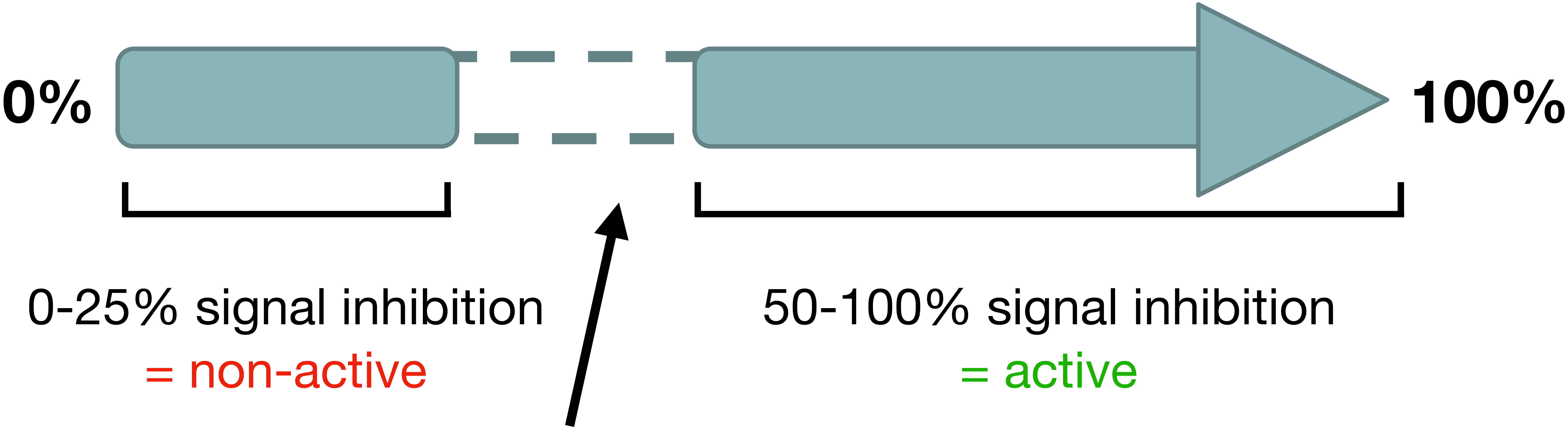
0%

100%

0-25% signal inhibition
= non-active

50-100% signal inhibition
= active

Thresholding Assay Data

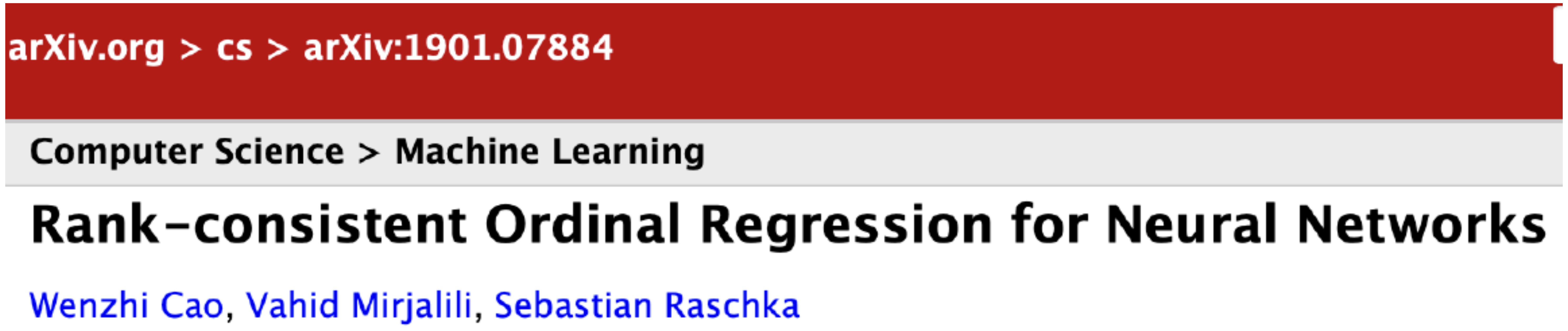


0%

100%

0-25% signal inhibition
= non-active

50-100% signal inhibition
= active



arXiv.org > cs > arXiv:1901.07884

Computer Science > Machine Learning

Rank-consistent Ordinal Regression for Neural Networks

Wenzhi Cao, Vahid Mirjalili, Sebastian Raschka

MLXTEND

scikit learn

`SequentialFeatureSelector`

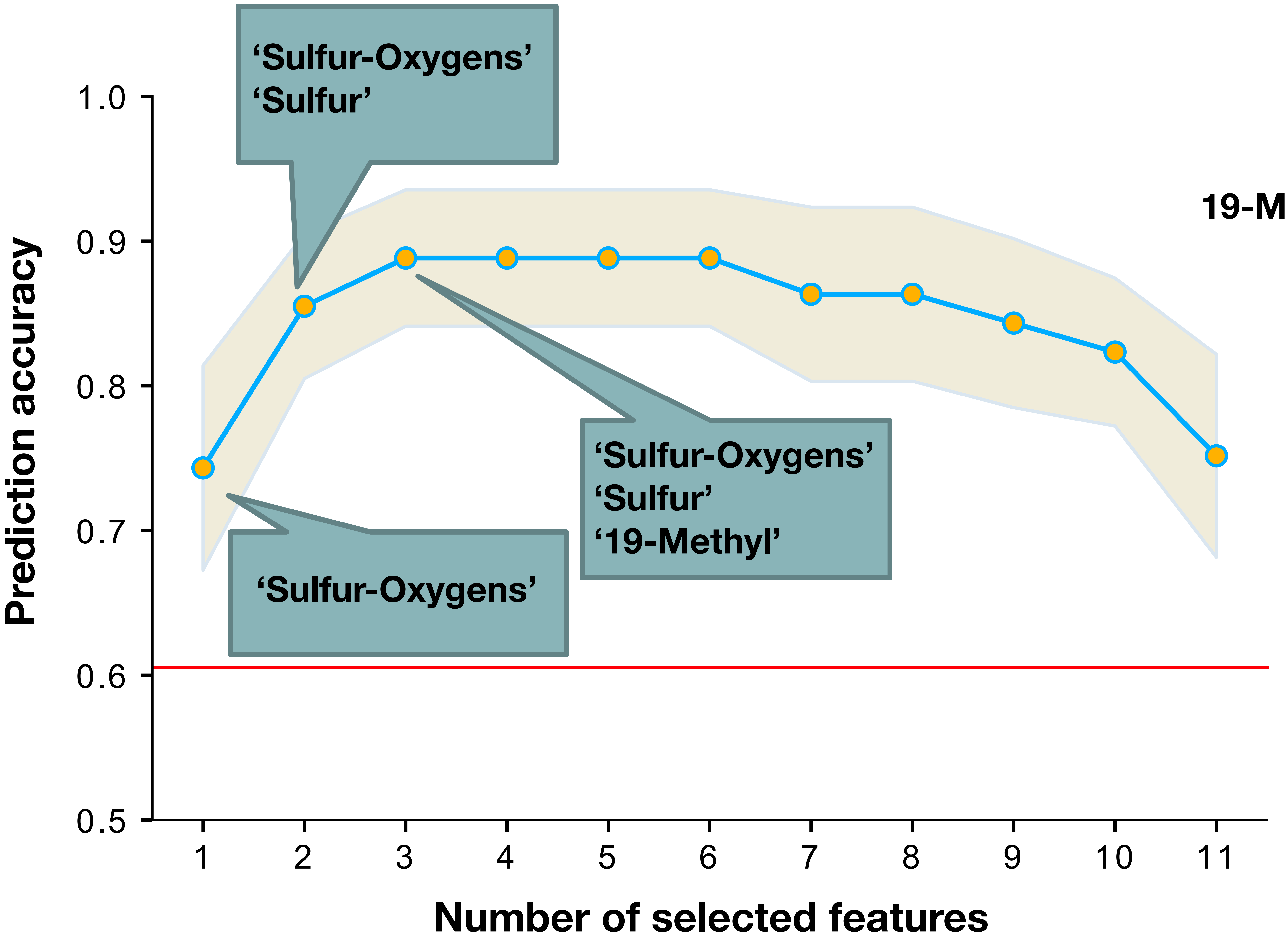
Sebastian Raschka (2018) *MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack*.
The Journal of Open Source Software 3.24.

http://rasbt.github.io/mlxtend/

`KNeighborsClassifier`

Pedregosa et al. (2011) *Scikit-learn: Machine learning in Python*.
Journal of Machine learning Research 2825-2830.

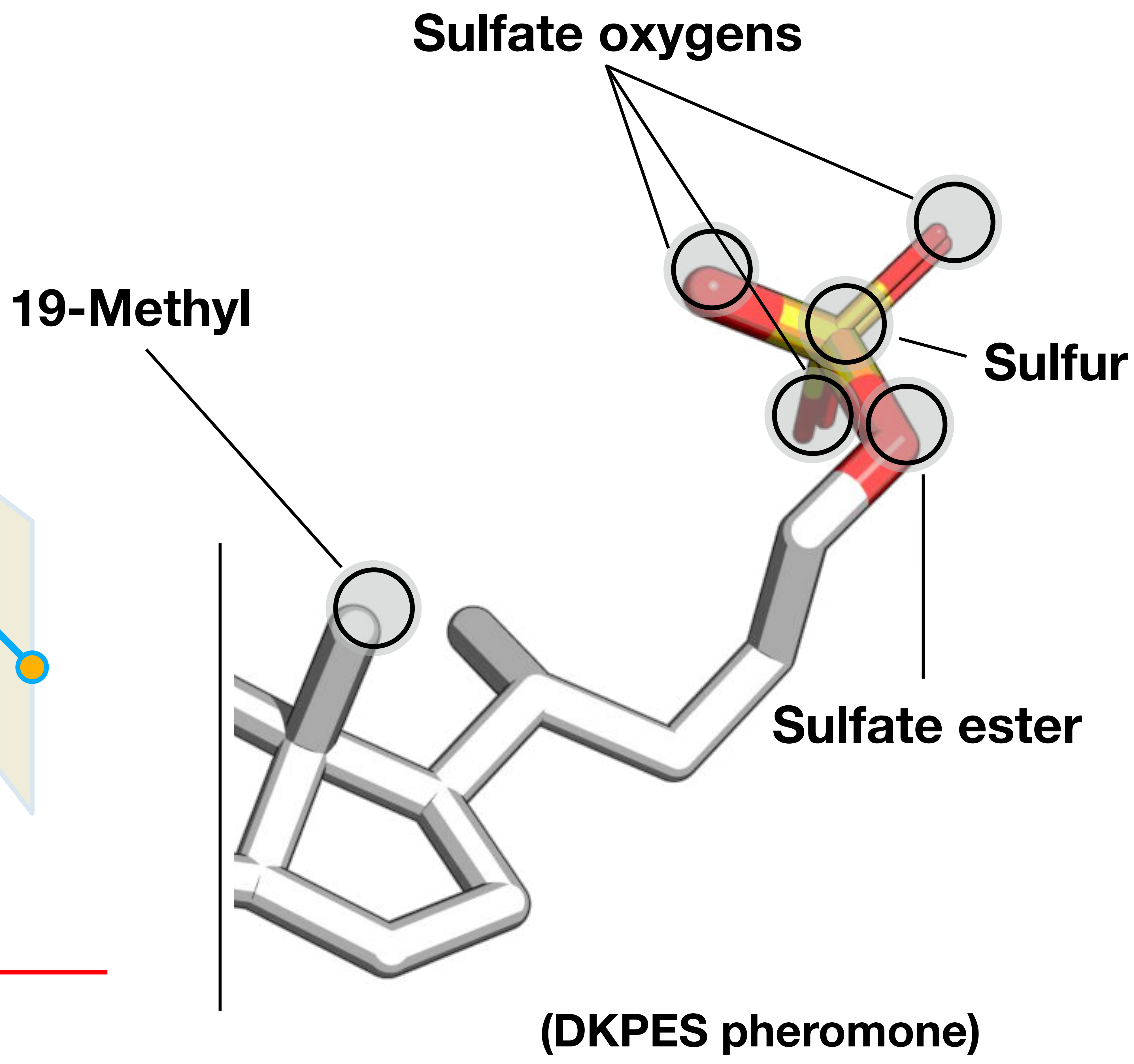
https://scikit-learn.org



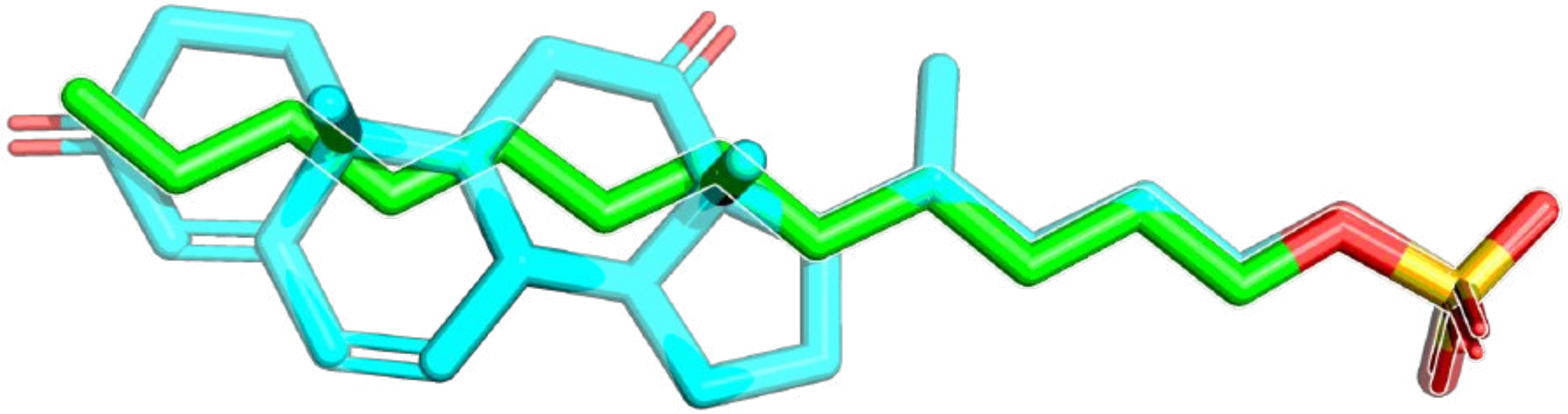
‘Sulfur-Oxygens’

‘Sulfur-Oxygens’
‘Sulfur’

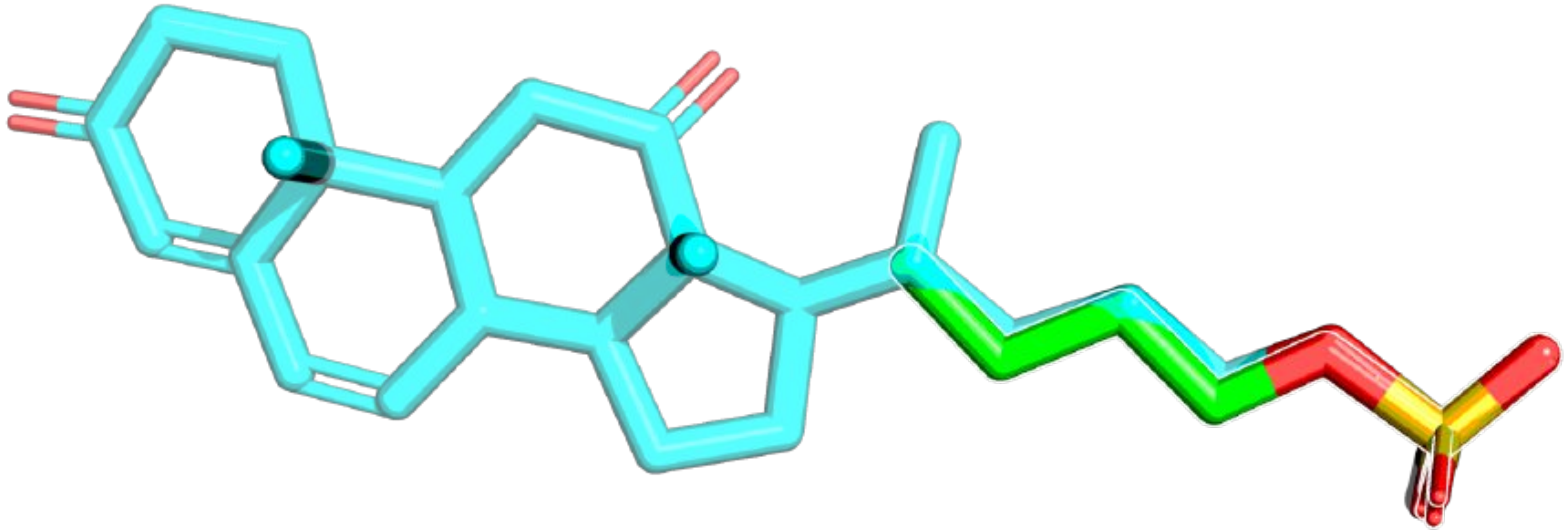
‘Sulfur-Oxygens’
‘Sulfur’
‘19-Methyl’



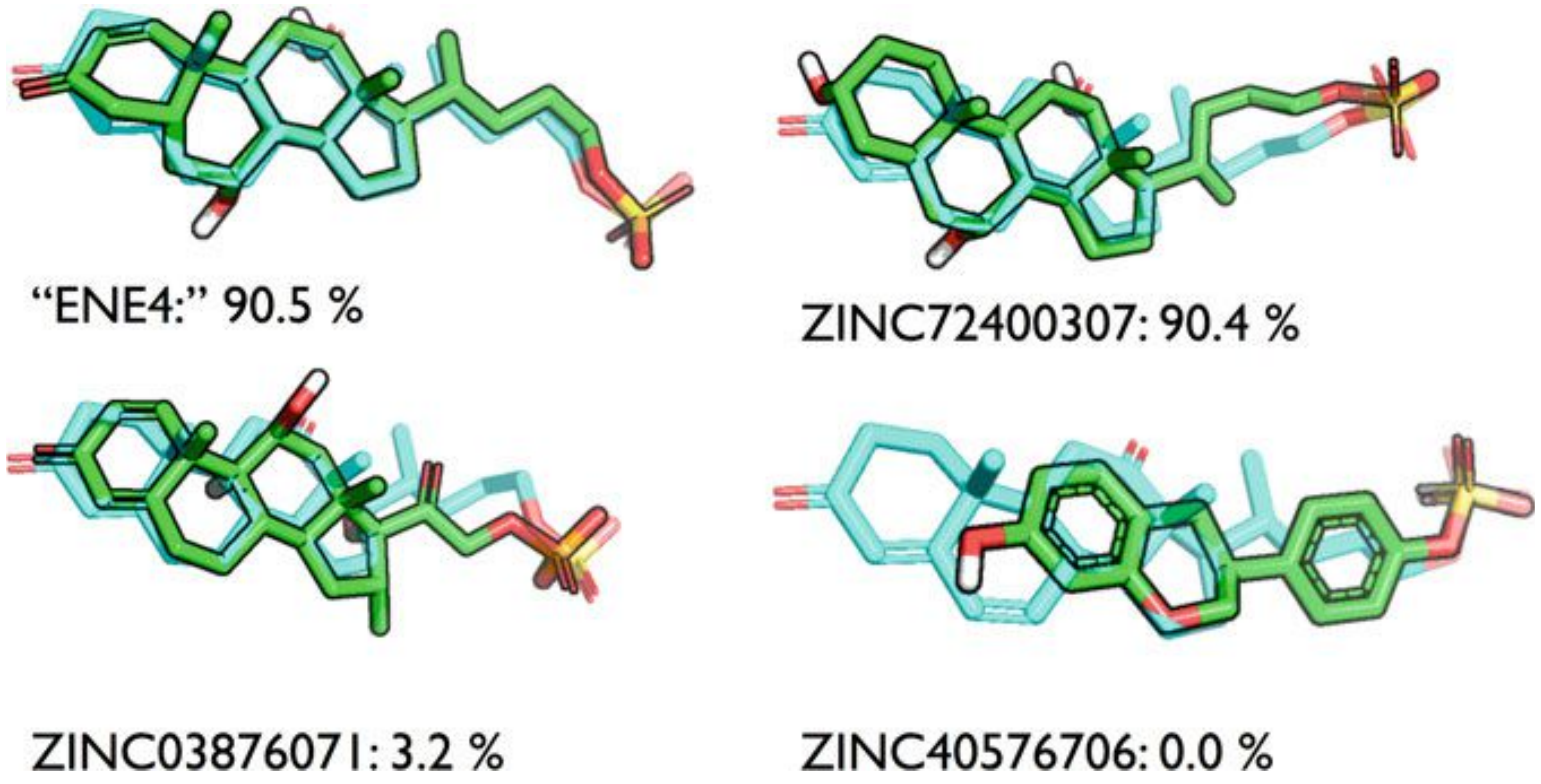
"Sulfate-tail" sufficient for bioactivity



69% signal inhibition



62% signal inhibition



"ENE4:" 90.5 %

ZINC72400307: 90.4 %

ZINC03876071: 3.2 %

ZINC40576706: 0.0 %

Fig. 5 3D structures and percent DKPES olfactory inhibition of the two most active molecules (actives, top row) and two low-activity molecules (non-actives, bottom row) from the screening set, shown in green as overlayed with the best-matching DKPES 3D conformer (cyan)

Sebastian Raschka, Leslie A. Kuhn, Anne M. Scott, and Weiming Li (2018) Computational Drug Discovery and Design: *Automated Inference of Chemical Group Discriminants of Biological Activity from Virtual Screening Data*. Springer. ISBN: 978-1-4939-7755-0

https://link.springer.com/protocol/10.1007/978-1-4939-7756-7_16

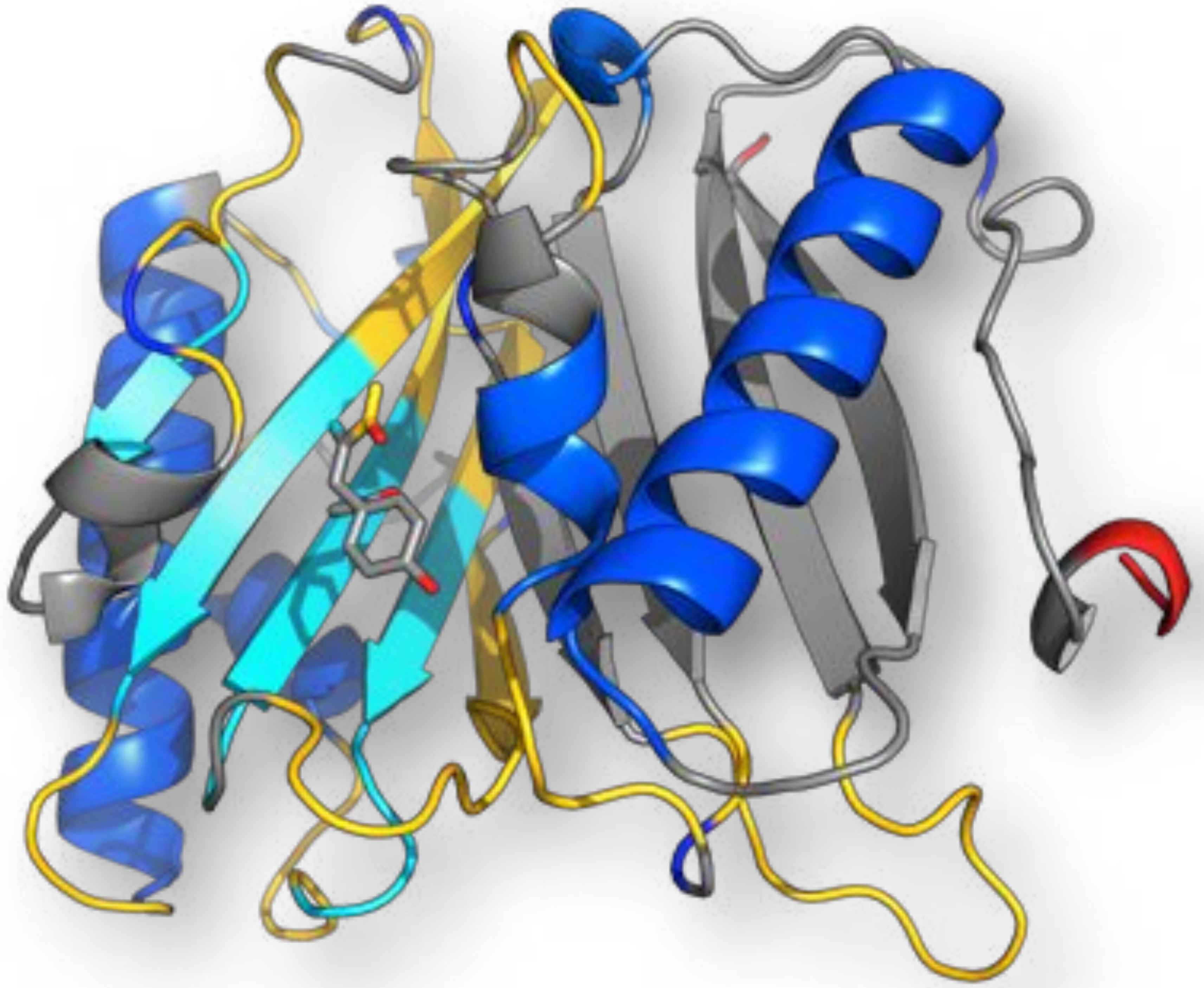
Case study 2

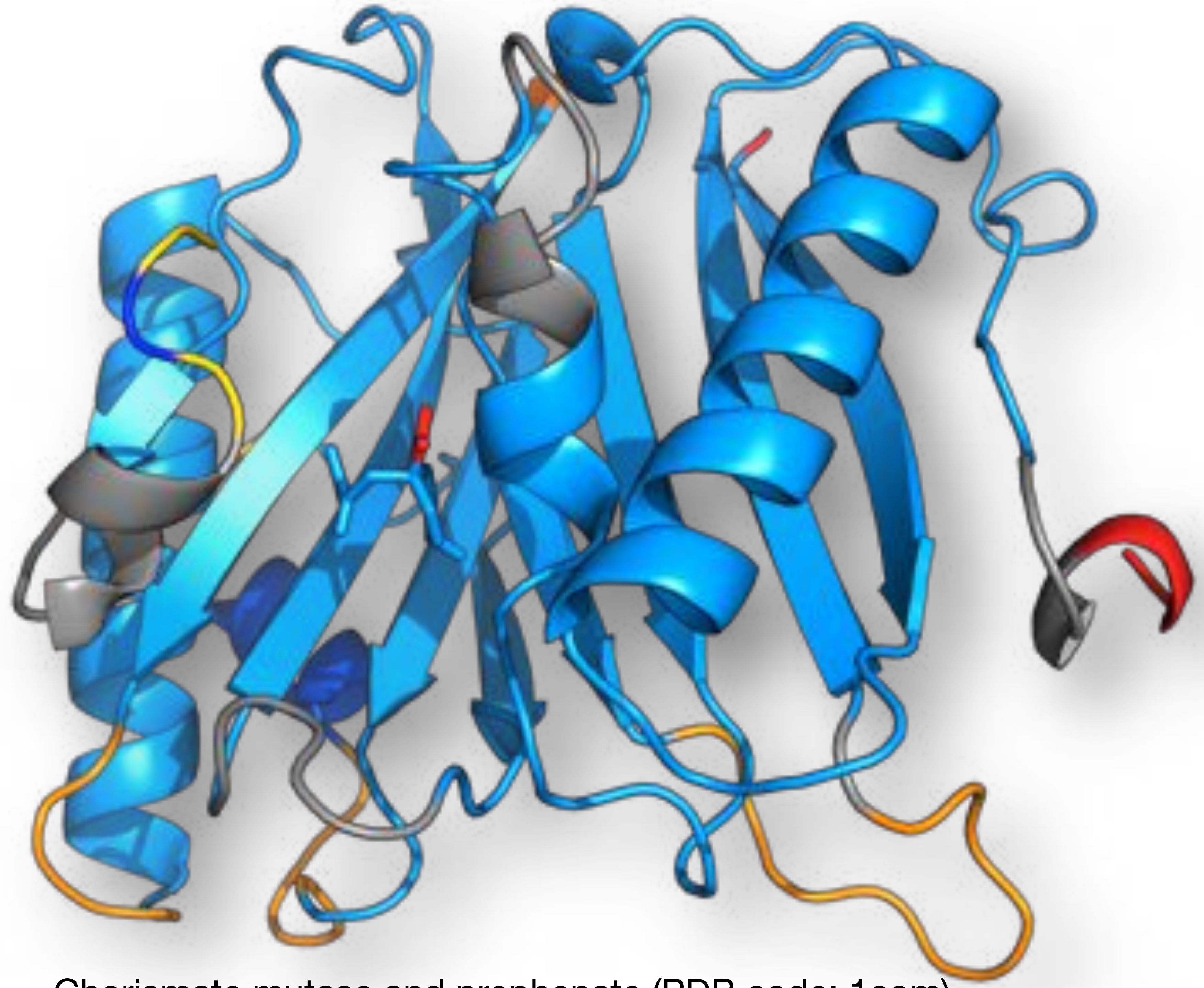
Predicting active state from structures with 96.6% accuracy (LOOCV)

"Flexibility Signatures of Class A GPCR Activation" (2019)
Joseph Bemister-Buffington, Alex J. Wolf, Sebastian Raschka, and Leslie A. Kuhn,
manuscript in preparation

“bad” docking
→ flexible binding pocket

near-native binding mode
→ rigid binding pocket





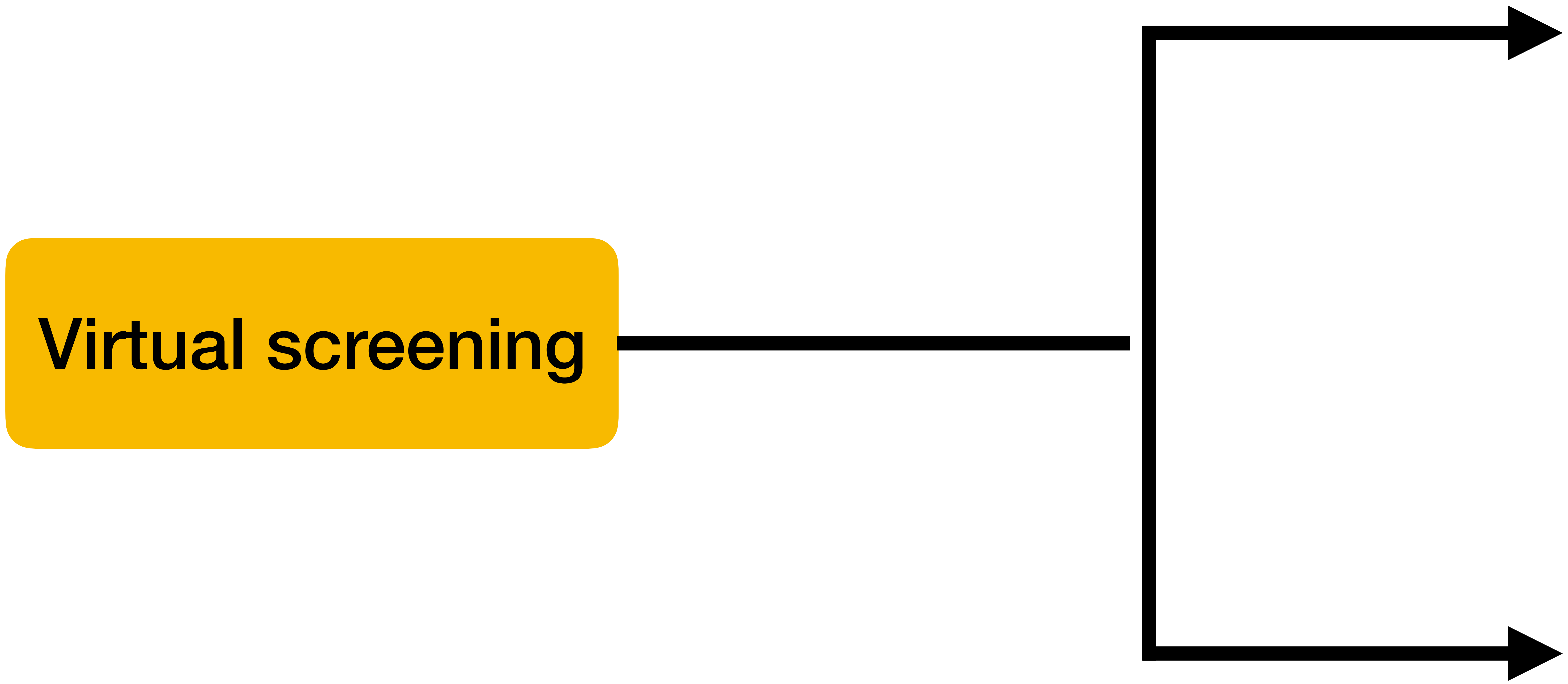
Chorismate mutase and prephenate (PDB code: 1com)

SiteInterlock

https://psa-lab.github.io/siteinterlock/

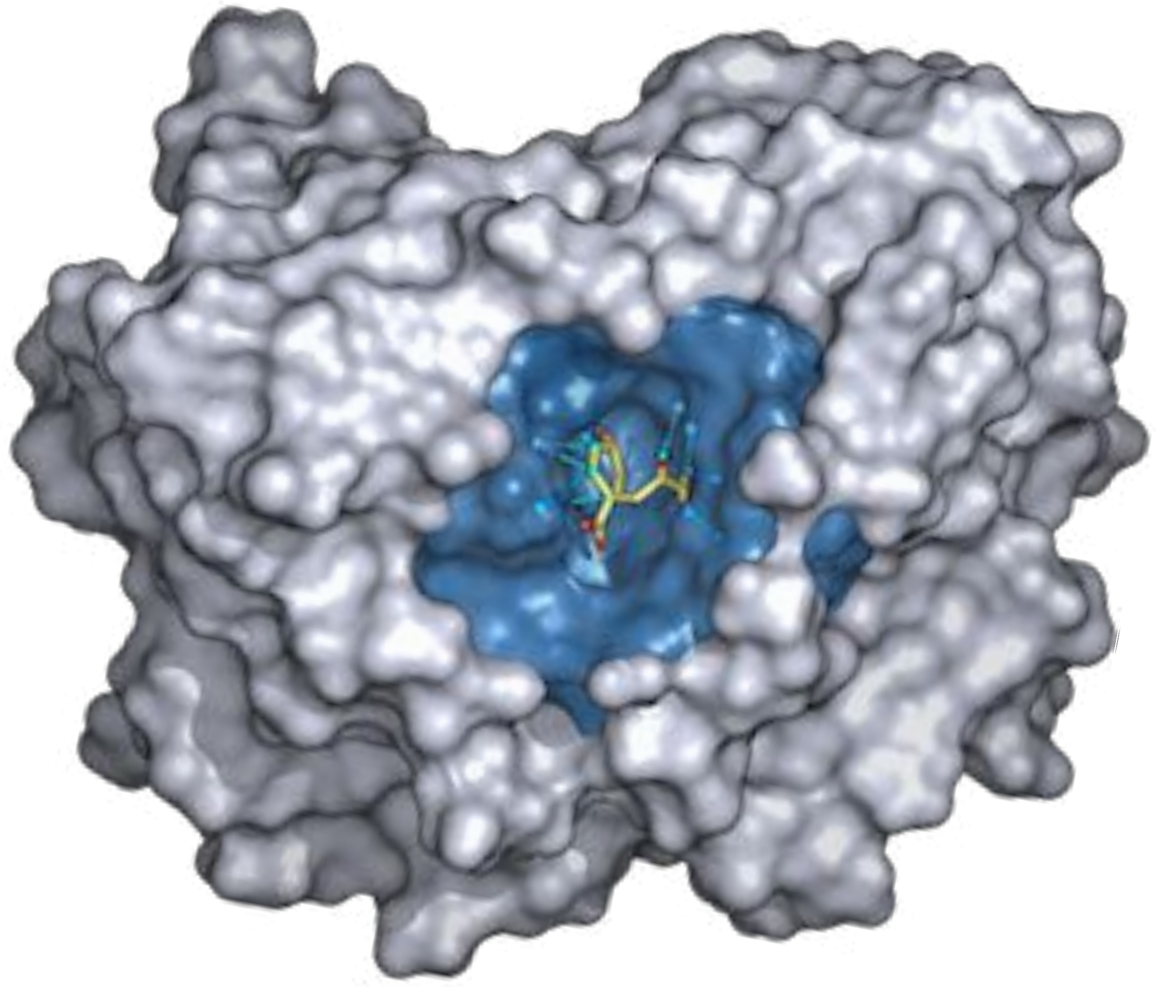
SiteInterlock: S. Raschka, J. Bemister-Buffington, L. A. Kuhn (2016)
Detecting the native ligand orientation by interfacial rigidity: SiteInterlock.
Proteins: Structure, Function and Bioinformatics 84.12: 1888-1901

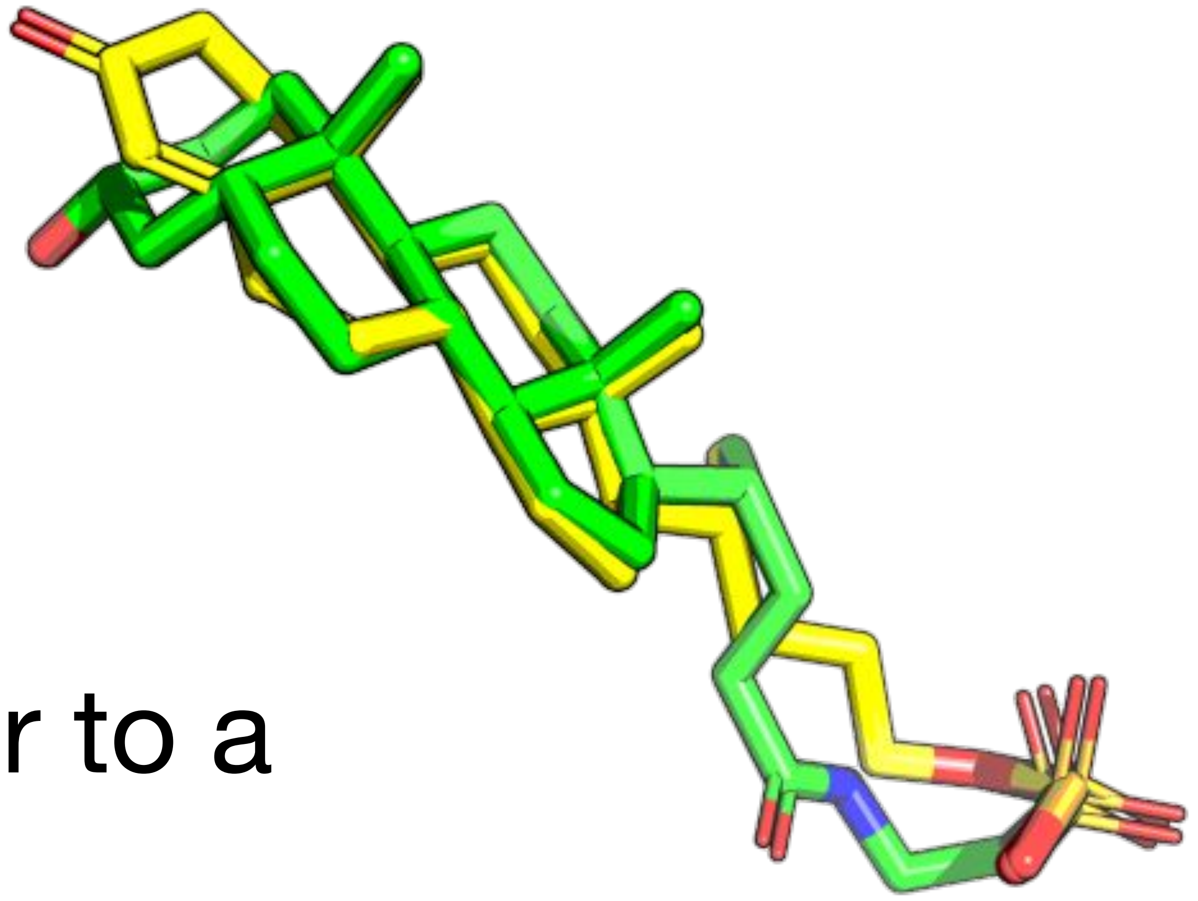
ProFlex: D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe (2001)
Protein Flexibility Predictions Using Graph Theory. Proteins: 44, 150-16



Virtual screening

Receptor structure-based



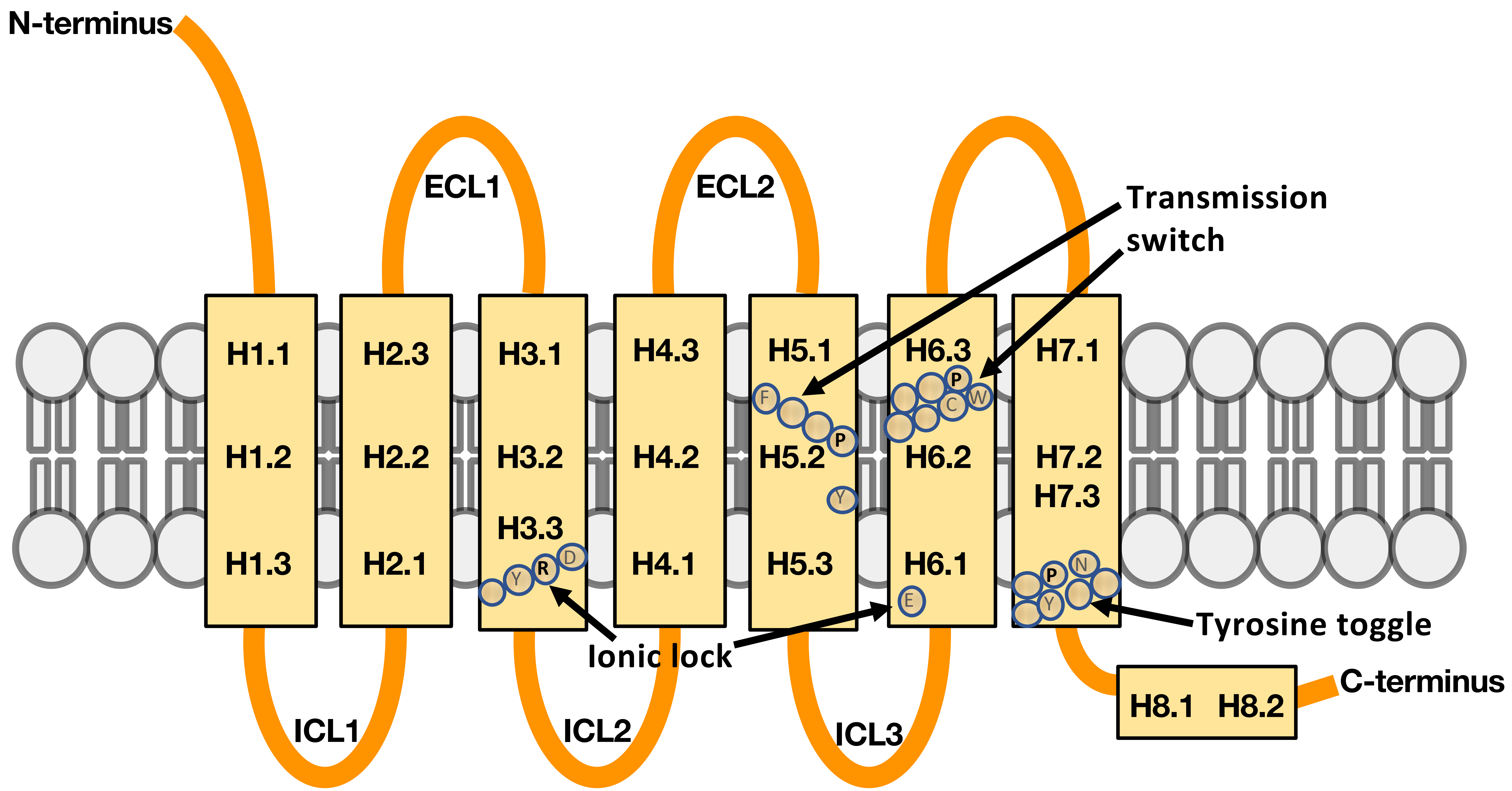


Small molecule-based

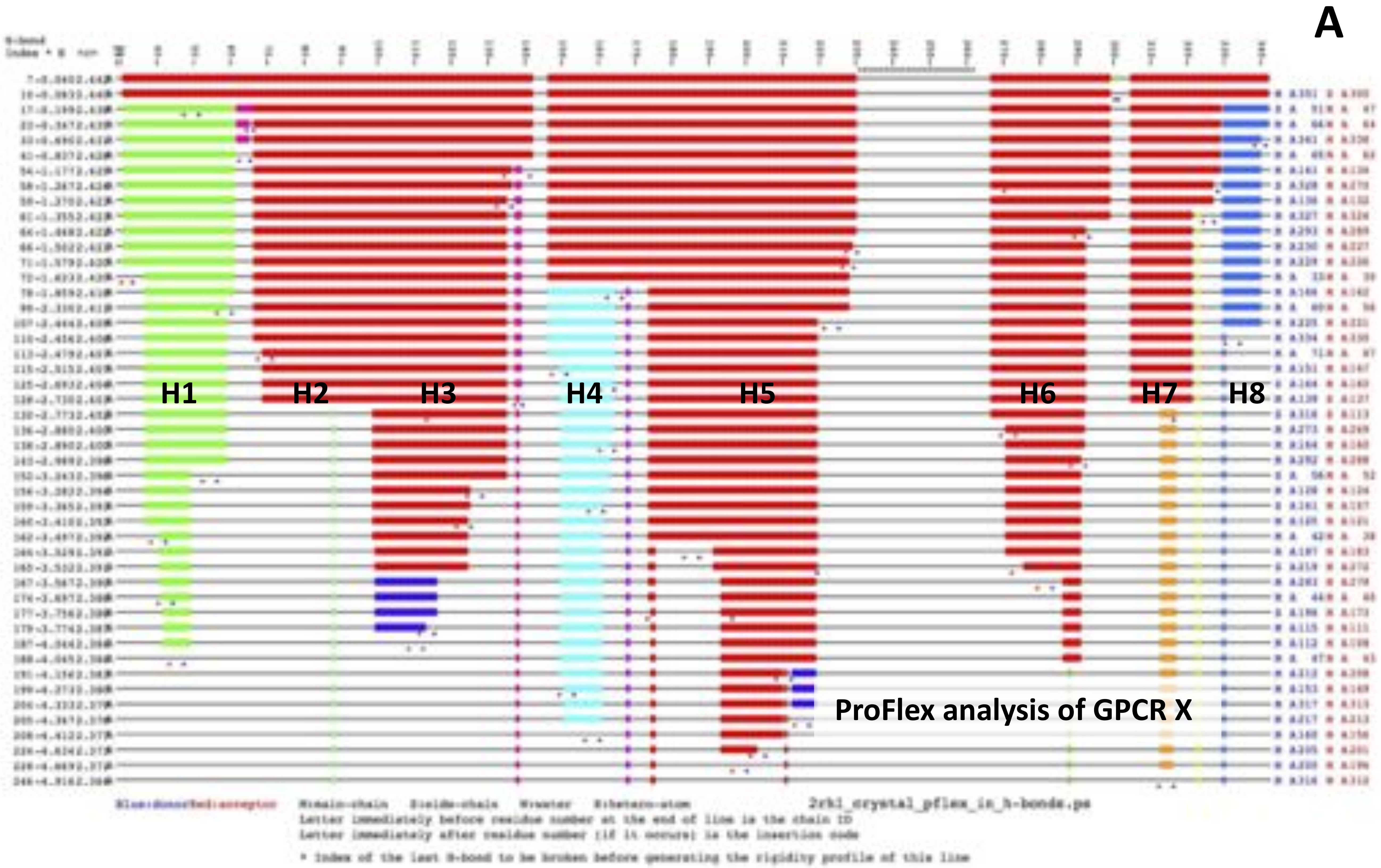
Assuming molecules similar to a
known binder are also
likely to bind the target receptor

Dataset of Active and Inactive GPCRs (here: only Class A)

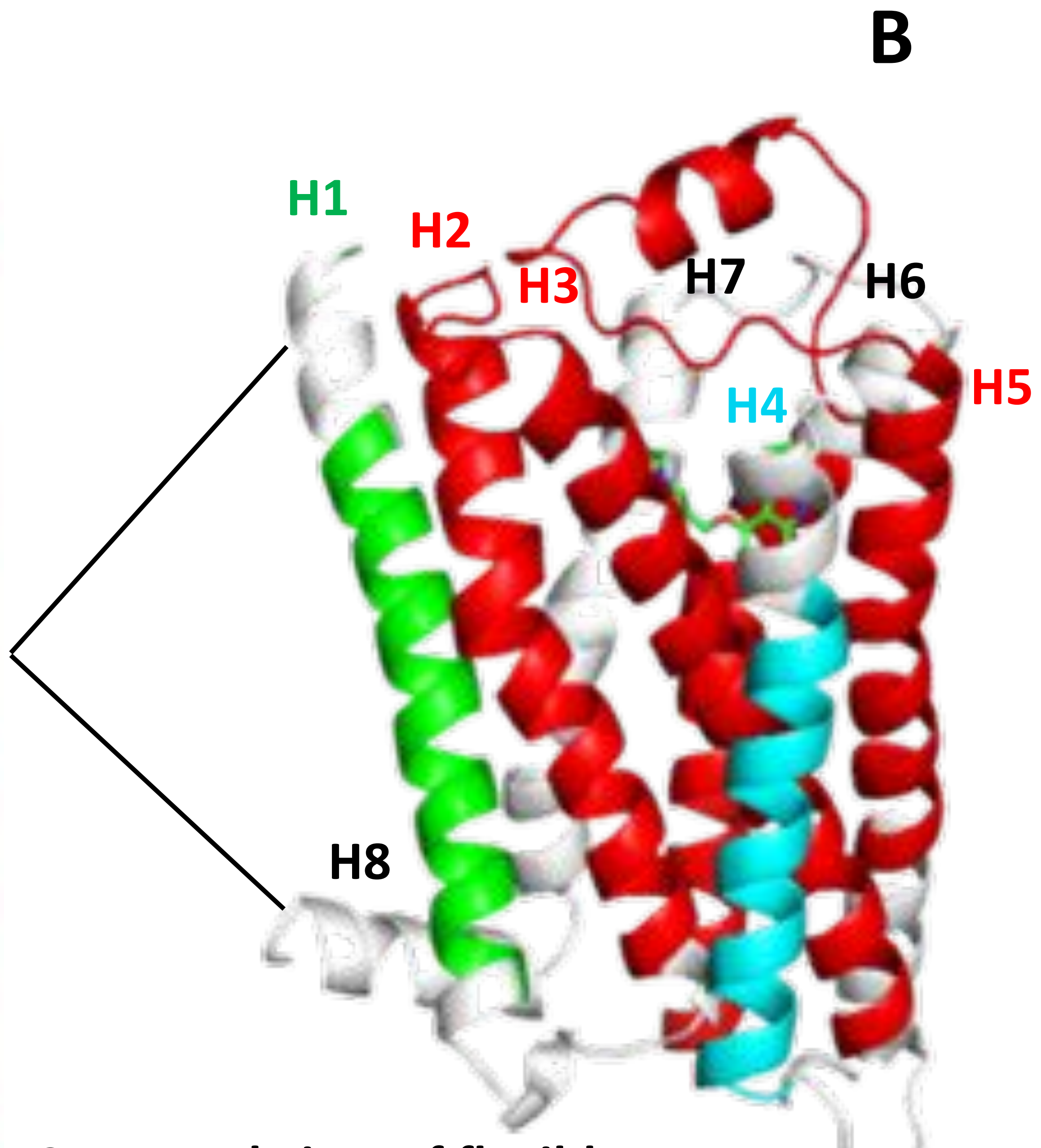
PDB ID	Activity*	Chain ID	Structure Description	Ligand Name	Organisms	Resolution (Å)	R(free)	R(work)
2VT4	0	A	Beta1 adrenergic receptor	4-{[(2s)-3-(Tert-butylamino)-2-hydroxypropyl]oxy}-3h-indole-2-carbonitrile	Meleagris gallopavo	2.7	0.268	0.212
3ODU	0	A	CXCR4 chemokine receptor	(6,6-Dimethyl-5,6-dihydroimidazo[2,1-b][1,3]thiazol-3-yl)methyl n,n'-dicyclohexylimidothiocarbamate	Homo sapiens	2.5	0.282	0.237
3V2Y	0	A	Lyso-phospholipid sphingosine 1-phosphate receptor	{(3r)-3-Amino-4-[(3-hexylphenyl)amino]-4-oxobutyl}phosphonic acid	Homo sapiens	2.8	0.272	0.229
3VW7	0	A	Human protease-activated receptor 1 (PAR1)	Ethyl [(1r,3ar,4ar,6r,8ar,9s,9as)-9-{(e)-2-[5-(3-fluorophenyl)pyridin-2-yl]ethenyl}-1-methyl-3-oxododecahydronaphtho[2,3-c]furan-6-yl]carbamate	Homo sapiens	2.2	0.235	0.218
3EML	0	A	A2A adenosine receptor	4-{2-[(7-Amino-2-furan-2-yl[1,2,4]triazolo[1,5-a][1,3,5]triazin-5-yl)amino]ethyl}phenol	Homo sapiens	2.6	0.231	0.196
...	...	...	...	...	...	...	...	...
3QAK	1	A	A2A adenosine receptor	ylpiperidin-4-yl)carbamoylamino]ethyl]purine-2-carboxamide	Homo sapiens	2.71	0.273	0.217
4IAR	1	A	5-HT1b	Ergotamine	Homo sapiens	2.7	0.261	0.223
4PXZ	1	A	Purinergic receptor P2Y12 receptor	2-(Methylsulfanyl)adenosine 5'-(trihydrogen diphosphate)	Homo sapiens	2.5	0.23	0.2
2YDV	1	A	A2A receptor	n-Ethyl-5'-carboxamido adenosine	Homo sapiens	2.6	0.258	0.233
3PQR	1	A	Metarhodopsin II	Retinal	Bos taurus	2.85	0.25	0.217
5C1M	1	A	Mu-opioid receptor	(2s,3s,3ar,5ar,6r,11br,11cs)-3a-Methoxy-3,14-dimethyl-2-phenyl-2,3,3a,6,7,11c-hexahydro-1h-6,11b-(epiminoethano)-3,5a-methanonaphtho[2,1-g]indol-10-ol	Mus musculus	2.1	0.221	0.185
4XES	1	A	Neurotensin receptor	Neurotensin chain B	Rattus norvegicus	2.6	0.28	0.23
5GLH	1	A	Endothelin receptor type B	Endothelin-1 peptide chain B	Homo sapiens	2.8	0.277	0.234
...	...	...	...	...	...	...	...	...



Circles represent residue positions of well-conserved GPCR motifs. The residues shown are those found in human CXCR4



ProFlex analysis of GPCR X

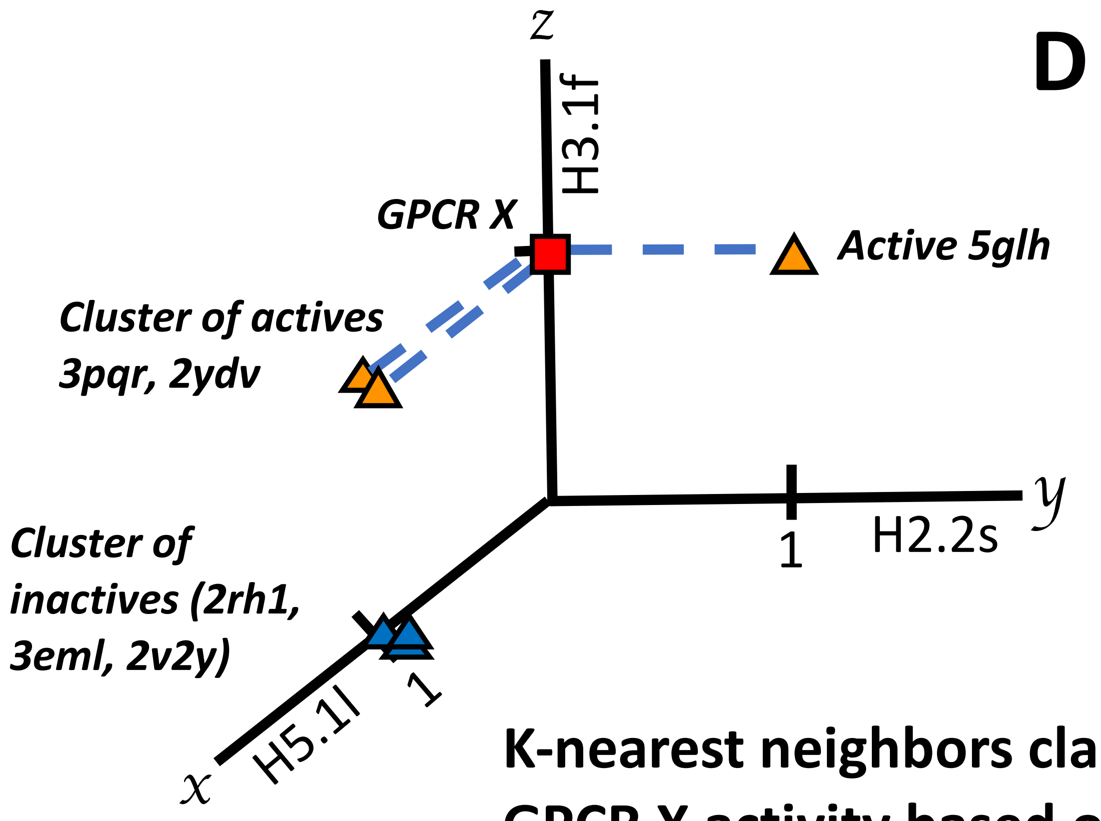


Structural view of flexible and rigid regions in GPCR X

Tabulation of key discriminatory flexible and rigid features of helices and loops in GPCR X and GPCRs of known activity

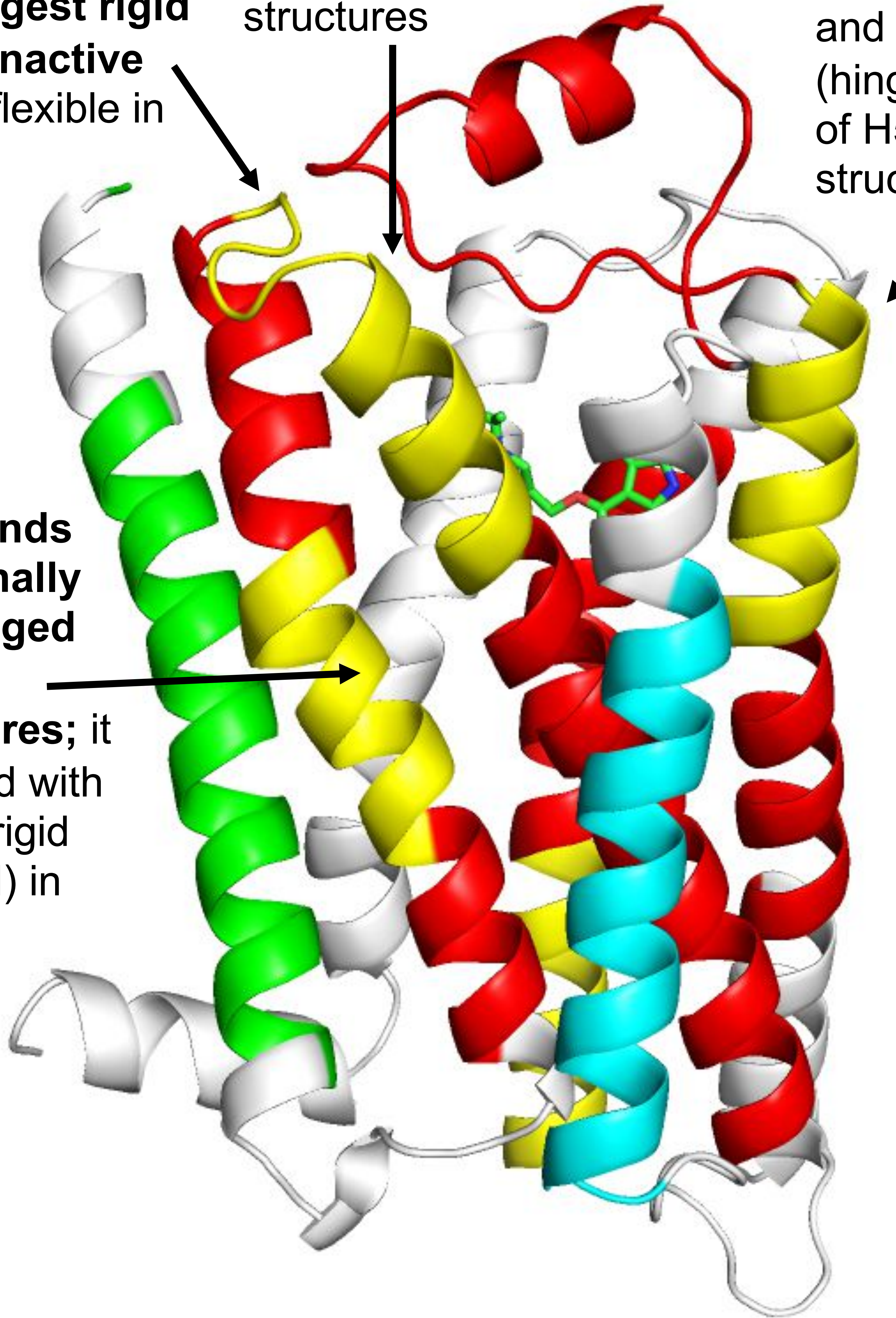
C

Activity	PDB	H5.1l	H2.2s	H3.1f	...
Inactive	2RH1	1	0	0	1
Inactive	3EML	1	0	0	0
Inactive	2V2Y	1	0	0	1
Active	5GLH	0	1	1	0
Active	3PQR	1	0	1	0
Active	2YDV	1	0	1	0
?	GPCR X	0	0	1	0


D
z
H3.1f
GPCR X
Active 5glh
Cluster of actives 3pqr, 2ydv
y
1
H2.2s
Cluster of inactives (2rh1, 3eml, 2v2y)
1
x
H5.1l

K-nearest neighbors classification of GPCR X activity based on known activity of 3 nearest neighbor GPCRs

GPCR X is predicted to be active, based on the activity of its nearest neighbors



ECL1 region (yellow) tends to be part of the scaffold-like largest rigid region (red) in inactive structures and flexible in active structures

H3.1 (yellow) tends to be flexible in active structures and part of the scaffold-like largest rigid region (red) in inactive structures

H5.1 (yellow) tends to be part of the scaffold-like largest rigid region (red) in inactive structures, and separately rigid (hinging relative to the rest of H5) or flexible in active structures

H2.2 region (yellow) tends to be a separate, internally rigid helical region hinged to the end of the helix (H2.3) in active structures; it tends to be mutually rigid with the scaffold-like largest rigid region of the GPCR (red) in inactive structures

We anticipate that ProFlex-based classification of GPCRs into active vs. inactive

will also be useful for ligand design: agonists vs antagonists

Current Trends and Outlook

Scoring Protein-Ligand Poses with 3D ConvNets

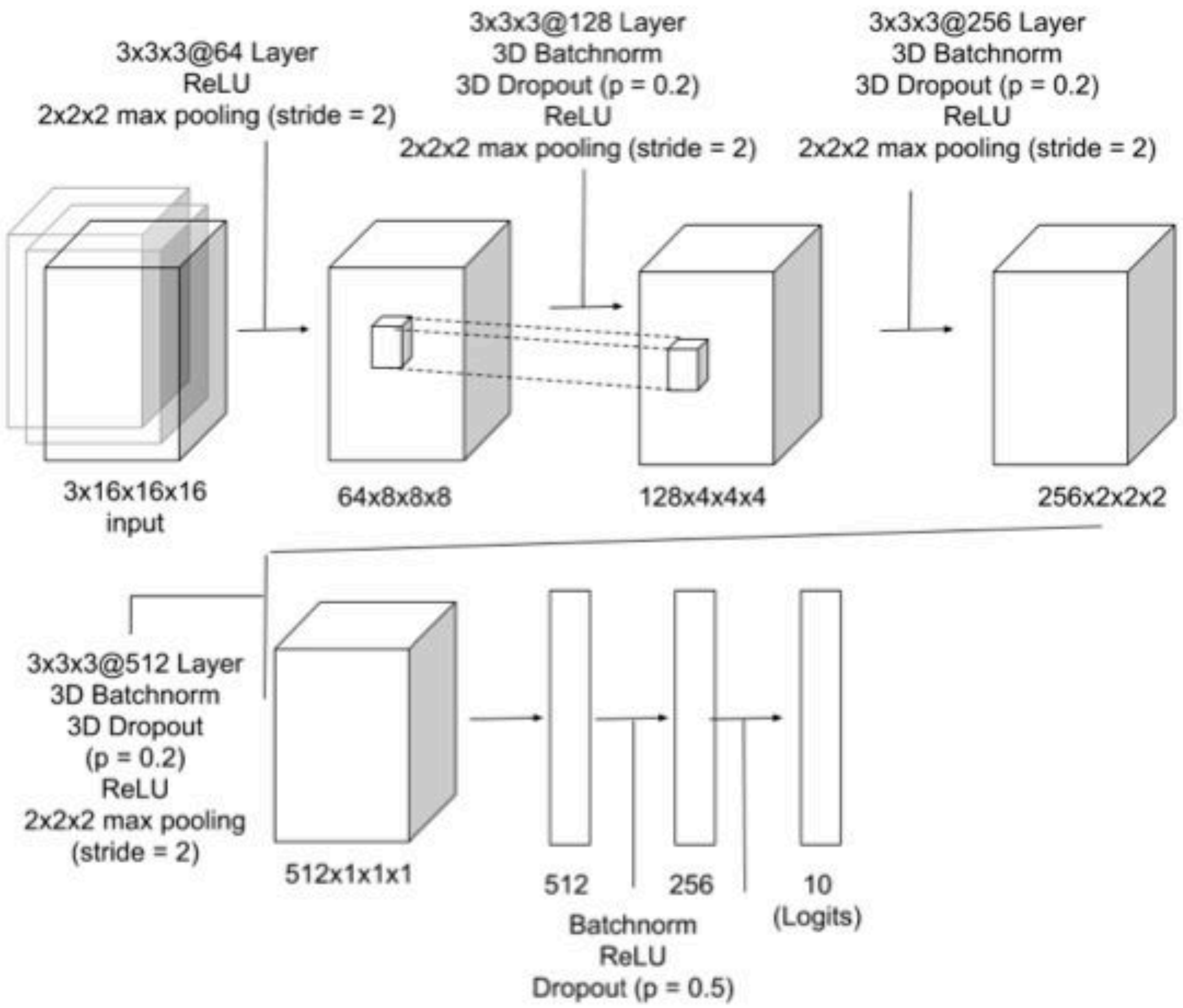
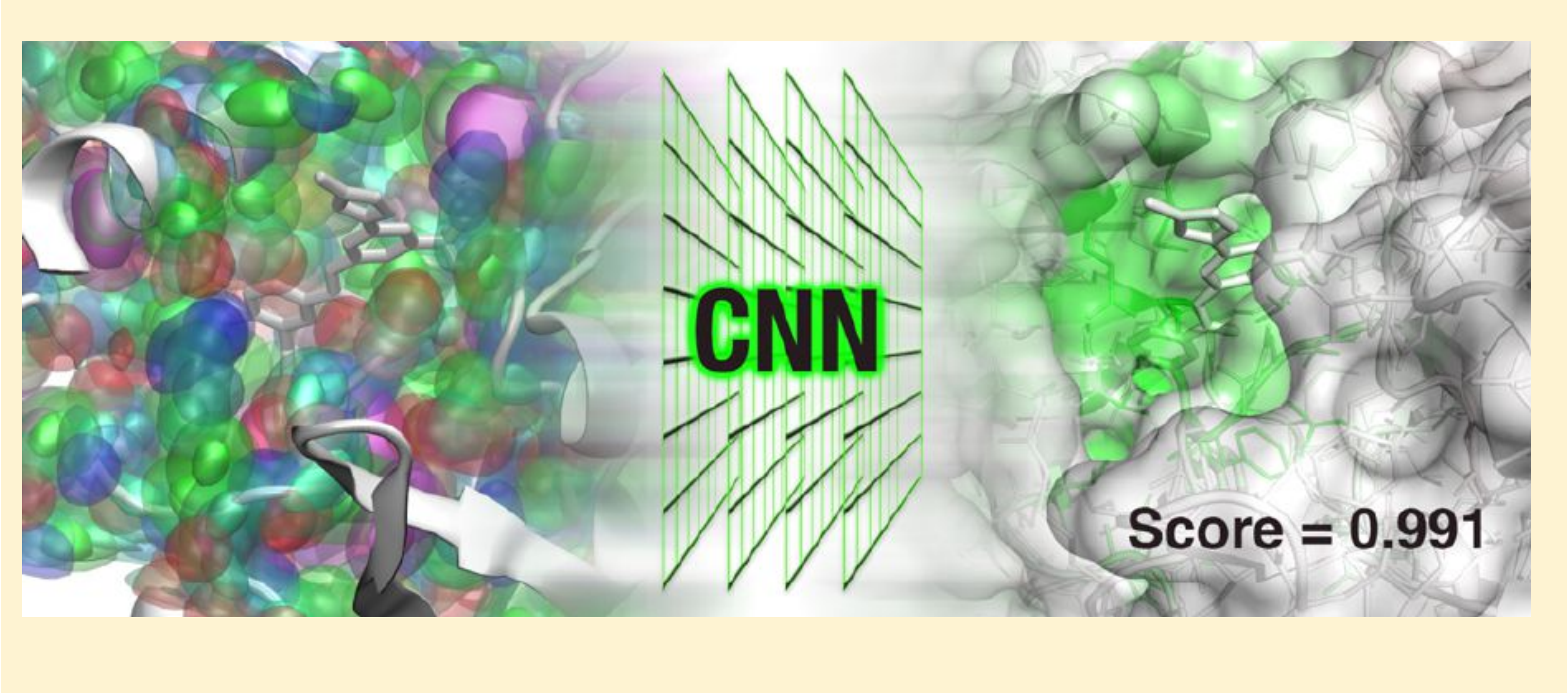


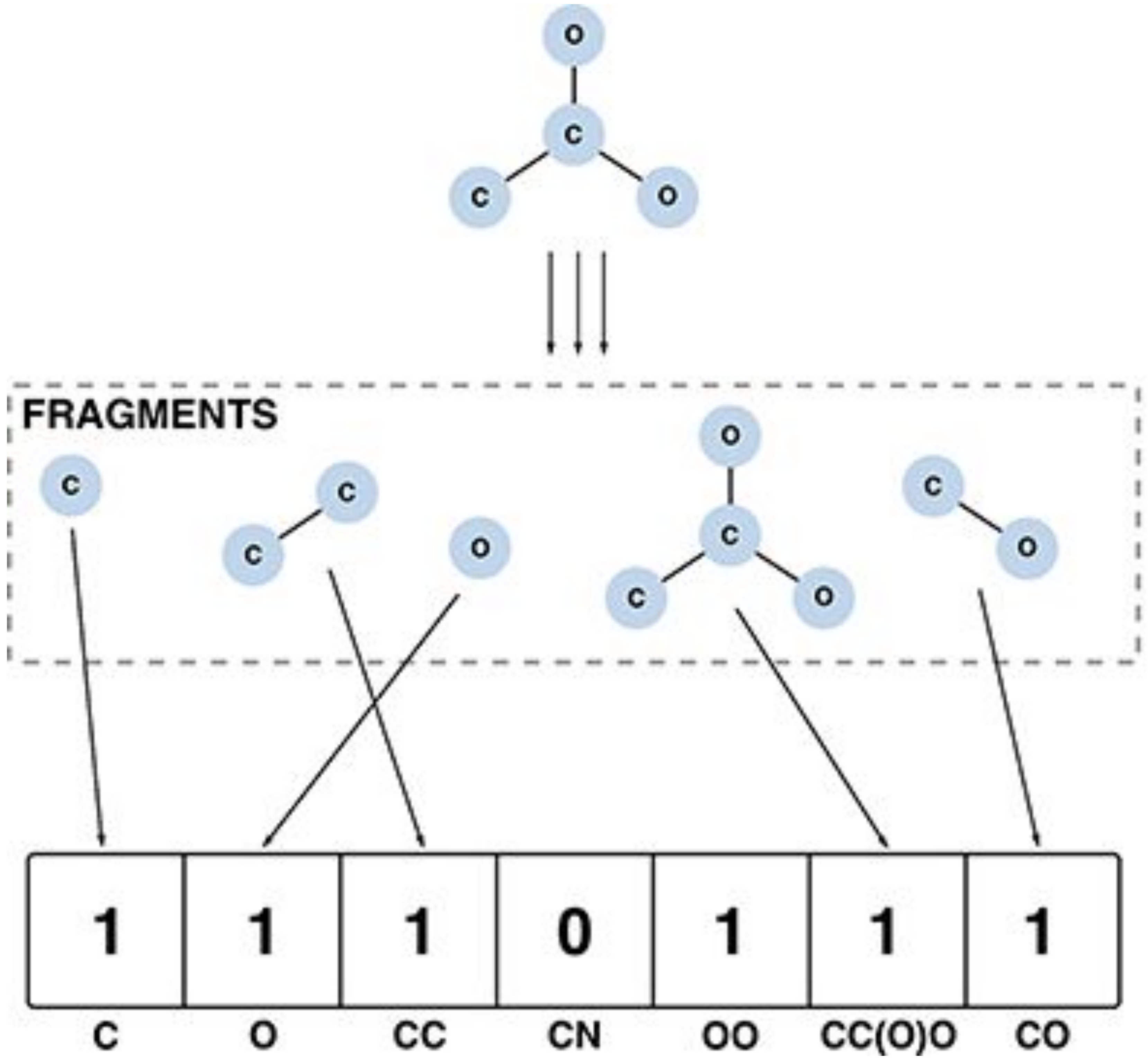
Fig. from Stat479 class project by
Sam Berglin, Zheming Lian, Jiahui Jiang



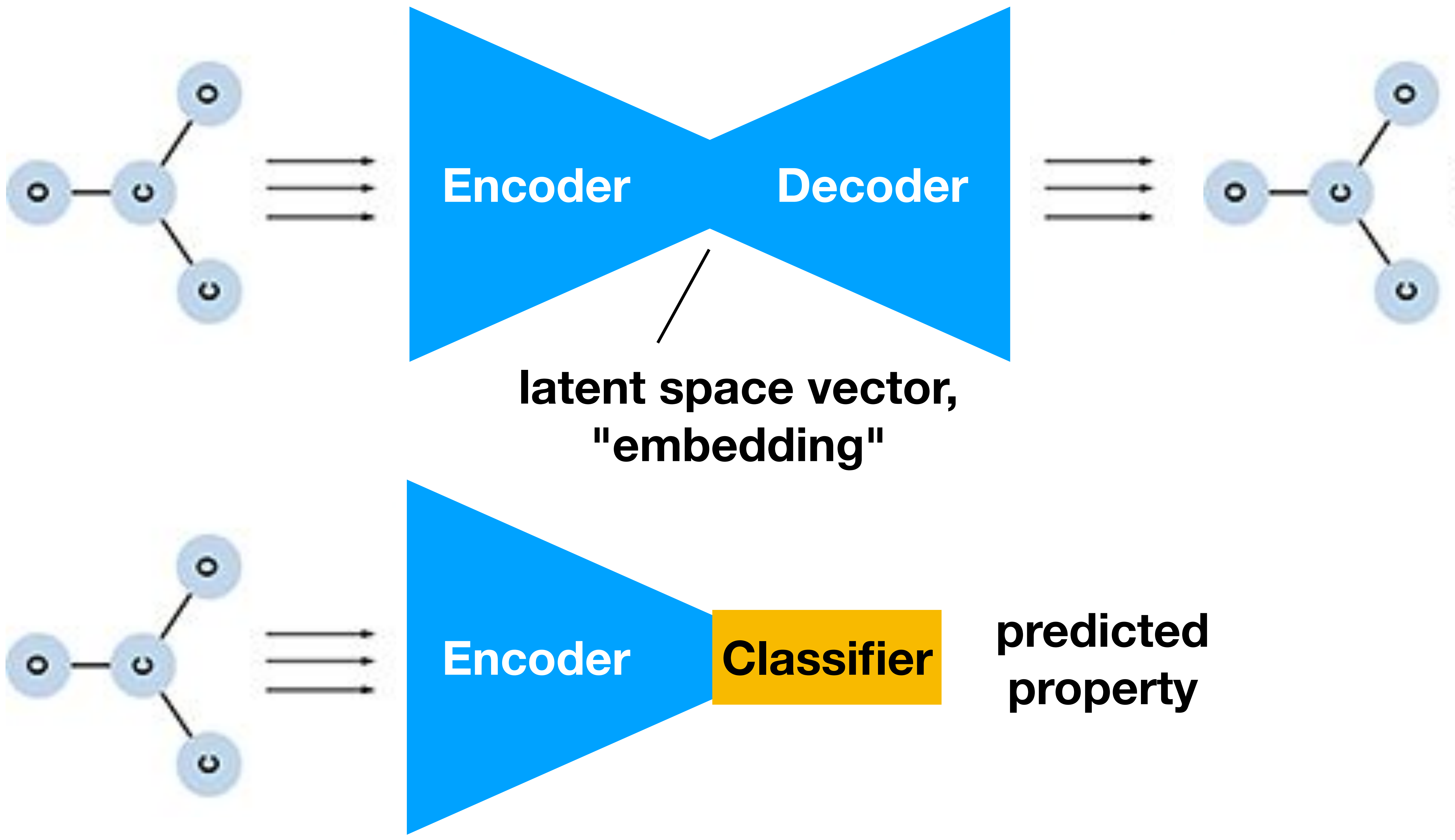
Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., & Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, *57*(4), 942-957.

"Neural Fingerprints"

Traditional fingerprints:



Representation learning:



Encoder

Decoder

latent space vector,
"embedding"

Encoder

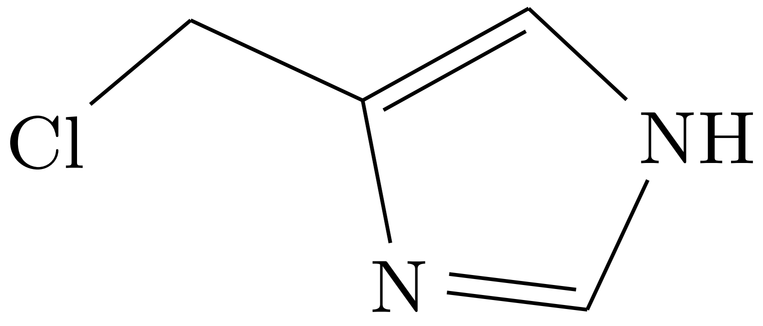
Classifier

predicted
property

Hop, Patrick, Brandon Allgood, and Jessen Yu. "Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts." *Molecular pharmaceutics* 15.10 (2018): 4371-4377.
https://pubs.acs.org/doi/full/10.1021/acs.molpharmaceut.7b01144

De Novo Design

Graph:



SMILES: ClCc1c[nH]cn1

One-hot encoding:

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

Fig. 3 Three representations of 4-(chloromethyl)-1H-imidazole. Depiction of a one-hot representation derived from the SMILES of a molecule. Here a reduced vocabulary is shown, while in practice a much larger vocabulary that covers all tokens present in the training data is used

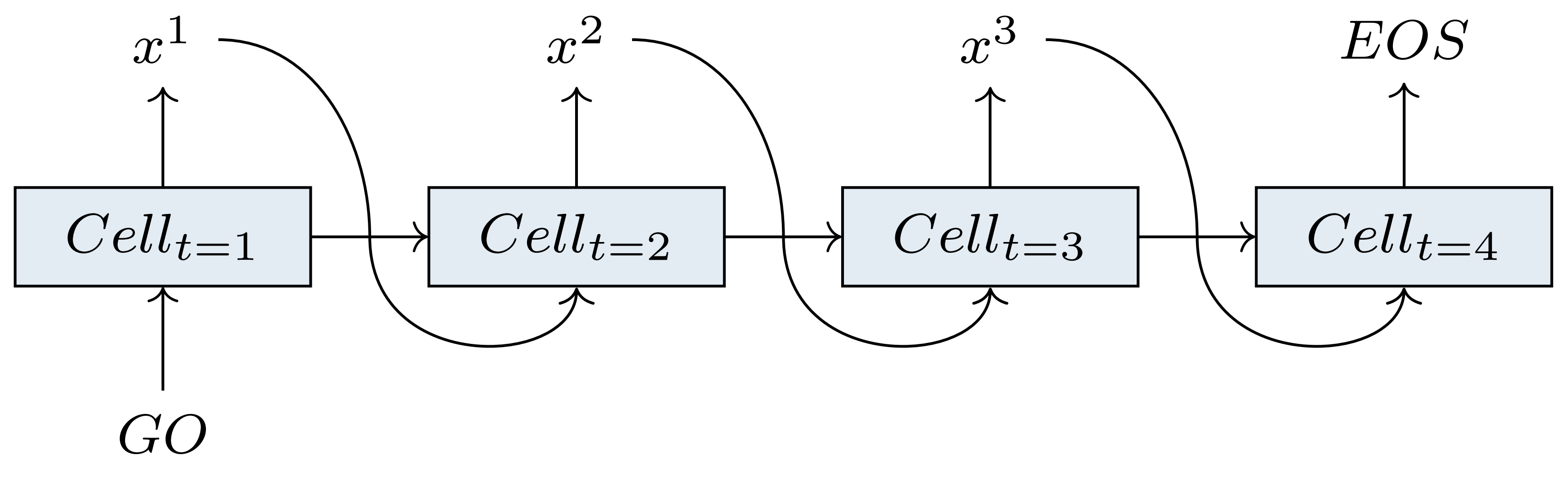


Fig. 2 Generating sequences. Sequence generation by a trained RNN. Every timestep *t* we sample the next token of the sequence x^t from the probability distribution given by the RNN, which is then fed in as the next input

Train recurrent neural net (RNN) to generate molecules (whole ChEMBL database)

Use Reinforcement Learning to fine-tune RNN to

1) **Generate molecules with a certain property**
2) **Generate analogs of a query molecule**
3) **Generate bioactive molecules**

Olivecrona, Marcus, et al. "Molecular de-novo design through deep reinforcement learning." *Journal of Cheminformatics* 9.1 (2017): 48.

https://www.biomedcentral.com/openurl?doi=10.1186/s13321-017-0235-x

My current research related to deep learning for drug discovery:

Thanks for attending!

Questions?

And thanks to my team!

Jitian Zhao
Zhongjie Yu (Statistics grad students)
Richard Yang
Yien Xu

Benjamin Kaufmann (BMI grad student)

sraschka@wisc.edu
http://stat.wisc.edu/~sraschka/