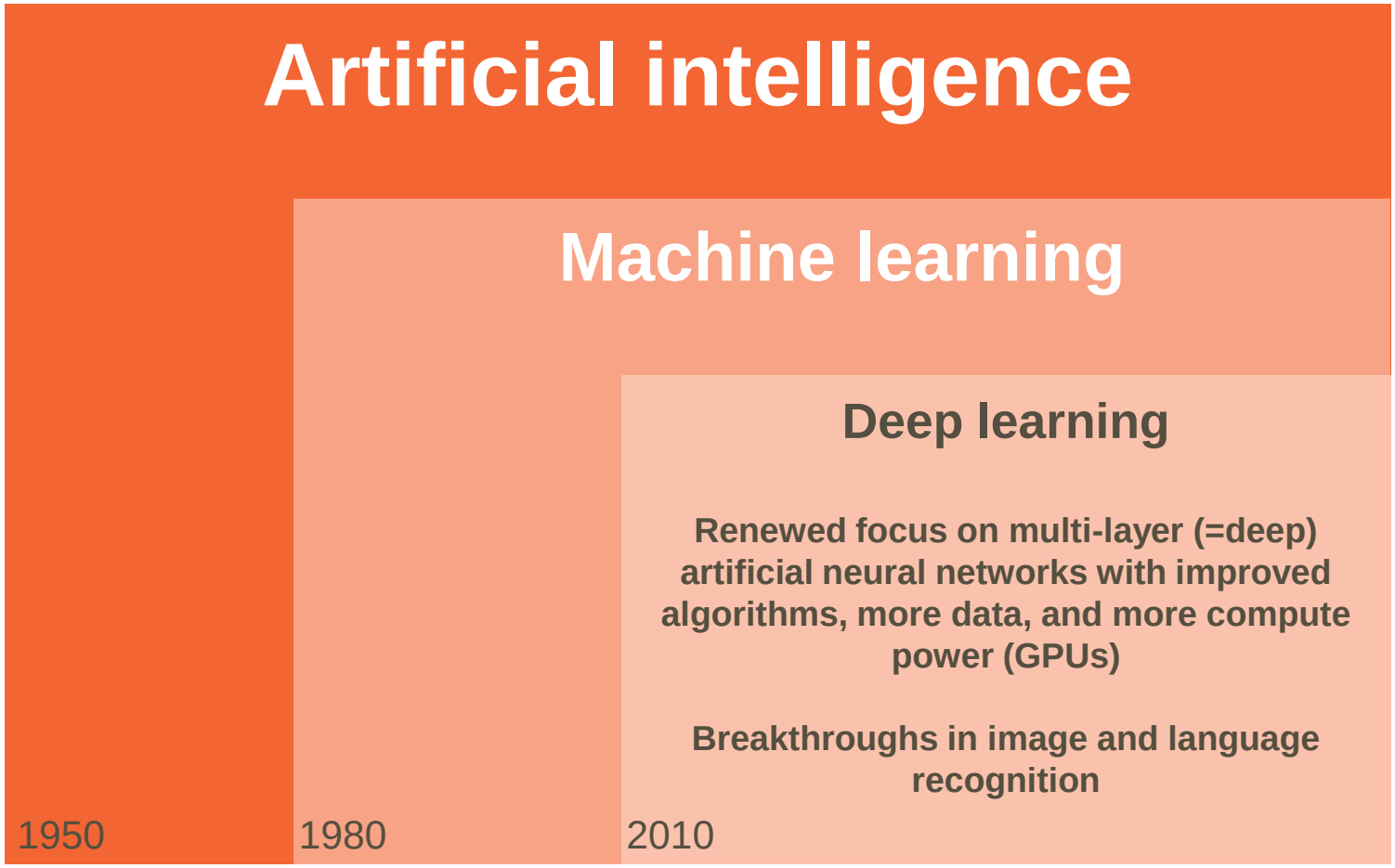
Accelerating drug discovery with deep neural networks

literature review

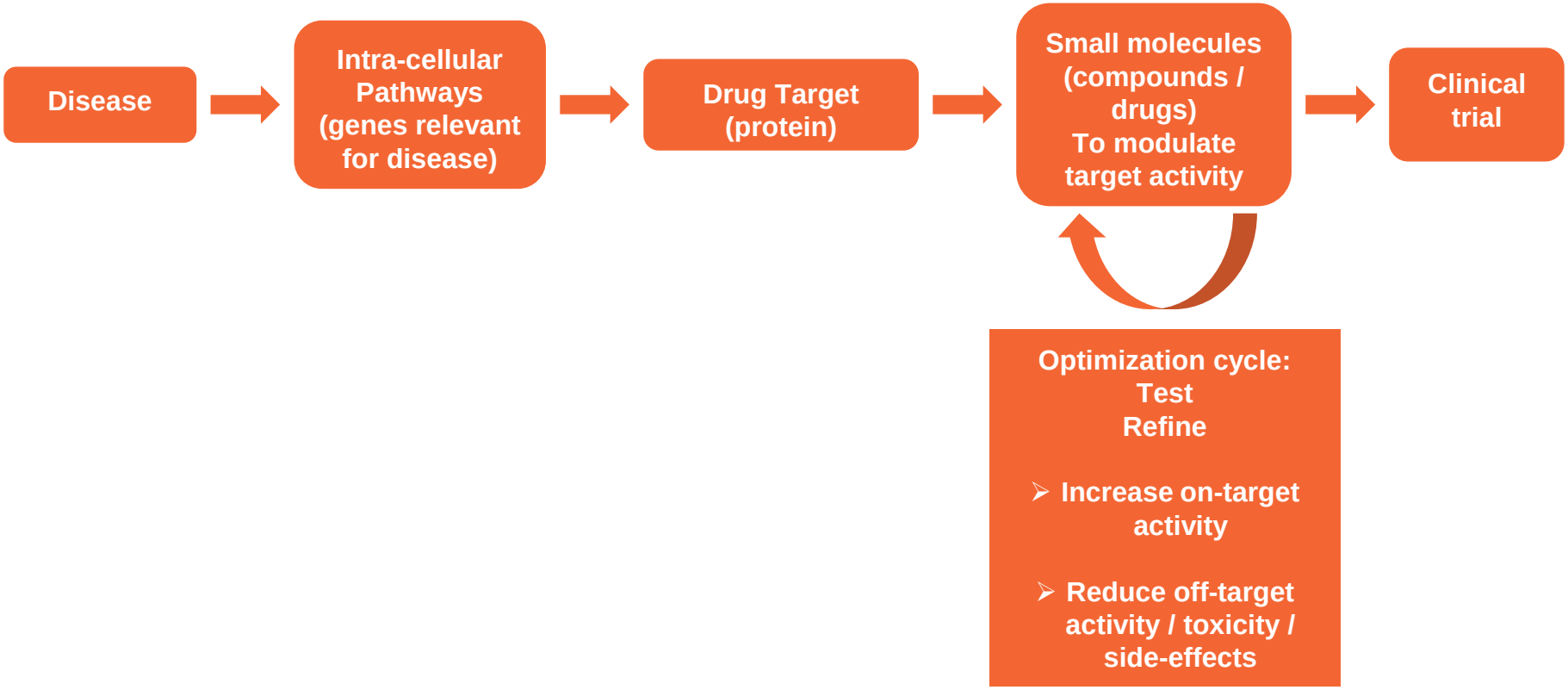
Tobias Sikosek
Senior Data Scientist
In Silico Unit
(Heidelberg)

Deep Learning



Drug Discovery in a nutshell

Preclinical drug discovery



Deep Learning in Drug Discovery

Learning from data to make better in silico predictions

- Target identification
  - Based on human **genetic variation (DNA)** associated with disease
  - Based on cellular **pathways / gene expression** associated with a disease

- Matching targets and small molecules with DL
  - Encode **protein structure**
  - Encode **small molecule**
  - **generate** new small molecules
  - Predict drug-target **interactions**

- Drug vs Biology: toxicity, side-effects
  - Predict **toxicity** of drugs from their chemical structure based on past **clinical failures**

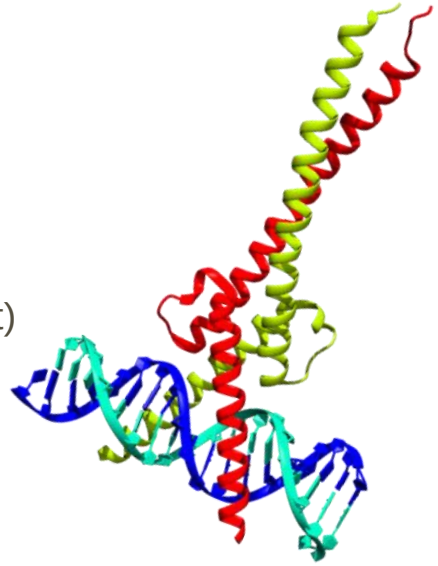
Target identification

protein that can be modified by drug to change disease state

Target identification

Serving patient subpopulations sharing common genetic markers for disease

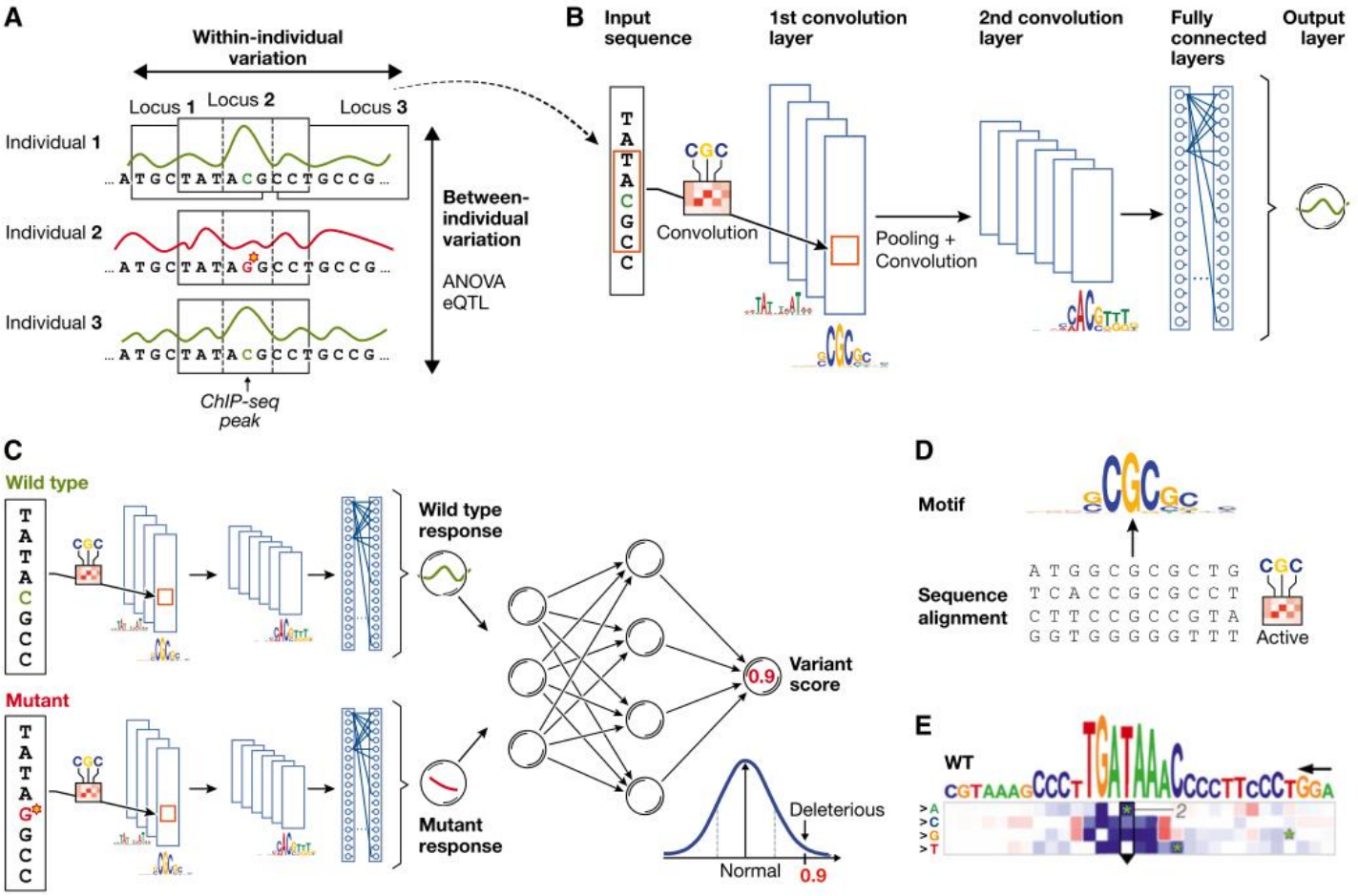
- Needle in a haystack problem:
  - **Genome wide association studies** statistically link regions within chromosomes to a particular disease / phenotype
  - Across human population, every chromosome region may contain many thousand SNVs (**single nucleotide variations**) – which one causes the disease?
  - Often SNVs lie within DNA regions bound by **transcription factors**, TFs (DNA-binding proteins that act as regulatory switches within the complex circuitry that controls all cell processes)
  - If an inherited change in that DNA region leads to decreased **TF binding** – a disease state of the cell can be the result
  - TFs are usually not direct drug targets, but may lead to the right target
- Deep Learning solution:
  - Input: DNA sequence segment
  - Output: binary classification (sequence contains TF-binding site – or not)



Crystal structure of Myc-Max recognizing DNA. PDB: 1NKP

Target identification

DNA-protein binding prediction

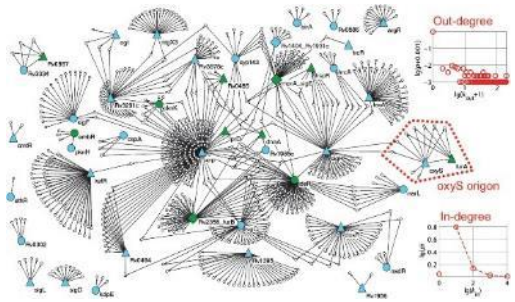


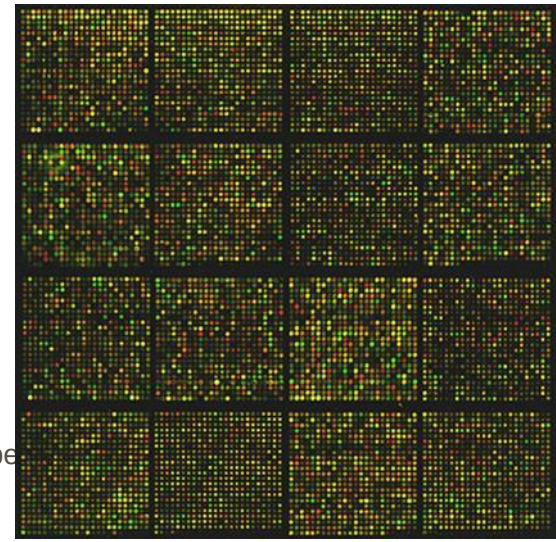
Angermueller, C., Pärnamaa, T., Parts, L. and Stegle, O. (2016) 'Deep learning for computational biology', *Molecular Systems Biology*, 12(7), p. 878

Target identification

Gene expression patterns reveal disease biology and pathways

- Complex network interaction problem:
  - Biology at the **cellular** level is the result of countless molecular **interactions** that can be described as **networks** (gene regulation, protein-protein interaction, metabolic reactions, protein modifications)
  - **Perturbations** in this complex system (**disease, environment, drugs**) can have highly **non-linear consequences** that are difficult to model or predict
  - Cellular data contain a lot of intrinisic **noise** (high time-dependence, dynamics, experimental variation, etc.)
  - The most popular experimental assay to capture complex cellular biology is **transcriptomics**, i.e. **expression** (=abundance/frequency of RNA copies made from DNA gene) patterns of all ~20000 **genes** – or cell-type specific subset.
  - Gene expression can be highly (anti-)corellated, i.e. When high expression of a gene causes increase or decrease of a range of other genes
  - Genes can be mapped to same **pathway** (causally linked to a common endpoint). Example: inherited genetic change associated with a disease changes gene expression with downstream effect along the pathway. Any gene (node) in the pathway could be target of a drug intervention to modify aberrant gene expression back to normal level.



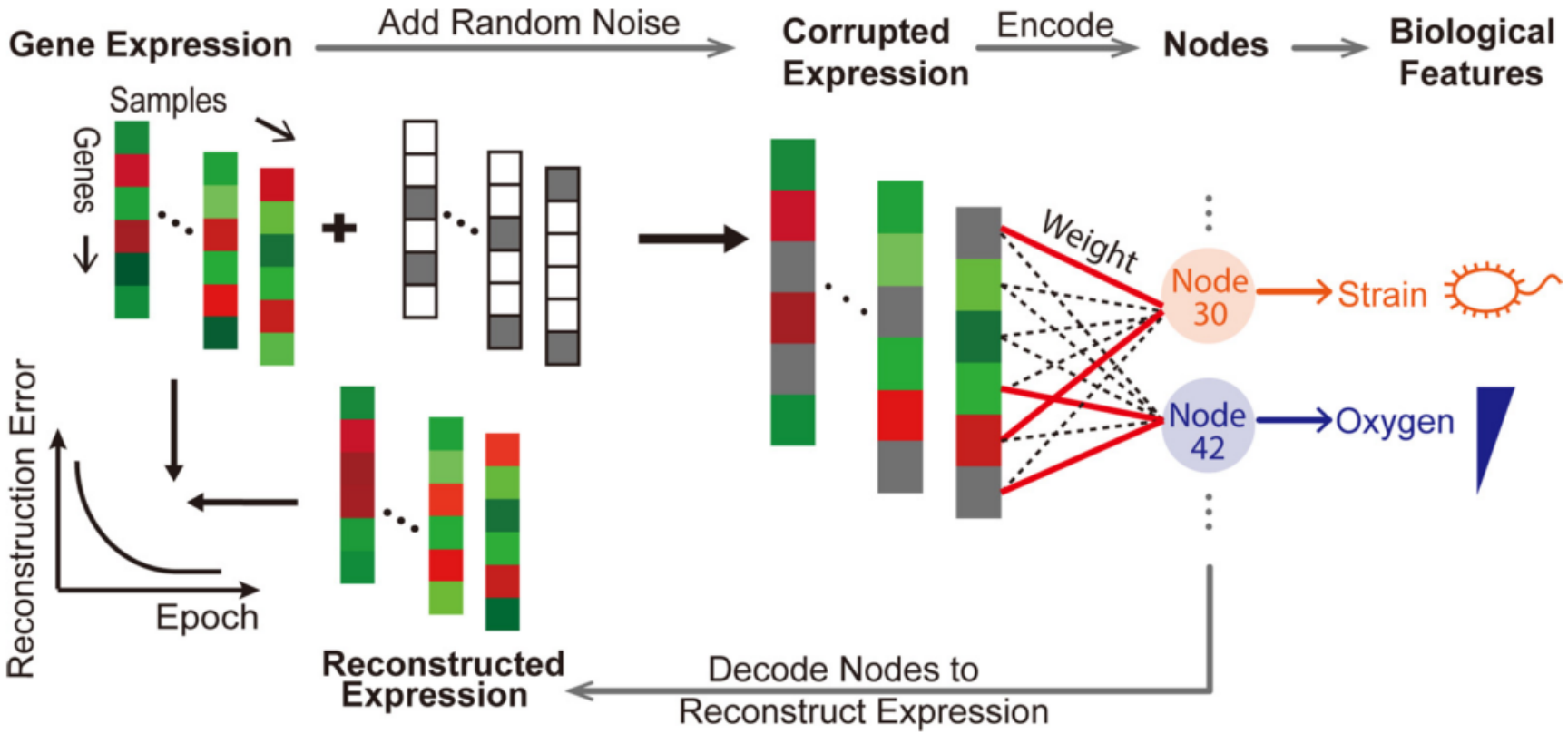


Balázsi, G., Heath, A. P., Shi, L. and Gennaro, M. L. (2008) 'The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest', *Molecular Systems Biology*, 4(225), pp. 1–8. ; https://commons.wikimedia.org/wiki/File:Mouse_cdna_microarray.jpg

Target identification

Gene expression patterns reveal disease biology and pathways

De-noising autoencoders signal/noise from gene expression data and provide lower-dimensional fingerprint of data (→ dimensionality reduction)

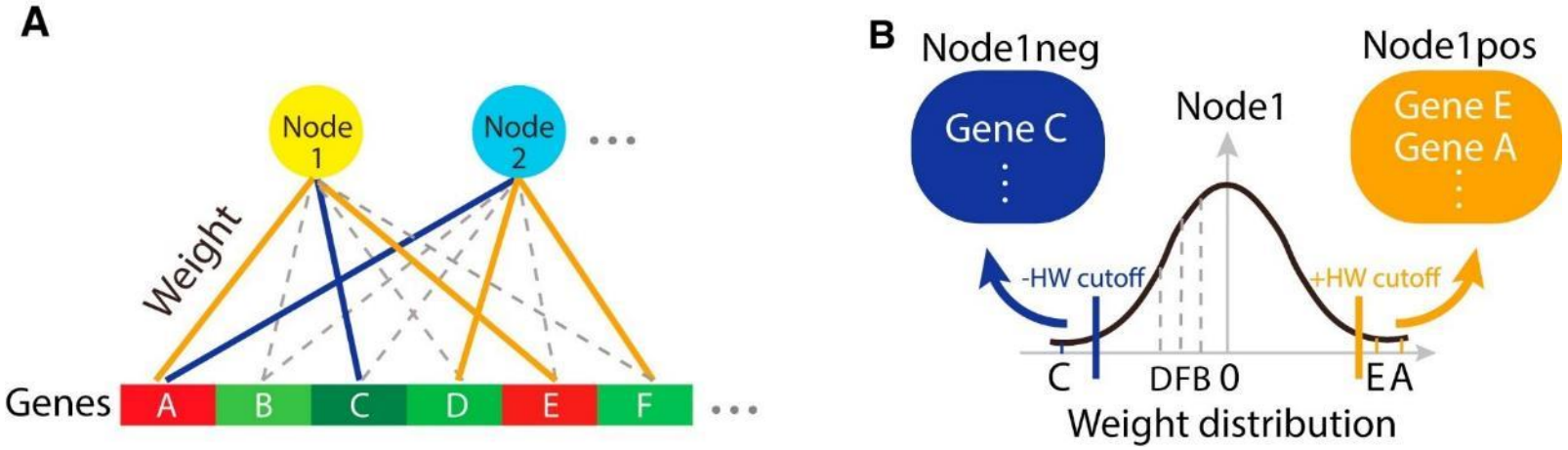


Tan, J., Hammond, J. H., Hogan, D. A. and Greene, C. S. (2016) 'ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions', *mSystems* 1(1), pp. e00025-15.

Target identification

Gene expression patterns reveal disease biology and pathways

- **Weights** (parameters) between input layer (genes) and hidden layer can be used to „label“ hidden nodes.
- Each **hidden node** is positively linked to subset of genes and negatively linked to other genes
- Each hidden node could in principle correspond to a cellular pathway (but is not restricted to any known **pathways**)
- Averaged results from **ensembles** of autoencoders yield improved results
- Outcome: which genes/pathways are most active in disease? → **potential drug targets**

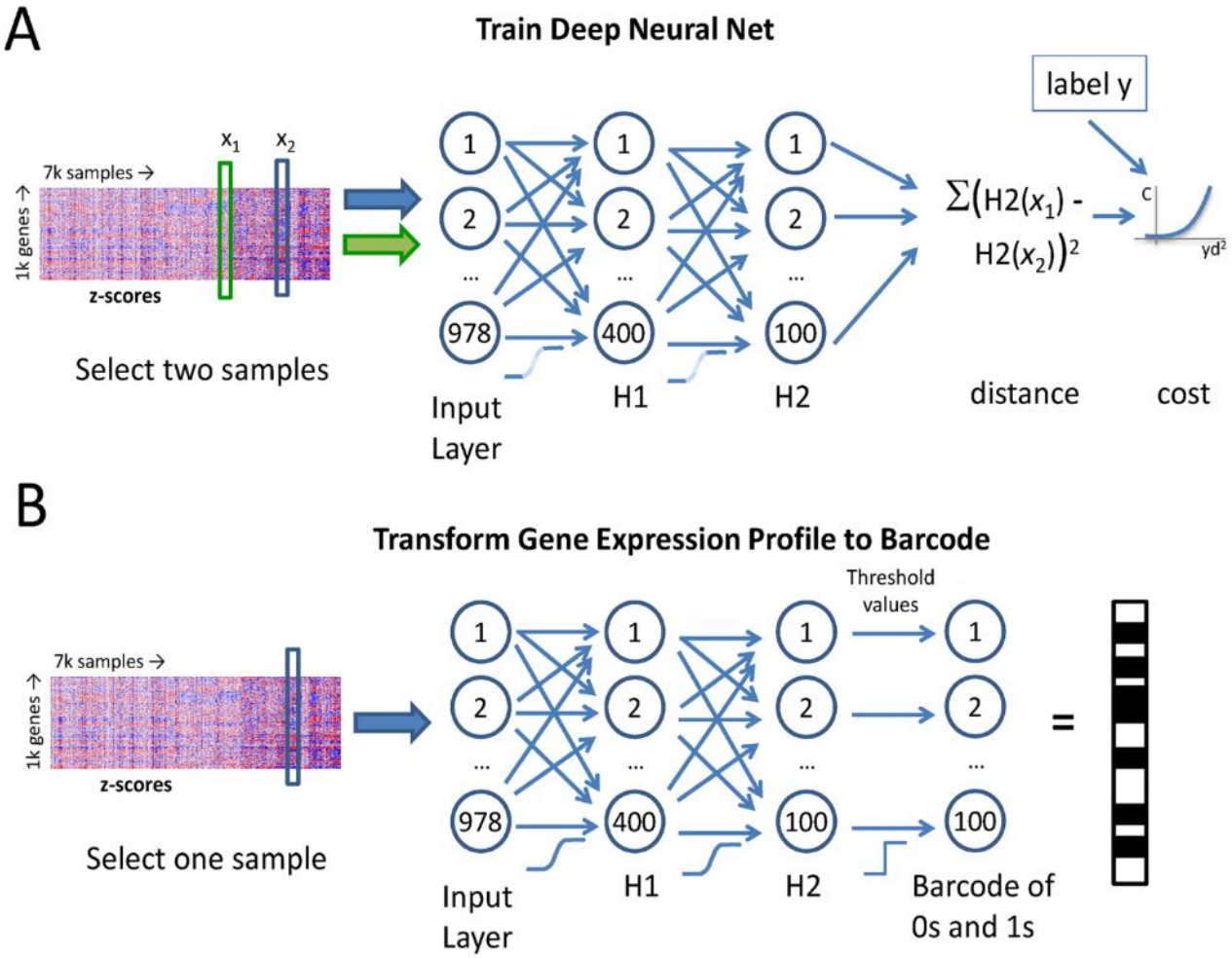


Tan, J., Doing, G., Lewis, K. A., Price, C. E., Chen, K. M., Cady, K. C., Perchuk, B., Laub, M. T., Hogan, D. A. and Greene, C. S. (2017) ‘Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks’, *Cell Systems*. 5(1), p. 63–71.e6.

Target identification

Barcodes from L1000 gene expression (drug perturbation) - method

- L1000 data: expression of ~1000 „landmark genes“ (minimal co-expression)
- Goal:
  - obtain **difference** profiles before and after **drug treatment**
  - condense information into length-100 **binary barcode**
- Calculate similarity between drugs based on L1000-barcodes



Filzen, T. M., Kutchukian, P. S., Hermes, J. D., Li, J. and Tudor, M. (2017) ‘Representing high throughput expression profiles via perturbation barcodes reveals compound targets’, *PLOS Computational Biology*. 13(2), p. e1005335.

Target identification

Barcodes from L1000 gene expression (drug perturbation) - application

– New unknown compounds with **verified activity against MAPK pathway** were identified based on similarity of gene expression profiles to known actives

- t-SNE is a dimensionality reduction algorithm for visualization in 2D
- Z-scores are from L1000 input data
- 100D barcodes were generated by deep neural network
- Orange: known active compounds against MAPK pathway
- Circled: MAPK tool compounds

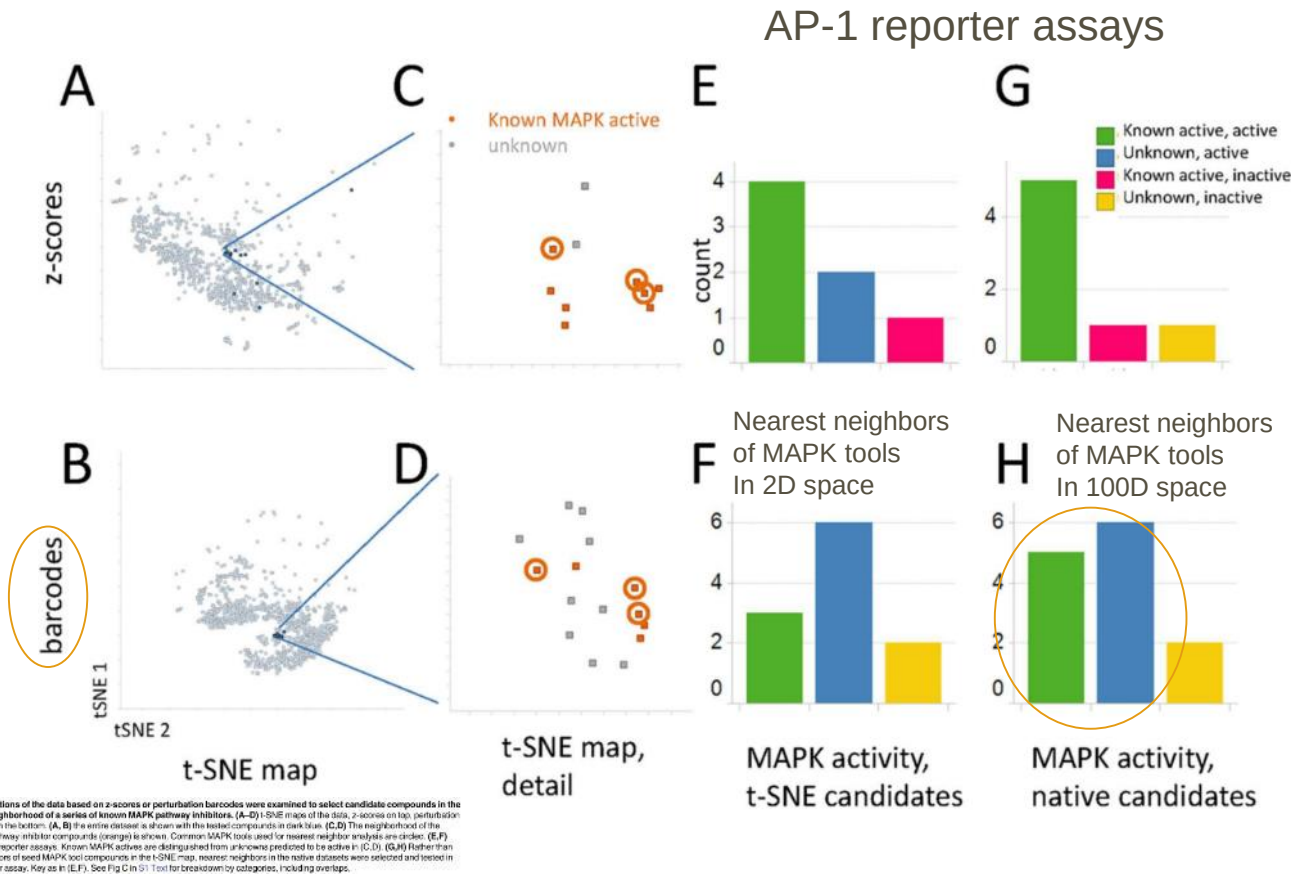


Fig 3. Visualizations of the data based on z-scores or perturbation barcodes were examined to select candidate compounds in the phenotypic neighborhood of a series of known MAPK pathway inhibitors. (A–D) t-SNE maps of the data, z-scores on top, perturbation barcode maps on the bottom. (A, B) the entire dataset is shown with the tested compounds in dark blue. (C,D) The neighborhood of the query MAPK pathway inhibitor compounds (orange) is shown. Common MAPK tools used for nearest neighbor analysis are circled. (E,F) Results of AP-1 reporter assays. Known MAPK actives are distinguished from unknowns predicted to be active in (C,D). (G,H) Rather than selecting neighbors of seed MAPK tool compounds in the t-SNE map, nearest neighbors in the native datasets were selected and tested in the AP-1 reporter assay. Key as in (E,F). See Fig C in S1 Text for breakdown by categories, including overlaps.

Filzen, T. M., Kutchukian, P. S., Hermes, J. D., Li, J. and Tudor, M. (2017) 'Representing high throughput expression profiles via perturbation barcodes reveals compound targets', *PLOS Computational Biology*. 13(2), p. e1005335. (MERCK)

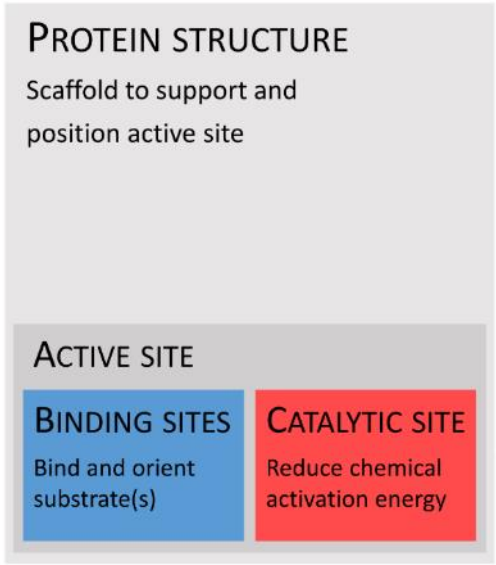
Protein structures

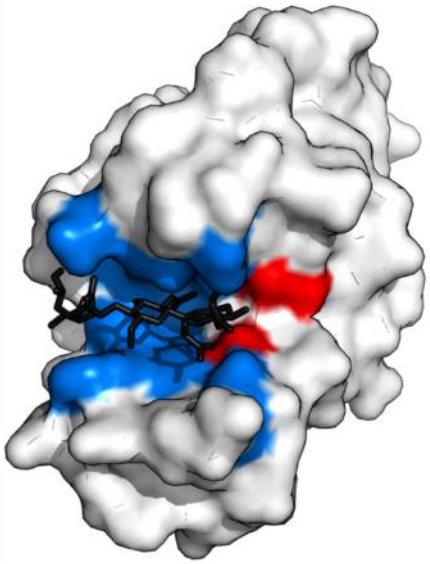
Representing drug targets at molecular detail

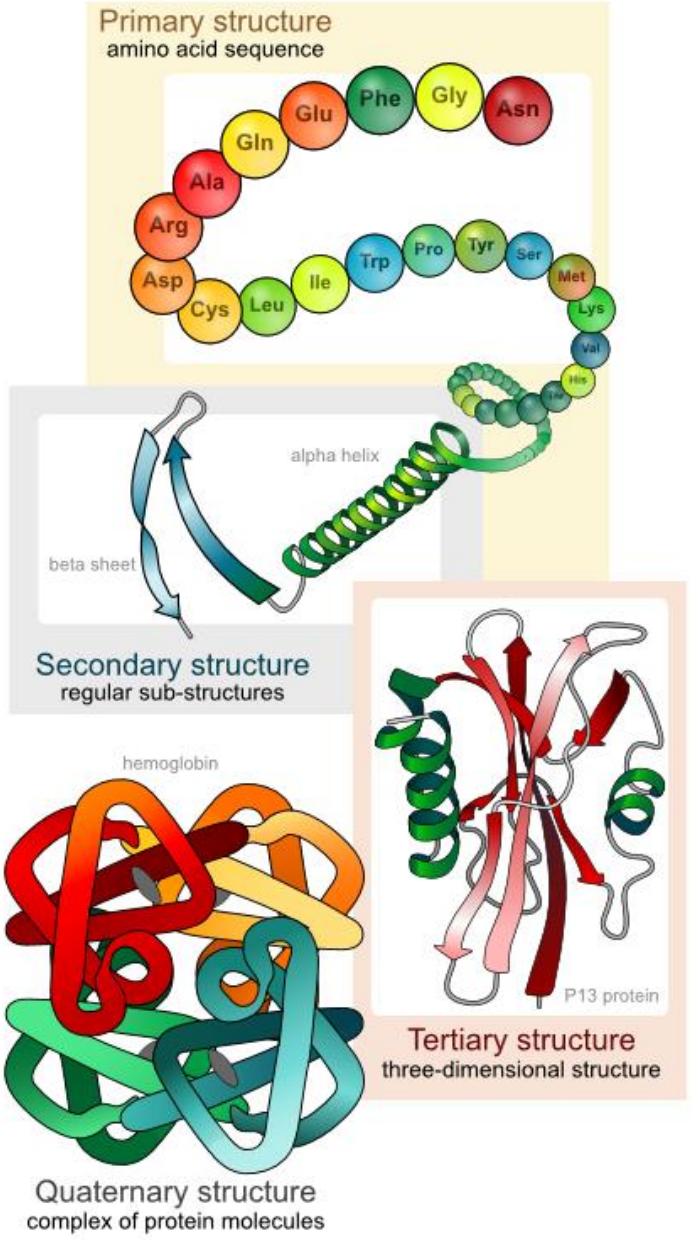
Protein structures

overview

- Most genes hold the instructions for making a particular type of protein
- Proteins are complex molecules that can be described at different levels of complexity:
  - Sequence of letters (amino acids, secondary structure)
  - List of 3D coordinates (multiple atoms per amino acid)
  - Interactions between proteins (and other molecules, e.g. drugs)







https://en.wikipedia.org/wiki/File:Main_protein_structure_levels_en.svg; https://en.wikipedia.org/wiki/Active_site#/media/File:Enzyme_structure.svg

Protein structures

Encoding protein sequences

- Challenge for deep learning:
  - length of protein sequence & size of 3D structure are **variable**
  - machine learning models often expect fixed-length input layer
- Variable-length protein → **fixed-length input**:
  - Break sequences into artificial chunks
    - Problem: often protein needs to be studied in its entirety
  - Choose input size <= longest sequence, buffer rest with „zeros“
    - Problem: wasteful

Protein structures

Encoding protein sequences

- ProtVec: borrows concepts from Natural Language Processing (NLP) – „Word2Vec“
  - Full protein sequence („sentence“) is broken down into three-letter „words“
  - Each sentence-vector can be represented as a linear combination of word-vectors

- Treat amino acid sequence as a „sentence“, AA triplets as „words“

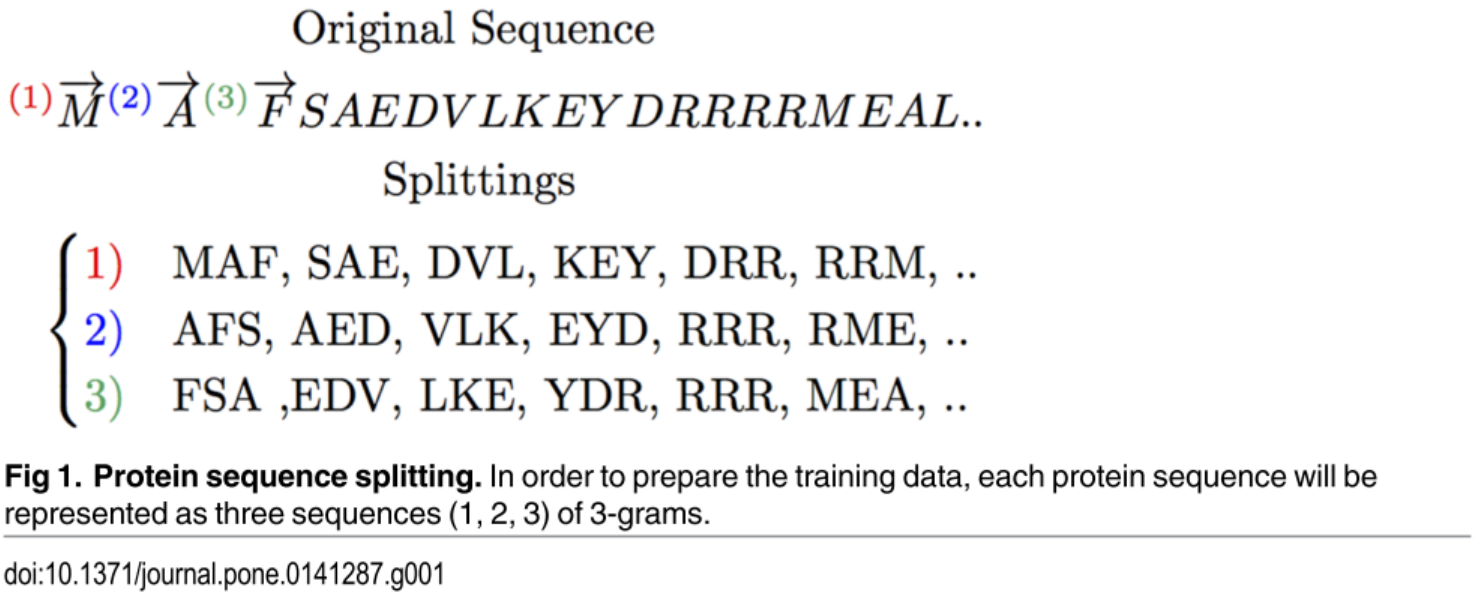


Fig 1. Protein sequence splitting. In order to prepare the training data, each protein sequence will be represented as three sequences (1, 2, 3) of 3-grams.

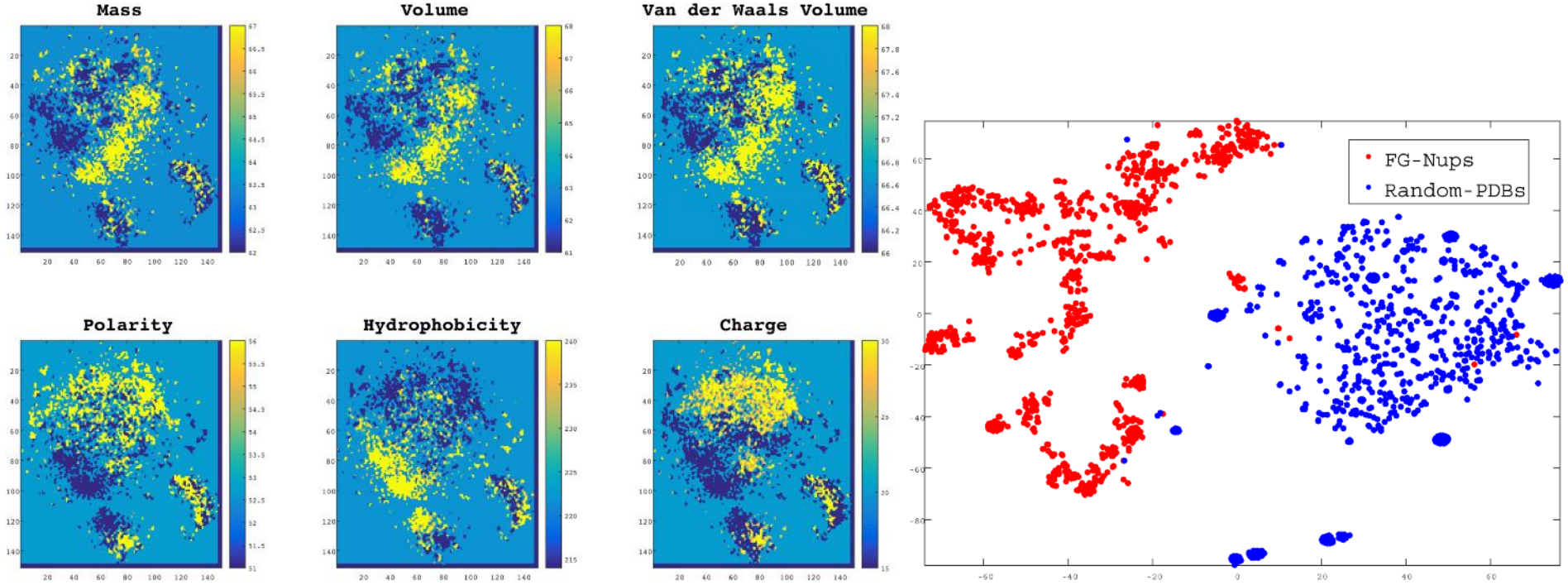
doi:10.1371/journal.pone.0141287.g001

Asgari, E. and Mofrad, M. R. (2015) ‘Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics’, *PLoS One*, 10(11), p. e0141287. doi: 10.1371/journal.pone.0141287.

Protein structures

Encoding protein sequences

- t-SNE: **2D maps** of protein space with ProtVec as input (derived from AA sequence only)
- Accurately clusters proteins based on **phys-chem properties** (left) and **disorder (proteins with no stable structure)** (right)

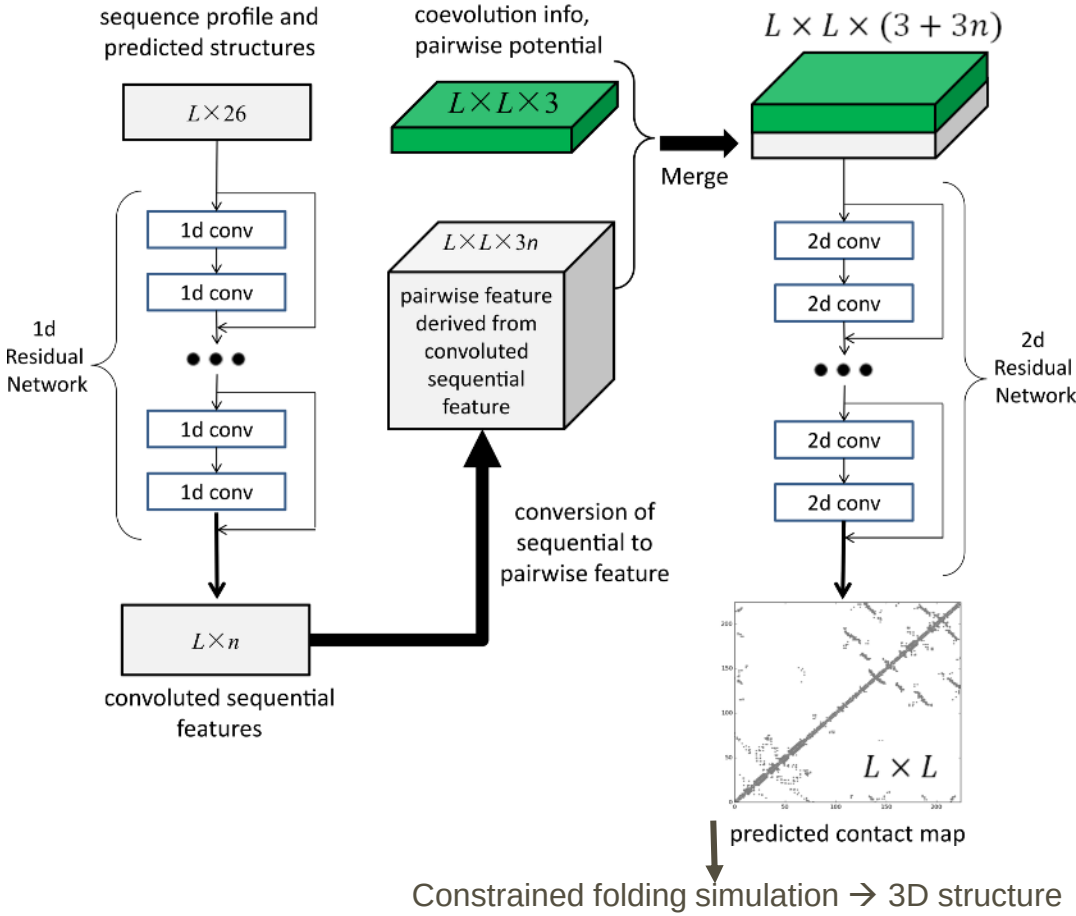


Asgari, E. and Mofrad, M. R. (2015) 'Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics', *PLoS One*, 10(11), p. e0141287. doi: 10.1371/journal.pone.0141287.

Protein structures

Predict protein structure based on sequence (and derived features)

- **Input** features: L=sequence length of protein
  - **Sequential** (L x 26)
    - Position-specific scoring matrix (20)
    - Predicted 3-state secondary structure (3)
    - Predicted 3-state solvent accessibility (3)
  - **Pairwise** (LxLx3)
    - Co-evolutionary information (CCMpred)
    - Mutual information
    - Mijazawa-Jernigan contact potential
- **Output**:
  - Pairwise amino-acid contact map

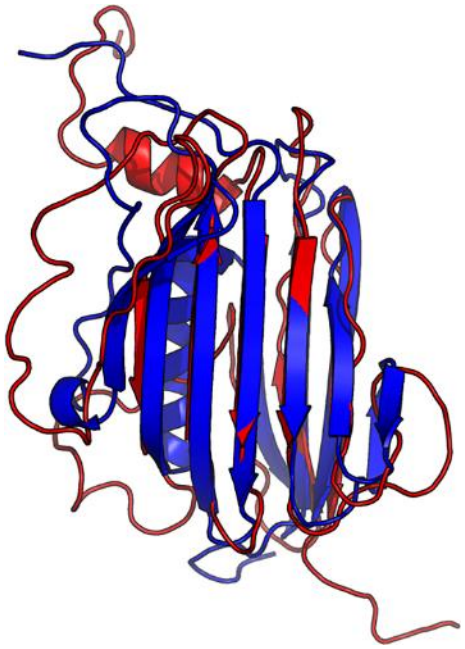


Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J. (2017) 'Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model', *PLOS Computational Biology*. Edited by A. Schlessinger, 13(1), p. e1005324. doi: 10.1371/journal.pcbi.1005324.

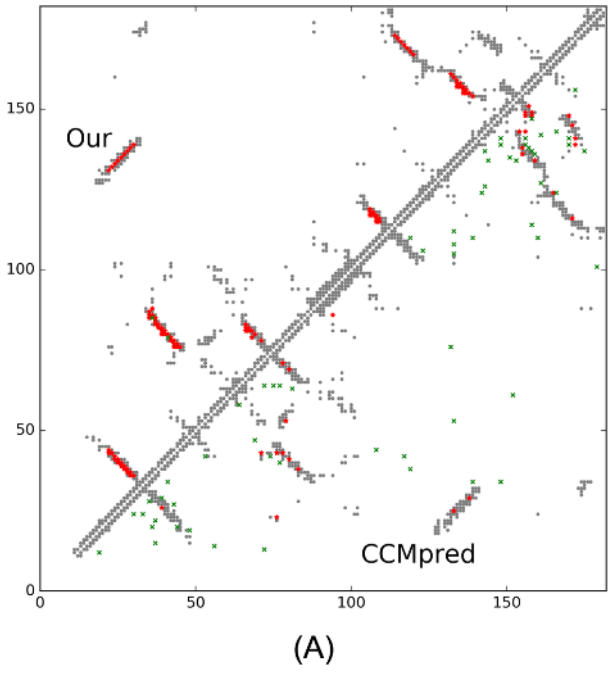
Protein structures

Predict protein structure based on sequence (and derived features)

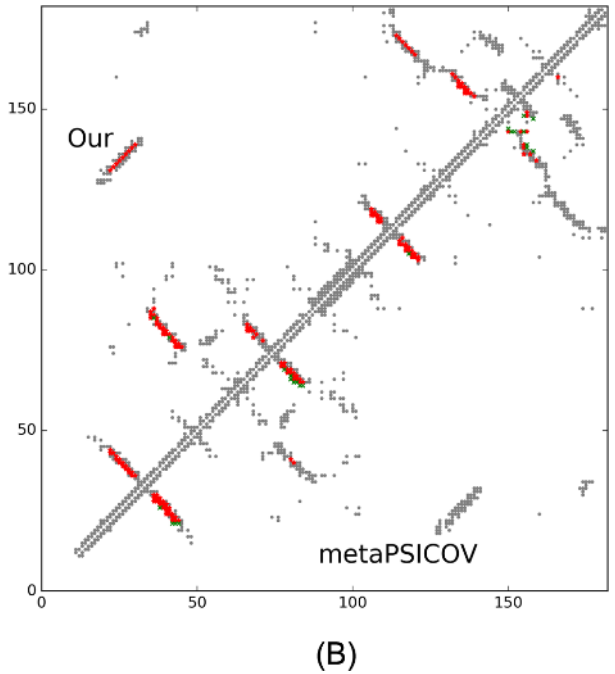
Improved prediction of long-range contacts (distant along sequence, close in 3D)



Superimposition between predicted model (red) and its native structure (blue) for the CAMEO test protein (PDB ID 2nc8 and chain A).



(A)



(B)

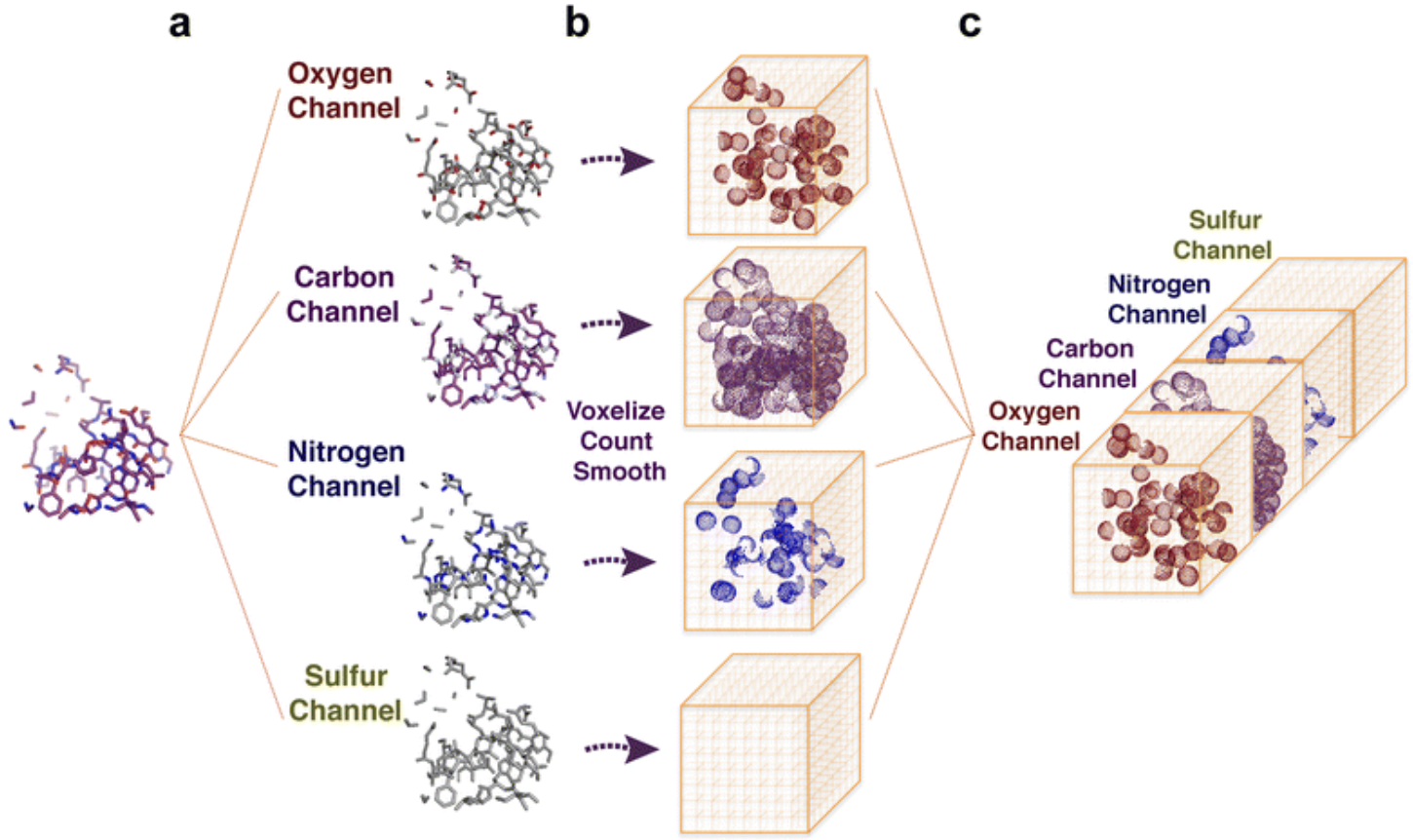
Overlap between top L/2 predicted contacts (in red or green) and the native contactmap (in grey) for CAMEOtarget 2nc8A. Red (green) dots indicate correct (incorrect) prediction. (A) The comparison between our prediction (in upper-left triangle) and CCMpred (in lower-right triangle). (B) The comparison between our prediction (in upper-left triangle) and MetaPSICOV (in lower-right triangle).

Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J. (2017) 'Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model', *PLOS Computational Biology*. Edited by A. Schlessinger, 13(1), p. e1005324. doi: 10.1371/journal.pcbi.1005324.

Protein structures

Amino acid substitutions: 3D atomic coordinates; 3D Conv Net (3D-CNN)

- Focus box on atomic coordinates of amino acid; four atom types (C,O,N,S)
- No „hand-engineered“ features, i.e. the network determines relevant features from raw data

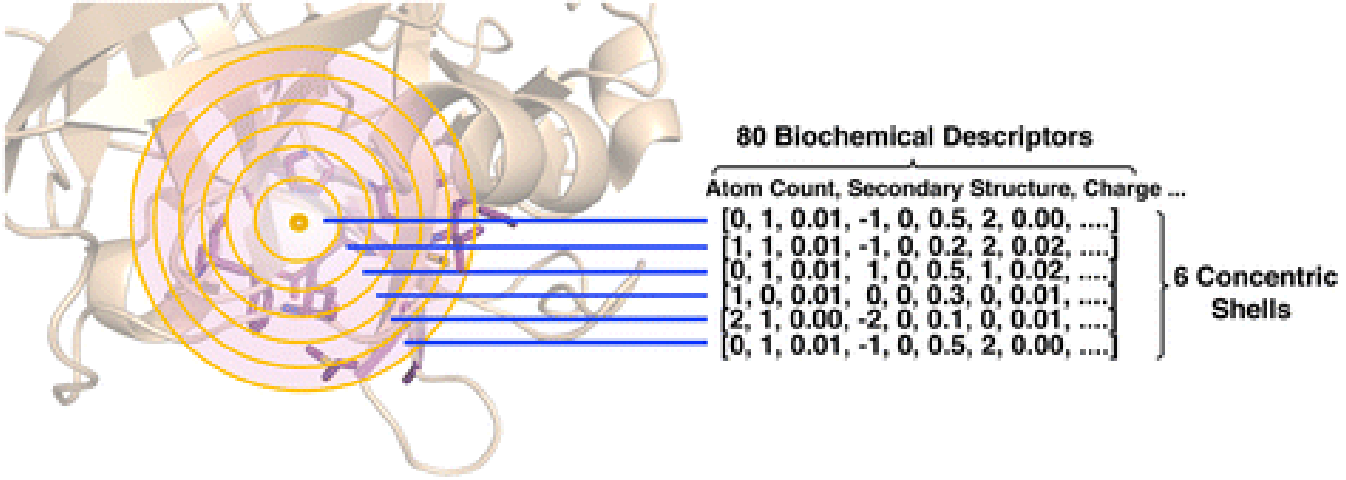


Torng, W. and Altman, R. B. (2017) ‘3D deep convolutional neural networks for amino acid environment similarity analysis’, *BMC Bioinformatics*. 18(1), p. 302.

Protein structures

Amino acid substitutions: 3D atomic coordinates; Conv Net

- Alternative method:
  - Convert local environment of 3D point into numeric vector
  - Exact structure not preserved

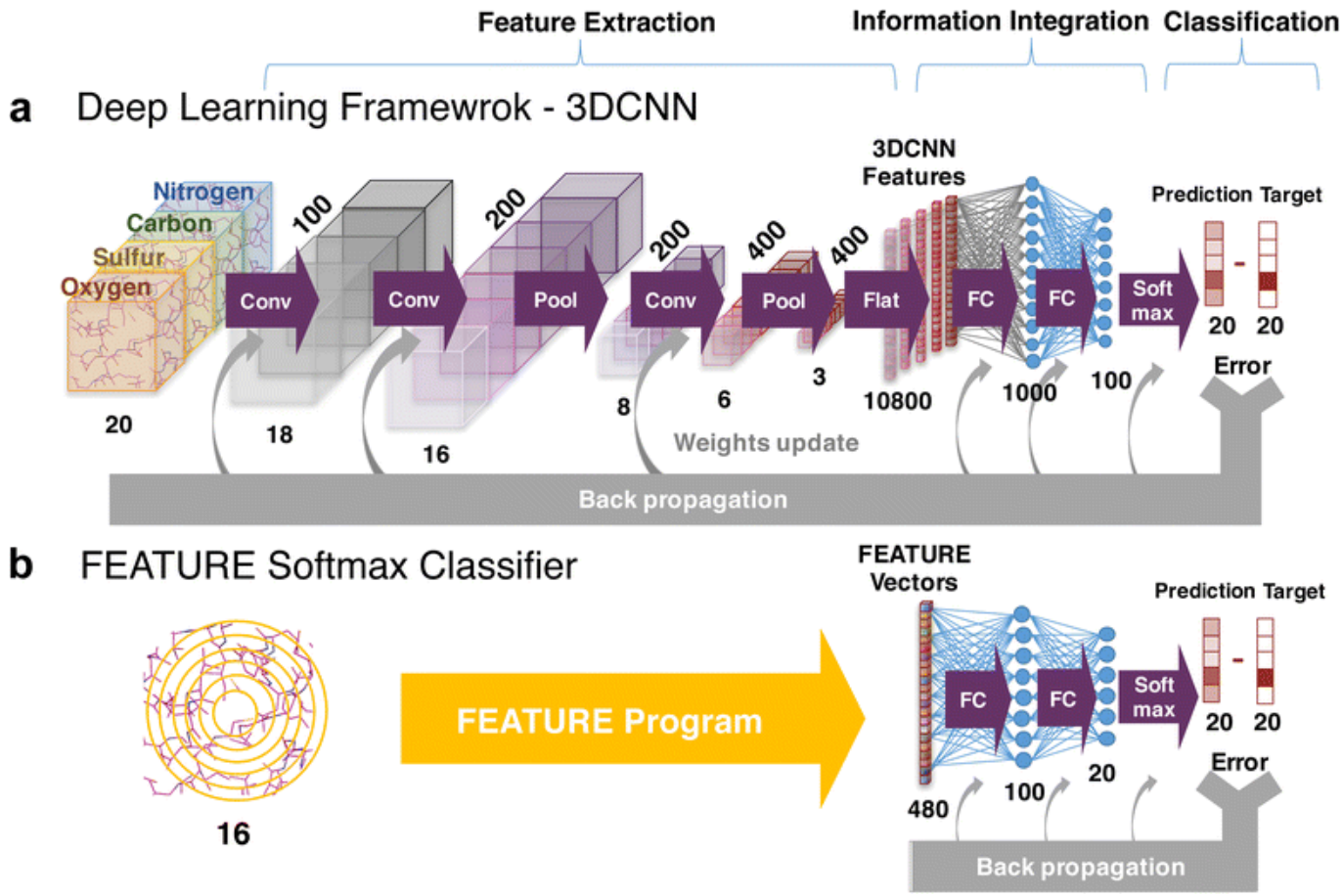


Torng, W. and Altman, R. B. (2017) '3D deep convolutional neural networks for amino acid environment similarity analysis', *BMC Bioinformatics*. 18(1), p. 302.

Protein structures

Amino acid substitutions: 3D atomic coordinates; Conv Net

– Learn to predict effect of mutations on protein structure (two alternative approaches)

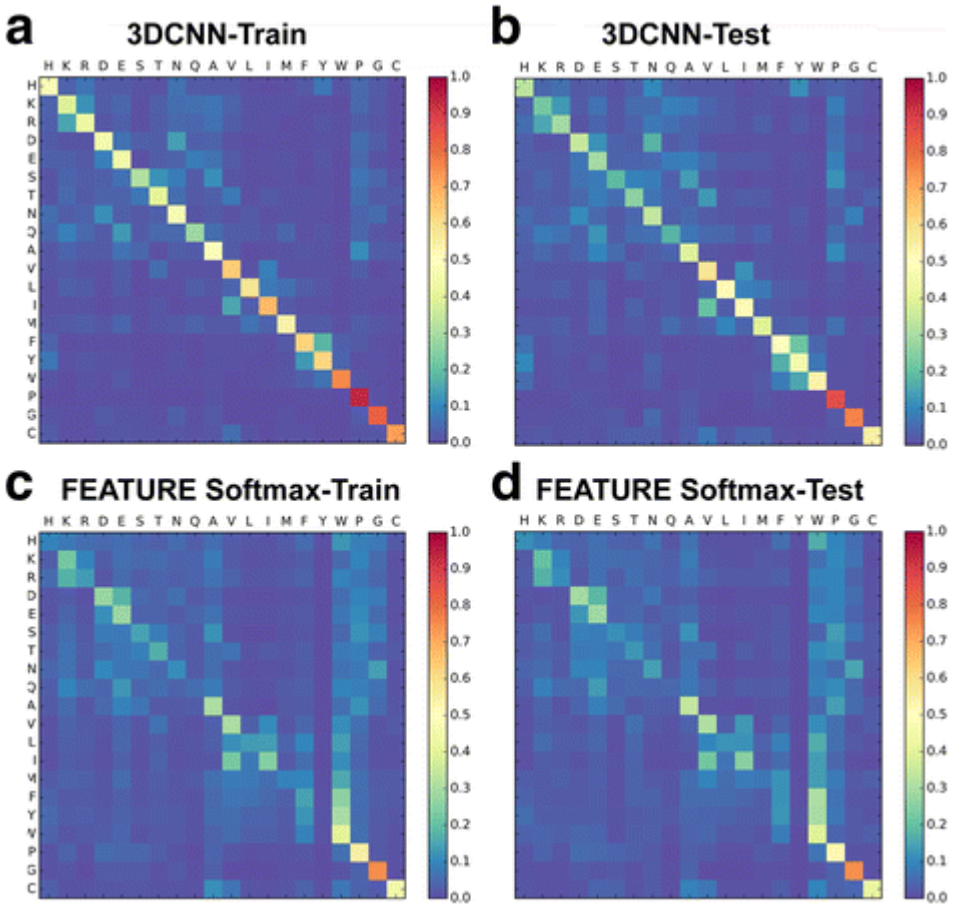


Torng, W. and Altman, R. B. (2017) '3D deep convolutional neural networks for amino acid environment similarity analysis', *BMC Bioinformatics*. 18(1), p. 302.

Protein structures

Amino acid substitutions: 3D atomic coordinates; Conv Net

3D Conv nets are superior for predicting amino acid changes



Confusion matrices for predictions of the 20 amino acid microenvironments.

Heatmap: probability of examples of true label i being predicted as label j.

Torng, W. and Altman, R. B. (2017) '3D deep convolutional neural networks for amino acid environment similarity analysis', *BMC Bioinformatics*. 18(1), p. 302.

Protein structures

Amino acid substitutions: 3D atomic coordinates; Conv Net

3DCNN agrees on which T4 lyzozyme mutants are **destabilizing or neutral**
Comparison: predicted vs actual amino acid at given position for wildtype and mutant

Class	Variant	Wild type PDB ID	Wild True	Wild Predicted	Mutant PDB ID	Mutant True	Mutant Predicted
Destabilizing	M102K	2LZM	MET	MET	1L54	LYS	MET
Destabilizing	L99G	2LZM	LEU	LEU	1QUD	GLY	LEU
Destabilizing	V149S	2LZM	VAL	VAL	1G06	SER	ILE
Destabilizing	V149C	2LZM	VAL	VAL	1G07	CYS	VAL
Destabilizing	V149G	2LZM	VAL	VAL	1G0P	GLY	VAL
Destabilizing	T157I	2LZM	THR	THR	1L10	ILE	THR
Destabilizing	G156D	2LZM	GLY	GLY	1L16	ASP	GLY
Destabilizing	R96H	2LZM	ARG	ARG	1L34	HIS	ARG
Destabilizing	D92N	2LZM	ASP	ASP	1L55	ASN	ASP
Destabilizing	I3P	2LZM	ILE	ILE	1L97	PRO	ILE
Destabilizing	V87M	2LZM	VAL	VAL	1CU3	MET	PHE
Destabilizing	R96N	2LZM	ARG	ARG	3CDT	ASN	ARG
Destabilizing	R96D	2LZM	ARG	ARG	3C8Q	ASP	ARG
Destabilizing	R96W	2LZM	ARG	ARG	3FI5	TRP	PHE
Destabilizing	R96Y	2LZM	ARG	ARG	3C80	TYR	ARG
Destabilizing	M102L	2LZM	MET	MET	1L77	LEU	ILE
Destabilizing	M106K	2LZM	MET	LEU	231L	LYS	MET
Destabilizing	M120K	2LZM	MET	GLU	232L	LYS	TYR
Destabilizing	I3Y	2LZM	ILE	ILE	1L18	TYR	TYR
Neutral	G77A	2LZM	GLY	ALA	1L23	ALA	ARG
Neutral	A82P	2LZM	ALA	PRO	1L24	PRO	PRO
Neutral	A93T	2LZM	ALA	ASP	129L	THR	ASP
Neutral	T151S	2LZM	THR	GLU	130L	SER	ASP
Neutral	T26S	2LZM	THR	ALA	131L	SER	ALA
Neutral	V149M	2LZM	VAL	VAL	1CV6	MET	ILE
Neutral	S38N	2LZM	SER	SER	1L61	ASN	ASP
Neutral	S38D	2LZM	SER	SER	1L19	ASP	ASP
Neutral	N144D	2LZM	ASN	GLU	1L20	ASP	GLU
Neutral	T109N	2LZM	THR	GLU	1L59	ASN	ALA
Neutral	T109D	2LZM	THR	GLU	1L62	ASP	GLU
Neutral	N116D	2LZM	ASN	ASP	1L57	ASP	ASP
Neutral	M106I	2LZM	MET	LEU	1P46	ILE	VAL
Neutral	M120Y	2LZM	MET	GLU	1P6Y	TYR	TYR
Neutral	V149I	2LZM	VAL	VAL	1G0Q	ILE	ILE
Neutral	T152V	2LZM	THR	VAL	1G0L	VAL	VAL
Neutral	E108V	2LZM	GLU	ILE	1QUG	VAL	VAL
Neutral	S117F	2LZM	SER	ALA	1TLA	PHE	PHE
Neutral	M106L	2LZM	MET	LEU	234L	LEU	LEU
Neutral	I3V	2LZM	ILE	ILE	1L17	VAL	ILE
Neutral	M120L	2LZM	MET	GLU	233L	LEU	LEU

Torng, W. and Altman, R. B. (2017) '3D deep convolutional neural networks for amino acid environment similarity analysis', *BMC Bioinformatics*. 18(1), p. 302.

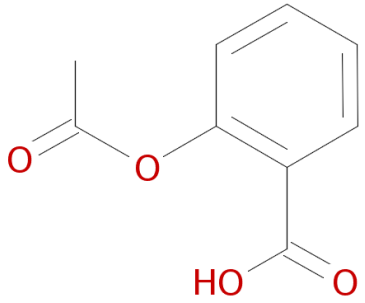
Small molecules

Representing small drug-like molecules for machine learning

Small molecules

Conventional representations

- Molecular structure graph



aspirin

- SMILES string
  - CC(=O)Oc1ccccc1C(=O)O
- Bit vector fingerprint
  - Different methods (MACCS, Morgan,...) → CDK toolkit, Python package RDKIT
  - Fixed length
  - 1 or 0
  - presence and absence of molecular features
  - can be used directly as input for ML

https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL25

Small molecules

Chemception: Learning chemistry from 2D drawings; Tox prediction

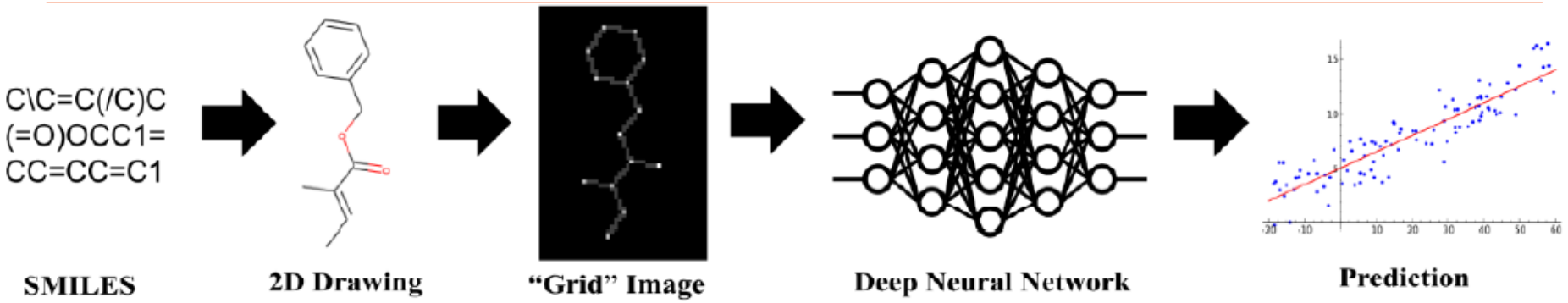


Table 3: Summary of Results for Tox21 trained on Chemception T1 network.

	Train AUC			Validation AUC			Test AUC			
nr-ahr	0.825	+/-	0.018	0.779	+/-	0.015	0.800	+/-	0.020	Y
nr-ar	0.843	+/-	0.010	0.797	+/-	0.049	0.757	+/-	0.029	Y
nr-ar-lbd	0.887	+/-	0.034	0.834	+/-	0.046	0.886	+/-	0.014	Y
nr-aromatase	0.801	+/-	0.010	0.759	+/-	0.027	0.799	+/-	0.016	Y
nr-er	0.747	+/-	0.020	0.710	+/-	0.023	0.694	+/-	0.013	Y
nr-er-lbd	0.824	+/-	0.029	0.765	+/-	0.036	0.762	+/-	0.009	Y
nr-ppar-gamma	0.791	+/-	0.038	0.742	+/-	0.025	0.819	+/-	0.015	Y
sr-are	0.724	+/-	0.009	0.702	+/-	0.025	0.654	+/-	0.009	N
sr-atad55	0.841	+/-	0.022	0.759	+/-	0.048	0.776	+/-	0.011	Y
sr-hse	0.776	+/-	0.032	0.732	+/-	0.013	0.717	+/-	0.018	N
sr-mmp	0.791	+/-	0.020	0.759	+/-	0.016	0.755	+/-	0.010	Y
sr-p53	0.844	+/-	0.034	0.782	+/-	0.036	0.776	+/-	0.011	Y
Tox21	**0.808**		**0.044**	**0.760**		**0.035**	**0.766**		**0.058**	

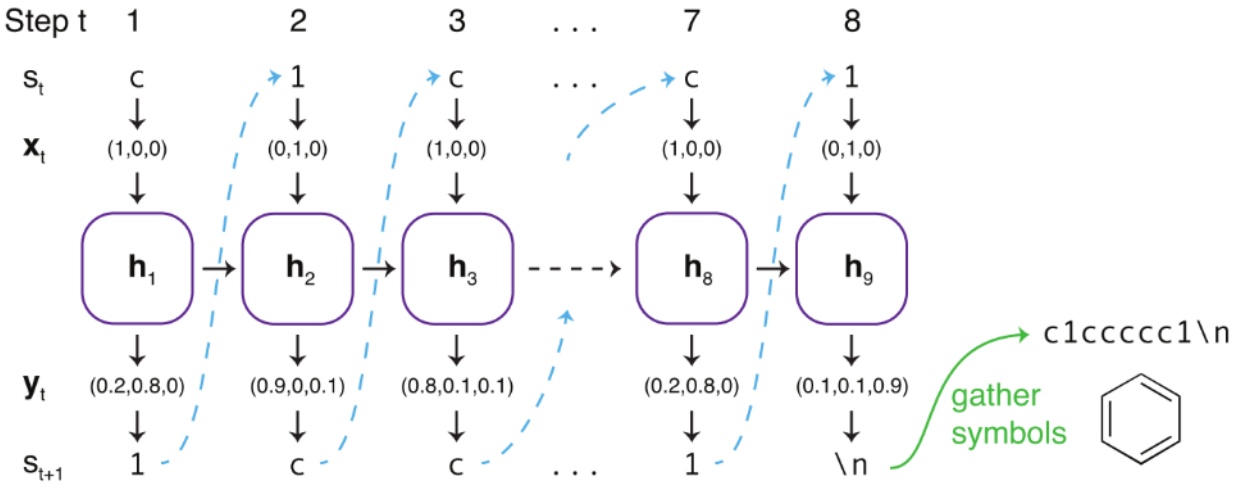
Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O. and Baker, N. (2017) 'Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models', pp. 1–38. Available at: http://arxiv.org/abs/1706.06689.

Small molecules

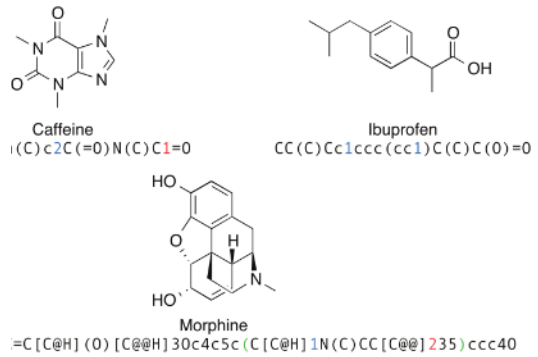
Generating novel compounds: recurrent neural networks

- Train RNN (recurrent neural network) model on SMILES strings from ChEMBL (1.4 M molecules)
- 72 M SMILES characters from vocabulary of 51 different characters (one-hot encoded)
- Apply filters to check that output is valid SMILES and drug-like properties (filters)

RNN: maps inputs x to outputs y, via hidden layer h connected to previous time steps (i.e. SMILES characters)



Example input molecules with SMILES strings



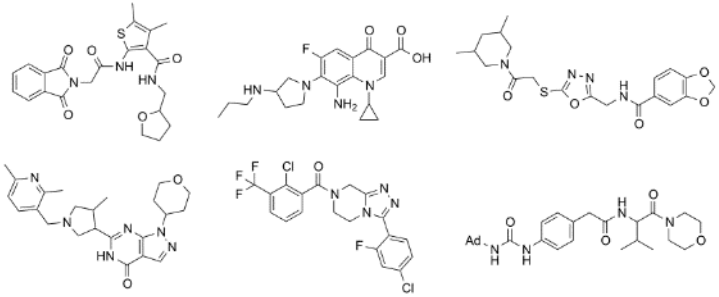
Segler, M. H. S., Kogej, T., Tyrchan, C. and Waller, M. P. (2017) 'Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks', pp. 1–17. Available at: http://arxiv.org/abs/1701.01329.

Small molecules

Generating novel compounds: recurrent neural networks

- Properties of novel molecules (848000):
  - Phys-chem descriptors very similar to ChEMBL
  - 75% suitable for high-throughput screen in Pharma
  - But new scaffolds (core molecular structure) proposed

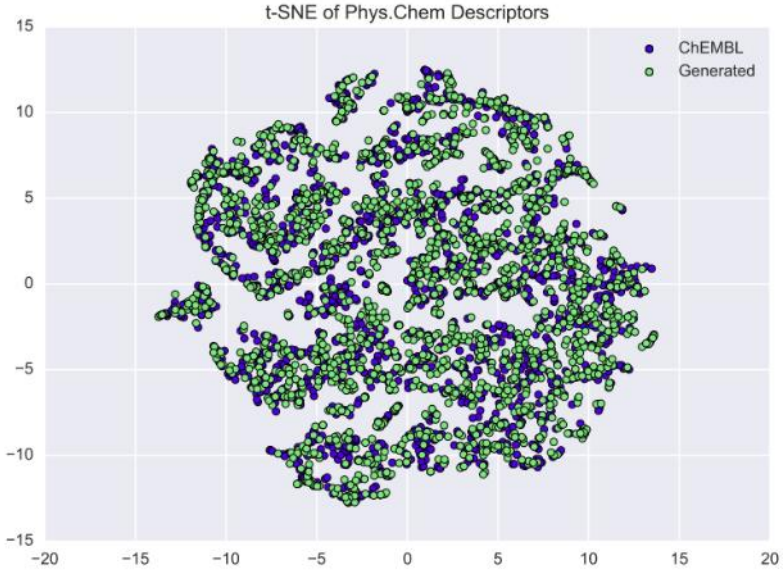
Example output molecules



Valid SMILES emerge over training time

Batch	Generated Example	valid
0	Oc.BK5i%ur+7oAFc7L3T=F8B5e=n)CS6RCTAR((OVCp1CApb)	no
1000	OF=CCC2OCCCC)C2)C1CNC2CCCCCCCCCCCCCCCCCCCCCCC	no
2000	O=C(N)C(=O)N(c1occc1OC)c2ccccc2OC	yes
3000	O=C1C=2N(c3cc(ccc3OC2CCC1)CCCc4cn(c5c(Cl)cccc54)C)C	yes

Generated molecules like ChEMBL



Segler, M. H. S., Kogej, T., Tyrchan, C. and Waller, M. P. (2017) 'Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks', pp. 1–17. Available at: http://arxiv.org/abs/1701.01329.

Small molecules

Generating novel compounds: recurrent neural networks

- Transfer learning:
  - Fine-tuning can be applied to create target-specific predictors
  - Take already trained model (1.4M from ChEMBL) and re-train on known actives for target protein of interest

Active molecules for specific target re-discovered after few additional training epochs with pre-trained model

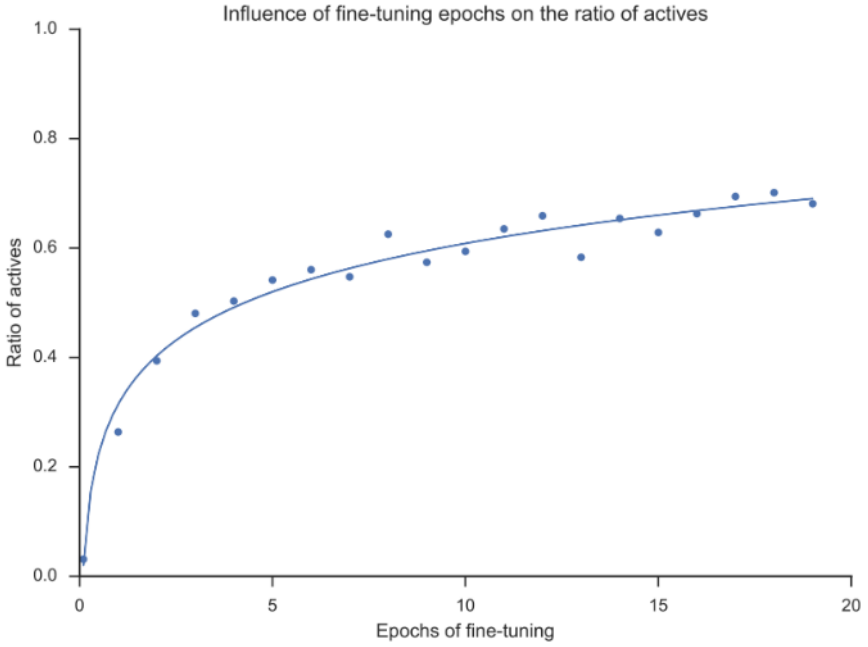


Table 2 Reproducting known actives in the *Plasmodium* test set. EOR: Enrichment over random.

#	pIC_{50}	Train.	Test	Gen. mols.	Reprod.	EOR
1	> 8	1239	1240	128,256	28%	66.9
2	> 8	100	1240	93,721	7%	19.0
3	> 9	100	1022	91,034	11%	35.7

Table 3 Reproducting known actives in the *Staphylococcus* test set. EOR: Enrichment over random.

Entry	*p*MIC	Train.	Test	Gen. mols.	Reprod.	EOR
1	> 3	1000	6051	51,052	14%	155.9
2	> 3	50	7001	70,891	2.5%	21.6
3^a	> 3	50	7001	85,755	1.8%	6.3
4^b	> 3	50	7001	285	0%	—
5^c	> 3	0	7001	60,988	6%	59.6

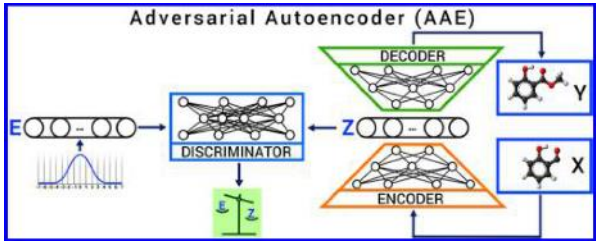
[a] Fine-tuning learning rate $= 10^{-4}$. [b] No Pretraining. [c] 8 Generate-Test cycles.

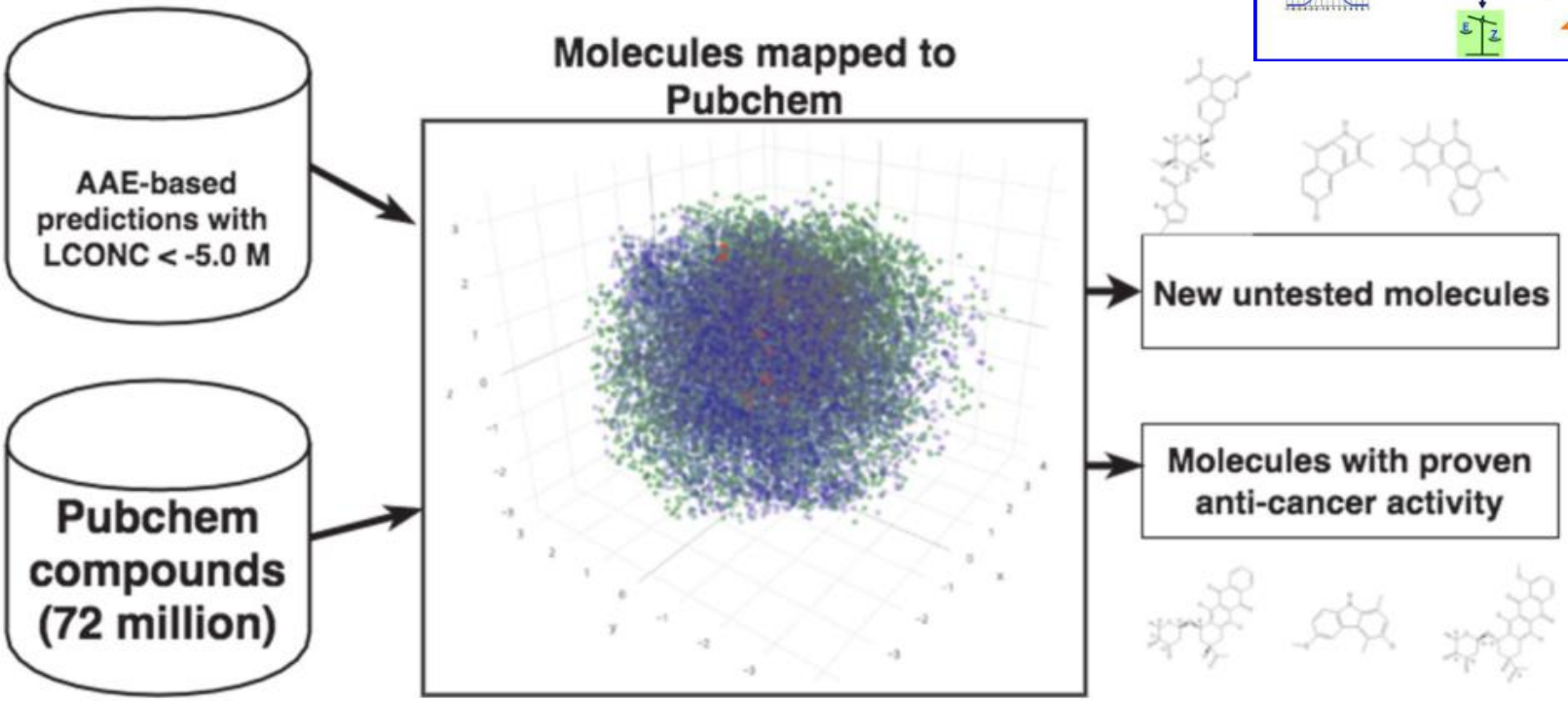
Segler, M. H. S., Kogej, T., Tyrchan, C. and Waller, M. P. (2017) 'Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks', pp. 1–17. Available at: http://arxiv.org/abs/1701.01329.

Small molecules

Generating novel compounds: Adversarial autoencoders

– Train on 6252 compounds profiled in MCF-7 cell lines (breast cancer)

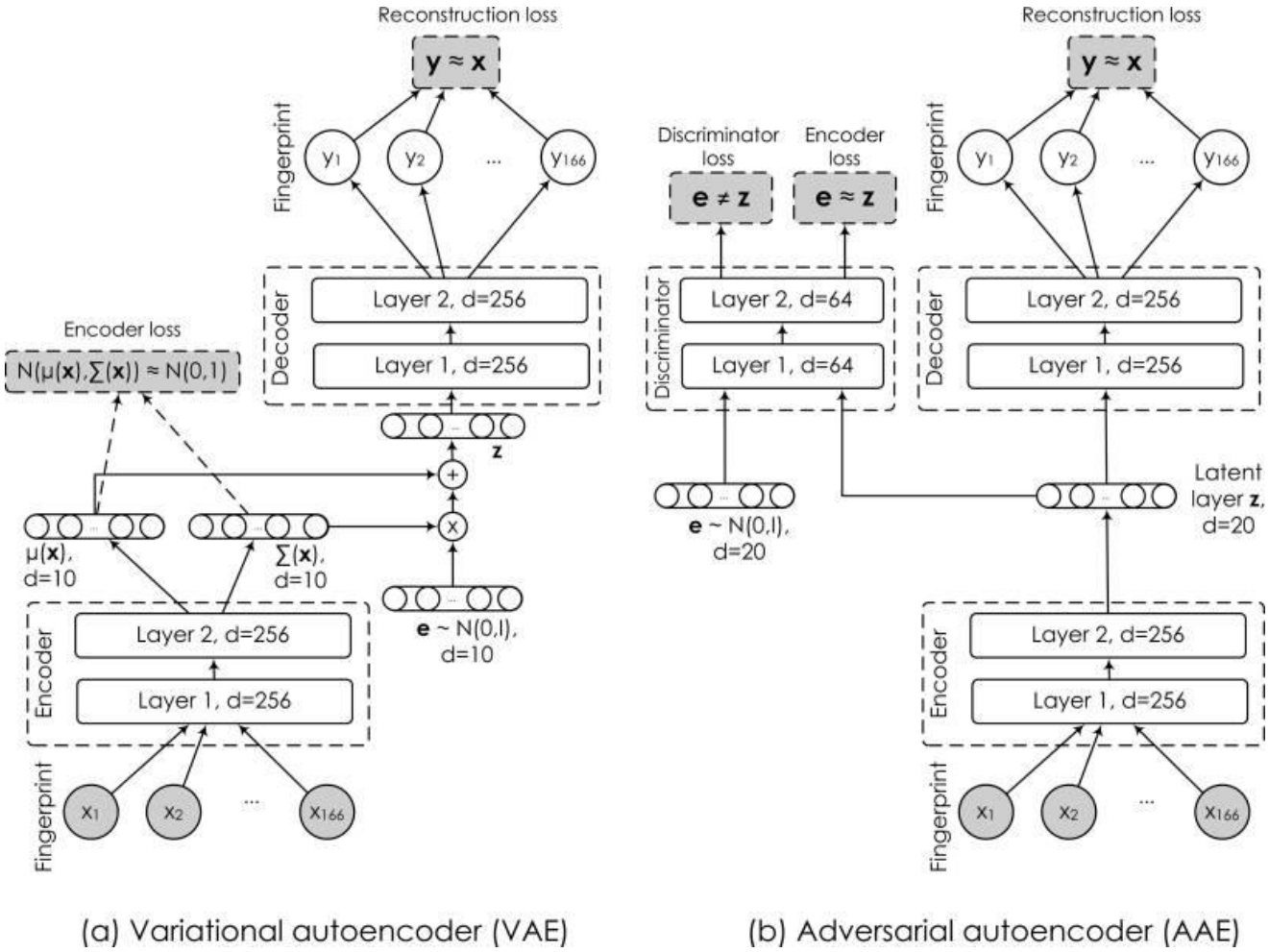




Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K. and Zhavoronkov, A. (2016) 'The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology', *Oncotarget*, 5(0). doi: 10.18632/oncotarget.14073.

Small molecules

Generating novel compounds: Adversarial autoencoders



Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. and Zhavoronkov, A. (2017) 'DruGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico', *Molecular Pharmaceutics*, 14(9), pp. 3098–3104.

Drug-target interactions

How does the small molecule bind to the target protein?

Drug-target interactions

DL-based scoring function of binding

- Again 3D convolution network
- focus on binding site
- Learn to score the binding interaction
- Compare against physics-based scoring function (AutoDock Vina)

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. and Koes, D. R. (2017) 'Protein-Ligand Scoring with Convolutional Neural Networks', *Journal of Chemical Information and Modeling*, 57(4), pp. 942–957. doi: 10.1021/acs.jcim.6b00740.

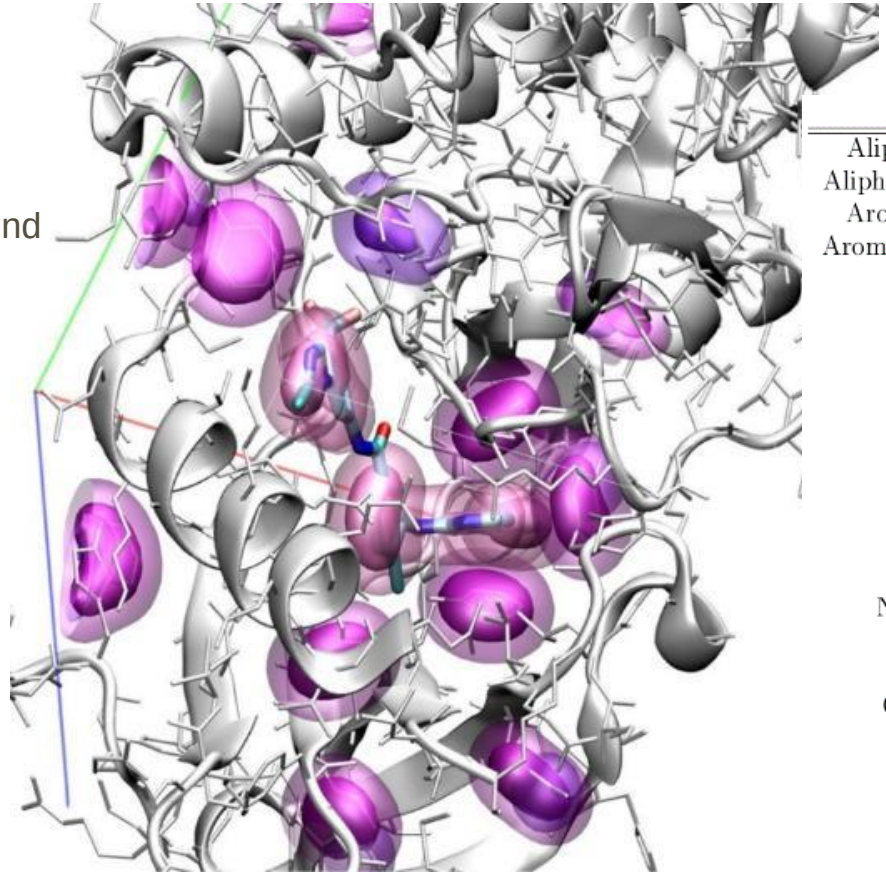
Drug-target interactions

DL-based scoring function of binding

Input: 3D grid 24x24x24

34 atom type channels

Density distribution around atom center



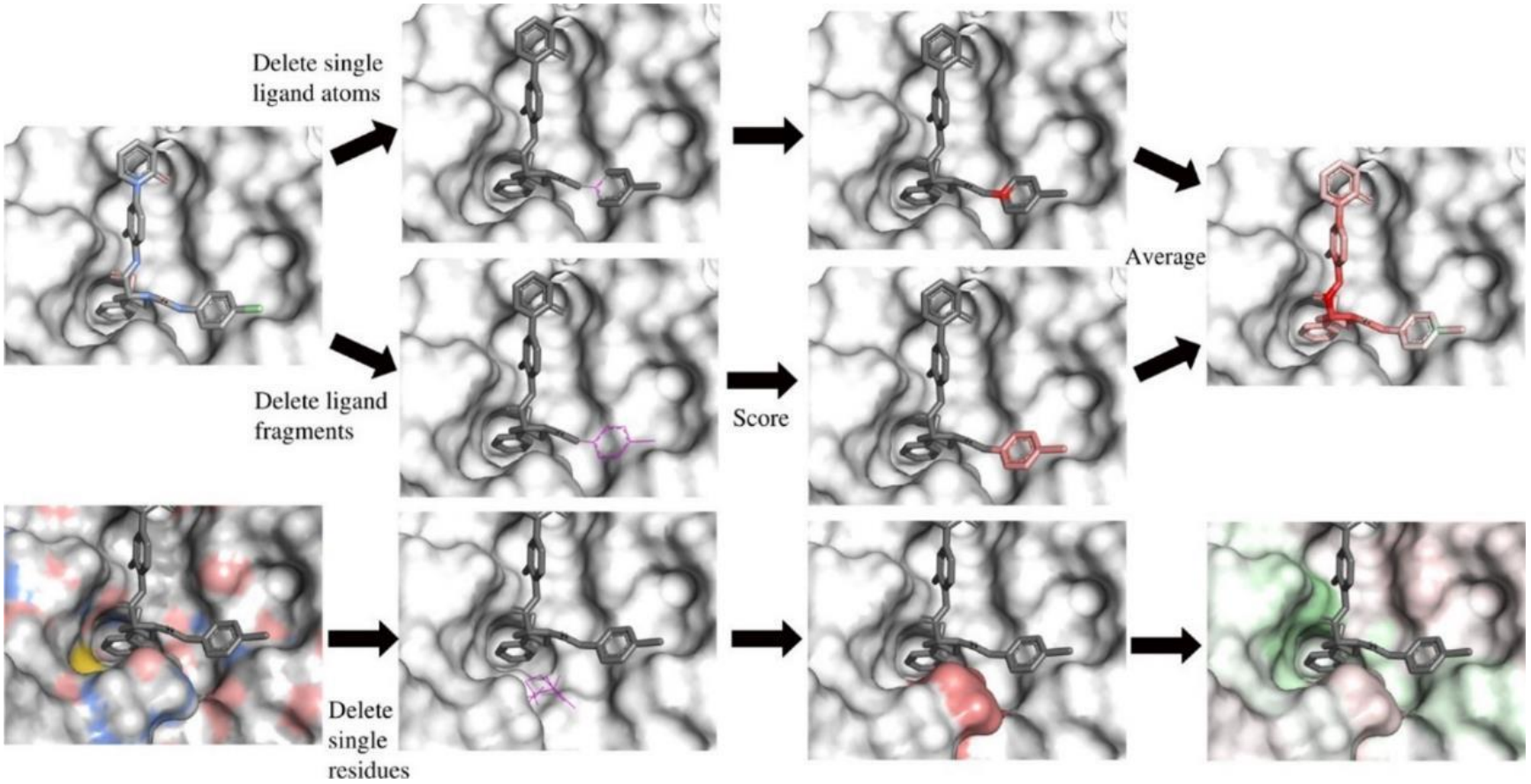
Type	Ligand	Receptor
AliphaticCarbonXSHydrophobe	Y	Y
AliphaticCarbonXSNonHydrophobe	Y	Y
AromaticCarbonXSHydrophobe	Y	Y
AromaticCarbonXSNonHydrophobe	Y	Y
Bromine	Y	N
Calcium	N	Y
Chlorine	Y	N
Fluorine	Y	N
Iodine	Y	N
Iron	N	Y
Magnesium	N	Y
Nitrogen	Y	Y
NitrogenXSAcceptor	Y	Y
NitrogenXSDonor	Y	Y
NitrogenXSDonorAcceptor	Y	Y
Oxygen	Y	N
OxygenXSAcceptor	Y	Y
OxygenXSDonorAcceptor	Y	Y
Phosphorus	Y	Y
Sulfur	Y	Y
SulfurAcceptor	Y	N
Zinc	N	Y

Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. and Koes, D. R. (2017) 'Protein–Ligand Scoring with Convolutional Neural Networks', *Journal of Chemical Information and Modeling*, 57(4), pp. 942–957. doi: 10.1021/acs.jcim.6b00740.

Drug-target interactions

DL-based scoring function of binding

Atom importance: which parts of the molecule and protein contribute most to binding score (strategy: modify input to understand contribution to output)

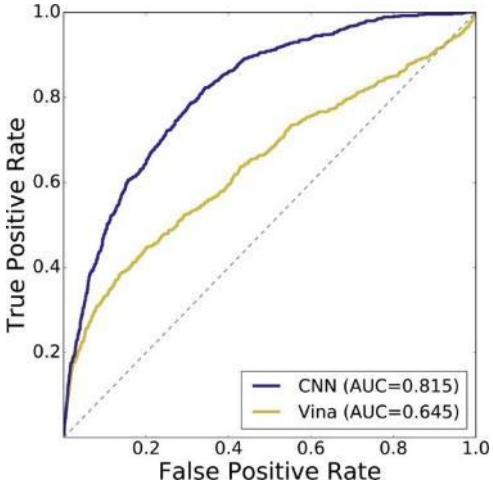


Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. and Koes, D. R. (2017) 'Protein–Ligand Scoring with Convolutional Neural Networks', *Journal of Chemical Information and Modeling*, 57(4), pp. 942–957. doi: 10.1021/acs.jcim.6b00740.

Drug-target interactions

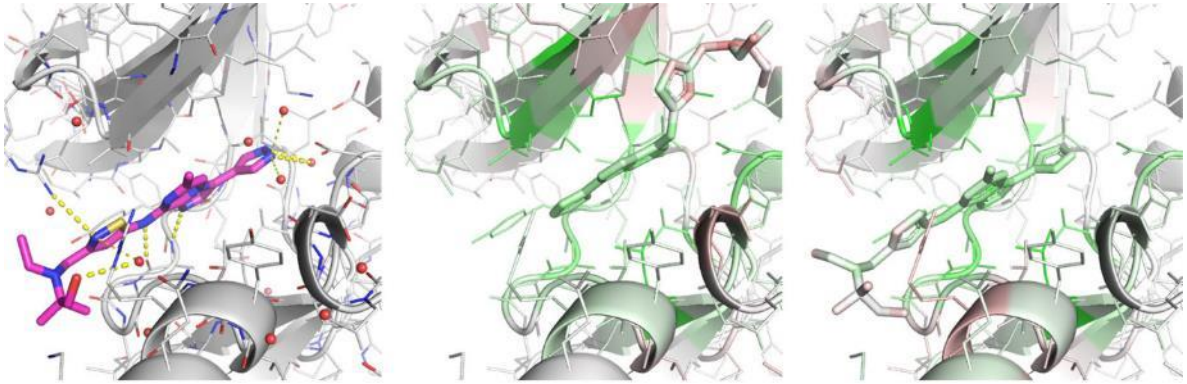
DL-based scoring function of binding

CNN wins at predicting "good" vs "bad" poses **across targets**

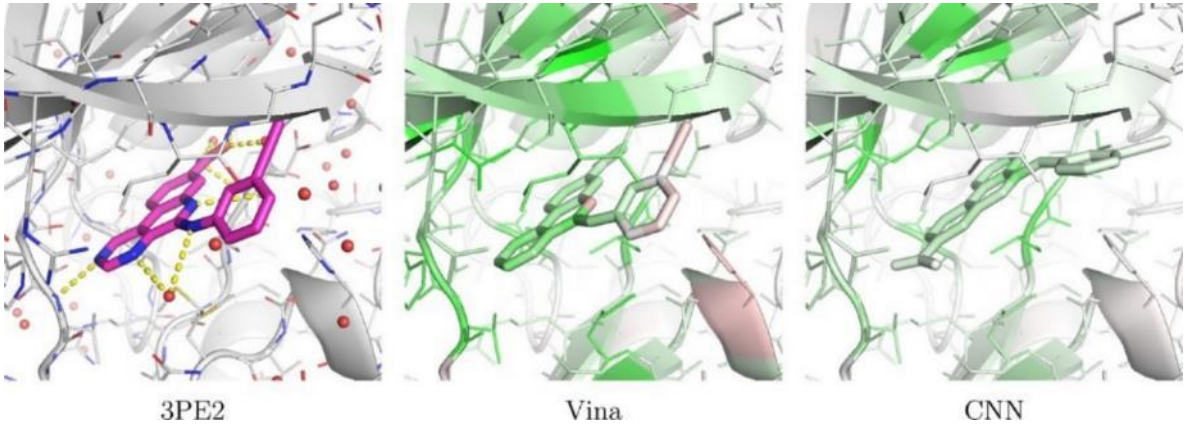


CNN loses to Vina at **same-target** predictions

Example: CNN wins



Example: CNN loses

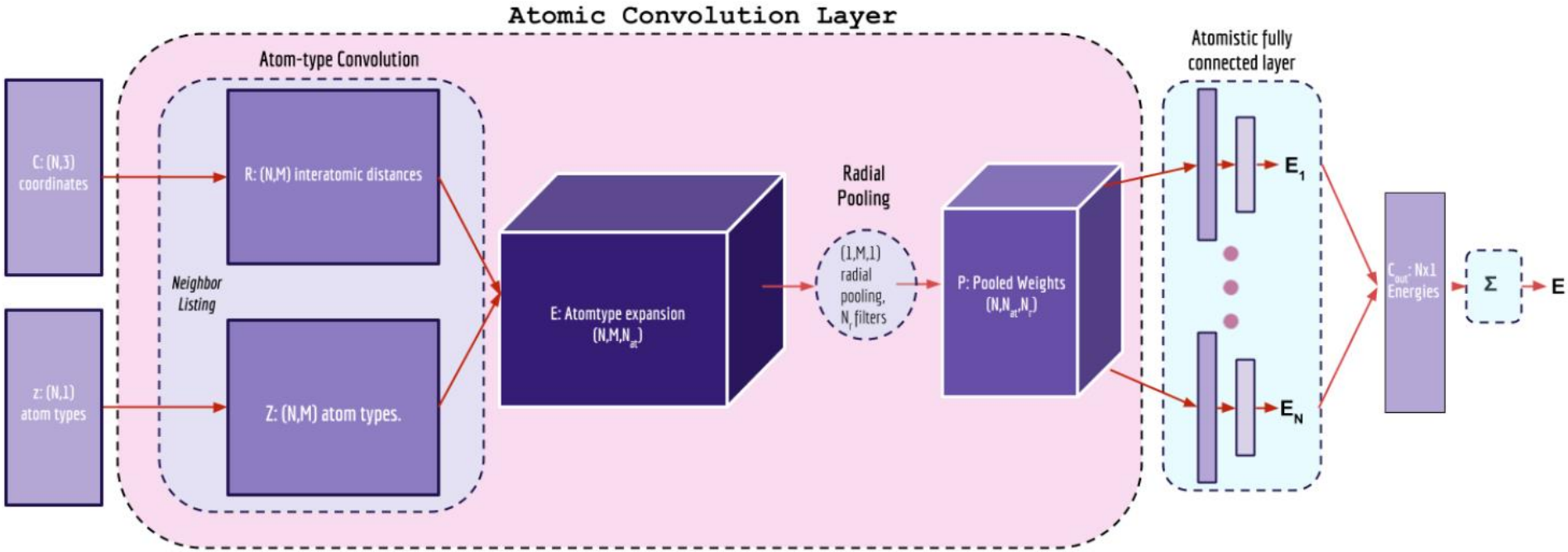


Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. and Koes, D. R. (2017) 'Protein–Ligand Scoring with Convolutional Neural Networks', *Journal of Chemical Information and Modeling*, 57(4), pp. 942–957. doi: 10.1021/acs.jcim.6b00740.

Drug-target interactions

Calculating binding free energy

- Characterize molecular neighborhood of each atom (distances to nearby atoms, atom types)
- Learn/predict energies of bound and unbound states → free energy (strength of drug binding)

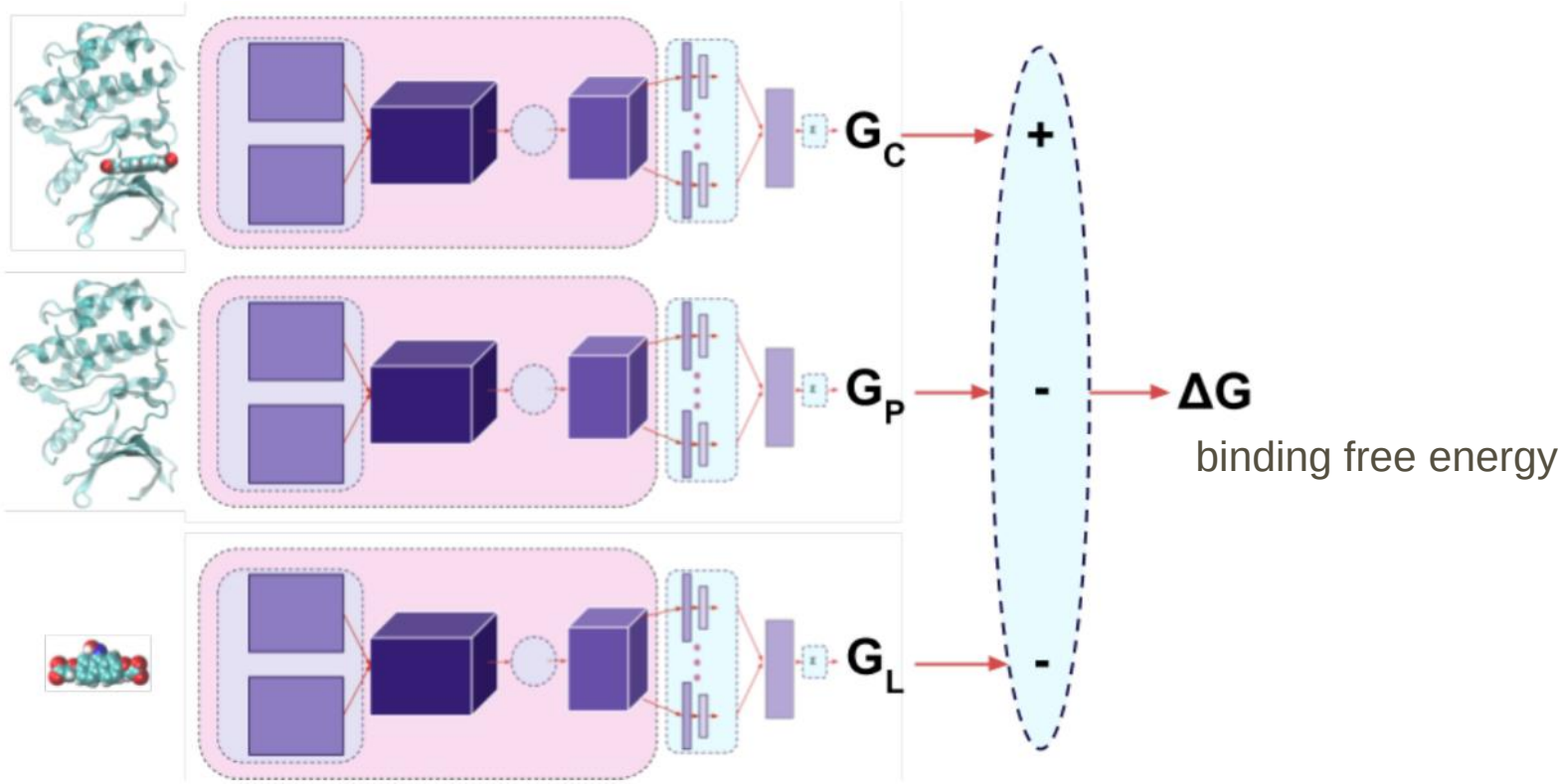


Gomes, J., Ramsundar, B., Feinberg, E. N. and Pande, V. S. (2017) 'Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity', pp. 1–17. Available at: http://arxiv.org/abs/1703.10603.

Drug-target interactions

Calculating binding free energy

- Characterize molecular neighborhood of each atom (distances to nearby atoms, atom types)
- Learn/predict energies of bound and unbound states → free energy (strength of drug binding)



Gomes, J., Ramsundar, B., Feinberg, E. N. and Pande, V. S. (2017) 'Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity', pp. 1–17. Available at: http://arxiv.org/abs/1703.10603.

Drug-target interactions

Calculating binding free energy

- Characterize molecular neighborhood of each atom (distances to nearby atoms, atom types)
- Learn/predict energies of bound and unbound states → free energy (strength of drug binding)

Compare against known free energies

	ACNN		GRID-RF		GRID-NN		GCNN		ECFP-RF		ECFP-NN	
Split	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Random	.912	**.448**	.969	.336	.963	.058	.676	.265	.920	.212	.942	.227
Stratified	.939	.116	.969	**.132**	.963	.165	.735	.064	.924	.071	.942	.077
Scaffold	.911	.043	.965	.109	.953	.067	.797	**.254**	.920	.218	.940	.206
Temporal	.923	.251	.972	**.287**	.957	.245	.744	.095	.925	.206	.952	.071

Table 1. Performance (Pearson R^2) on PDBBind core train/test sets.

	ACNN		GRID-RF		GRID-NN		GCNN		ECFP-RF		ECFP-NN	
Split	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Random	0.325	0.774	0.385	**0.741**	0.230	0.877	0.656	1.112	0.399	1.112	0.234	1.138
Stratified	0.282	0.997	0.339	0.990	0.205	**0.813**	0.556	0.995	0.410	0.901	0.223	1.115
Scaffold	0.410	0.993	0.338	1.397	0.211	1.630	0.516	**0.883**	0.438	1.003	0.221	0.909
Temporal	0.363	0.974	0.368	**0.860**	0.237	0.809	0.588	1.062	0.413	0.974	0.341	1.265

Table 2. Performance (MUE [kcal/mol]) on PDBBind core train/test sets.

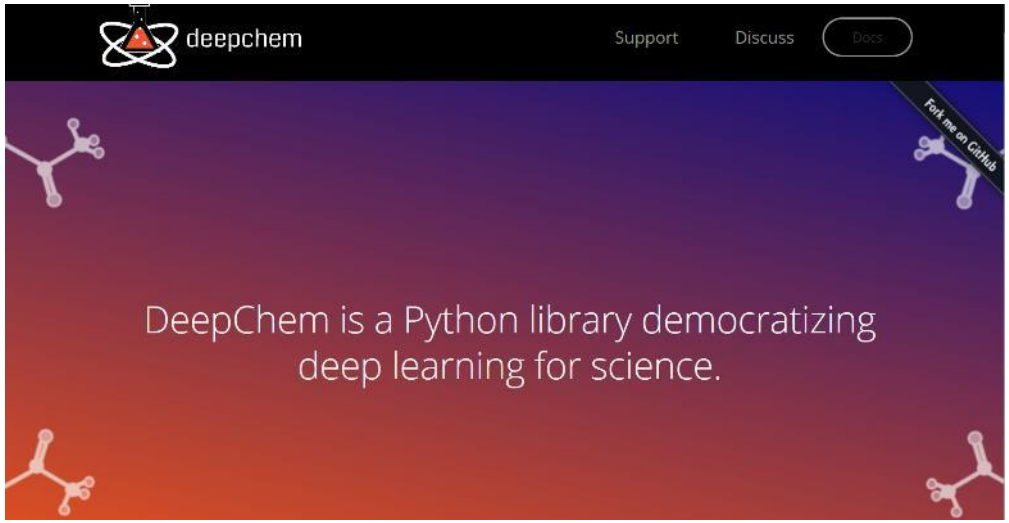
Uses DeepChem...

Gomes, J., Ramsundar, B., Feinberg, E. N. and Pande, V. S. (2017) 'Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity', pp. 1–17. Available at: http://arxiv.org/abs/1703.10603.

DeepChem

Deep Learning tools for drug discovery

- Vijay Pande lab (Stanford)
- Implementation of many useful deep learning techniques for small molecules / drug binding (e.g. Graph convolution, dataset stratification based on molecular scaffold, ...)



https://www.deepchem.io/

deepchem 1.3.1

Tutorials

The following tutorials show off various aspects or capabilities of DeepChem. They can be run interactively in Jupyter (IPython) notebook. Download the notebook files and open them in Jupyter:

```
$ jupyter notebook
```

DeepChem Tutorials

- Predicting Ki of Ligands to a Protein
- Conditional Generative Adversarial Network
- Creating a high fidelity DeepChem dataset from experimental data
- Graph Convolutions For Tox21
- MNIST with DeepChem and TensorGraph
- MNIST GAN
- Multitask Networks On MUV
- Pong in DeepChem with A3C
- Basic Protein-Ligand Affinity Models
- Tutorial: Use machine learning to model protein-ligand affinity.
- The protein-ligand complex view.
- Quantum Machinery with gdb1k
- SeqToSeq Fingerprint
- Modeling Solubility
- TensorGraph Mechanics

Contributing tutorials

Do you have a neat example of using DeepChem? Format your code into an IPython notebook and submit a pull request!

Source Back to top

Created using Sphinx 1.3.5.

Drug vs Biology

Complex effects of drugs inside an organism

Drug vs Biology

DeepTox: Toxicity Prediction using Deep Learning

– Predict tox from 2D/3D molecular features

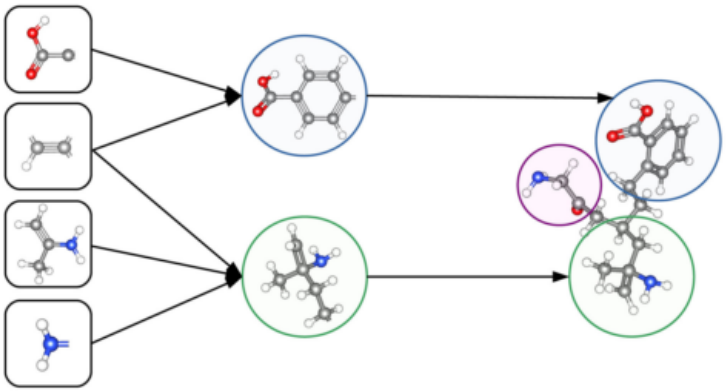


FIGURE 3 | Representation of a toxicophore by hierarchically related features. Simple features share chemical properties coded as reactive centers. Combining reactive centers leads to toxicophores that represent specific toxicological effects.

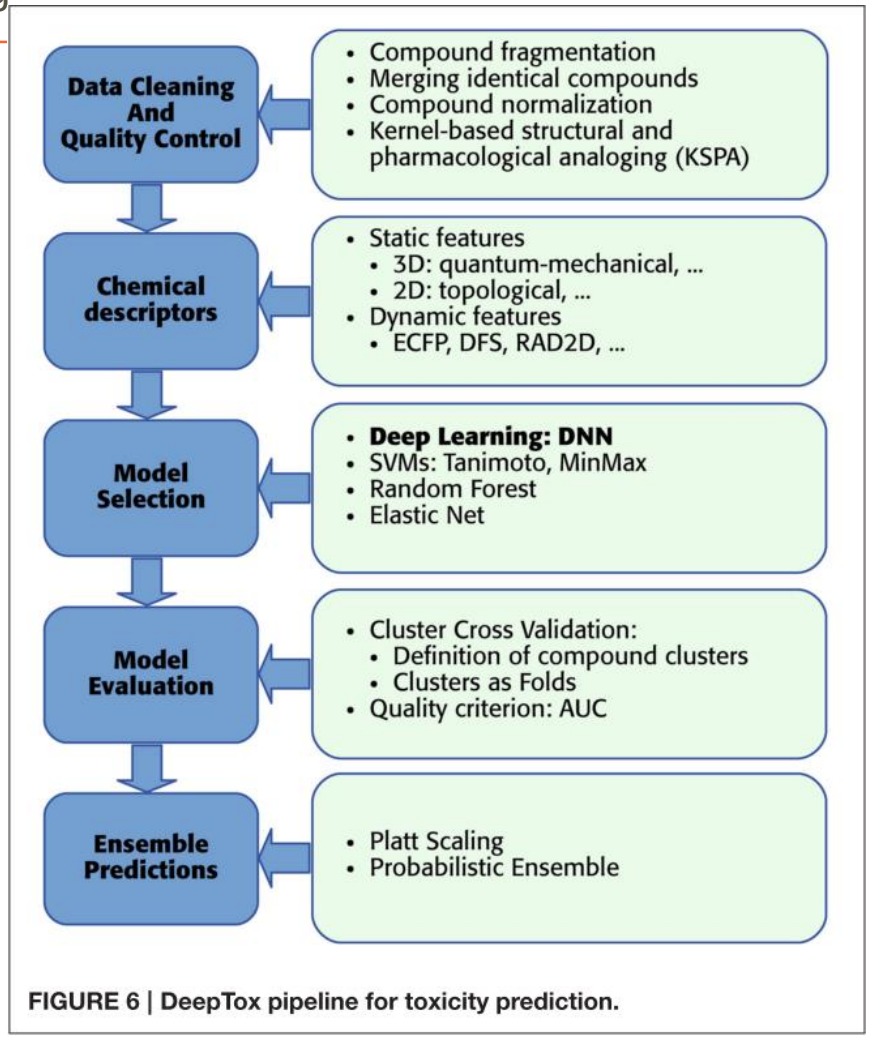


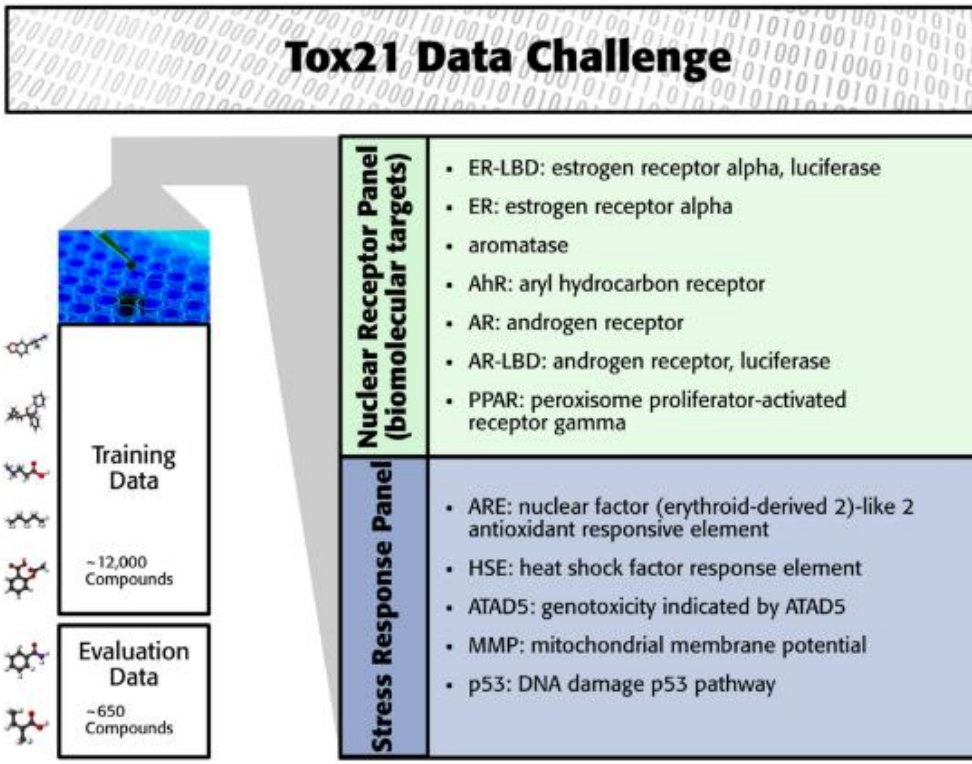
FIGURE 6 | DeepTox pipeline for toxicity prediction.

Mayr, A., Klambauer, G., Unterthiner, T. and Hochreiter, S. (2016) 'DeepTox: Toxicity Prediction using Deep Learning', *Frontiers in Environmental Science*, 3(February). doi: 10.3389/fenvs.2015.00080.

Drug vs Biology

DeepTox: Toxicity Prediction using Deep Learning

	AVG	NR	SR	AhR	AR	AR-LBD	ARE	Aromatase	ATAD5	ER	ER-LBD	HSE	MMP	p53	PPAR.g
our method	**0.846**	**0.826**	**0.858**	**0.928**	0.807	**0.879**	**0.840**	0.834	0.793	**0.810**	0.814	**0.865**	0.942	0.862	**0.861**
AMAZIZ	0.838	0.816	0.854	0.913	0.770	0.846	0.805	0.819	**0.828**	0.806	0.806	0.842	**0.950**	0.843	0.830
dmlab	0.824	0.811	0.850	0.781	**0.828**	0.819	0.768	**0.838**	0.800	0.766	0.772	0.855	0.946	**0.880**	0.831
T	0.823	0.798	0.842	0.913	0.676	0.848	0.801	0.825	0.814	0.784	0.805	0.811			
microsomes	0.810	0.785	0.814	0.901	–	–	0.804	–	0.812	0.785	**0.827**	-			
filipsPL	0.798	0.765	0.817	0.893	0.736	0.743	0.758	0.776	–	0.771	–	0.766			
Charite	0.785	0.750	0.811	0.896	0.688	0.789	0.739	0.781	0.751	0.707	0.798	0.852			
RCC	0.772	0.751	0.781	0.872	0.763	0.747	0.761	0.792	0.673	0.781	0.762	0.755			
frozenarm	0.771	0.759	0.768	0.865	0.744	0.722	0.700	0.740	0.726	0.745	0.790	0.752			
ToxFit	0.763	0.753	0.756	0.862	0.744	0.757	0.697	0.738	0.729	0.729	0.752	0.689			
CGL	0.759	0.720	0.791	0.866	0.742	0.566	0.747	0.749	0.737	0.759	0.727	0.775			
SuperTox	0.743	0.682	0.768	0.854	–	0.560	0.711	0.742	–	–	–	-			
kibutz	0.741	0.731	0.731	0.865	0.750	0.694	0.708	0.729	0.737	0.757	0.779	0.587			
MML	0.734	0.700	0.753	0.871	0.693	0.660	0.701	0.709	0.749	0.750	0.710	0.647			
NCI	0.717	0.651	0.791	0.812	0.628	0.592	0.783	0.698	0.714	0.483	0.703	0.858			
VIF	0.708	0.702	0.692	0.827	0.797	0.610	0.636	0.671	0.656	0.732	0.735	0.723			
Toxic Avg	0.644	0.659	0.607	0.715	0.721	0.611	0.633	0.671	0.593	0.646	0.640	0.465			
Swamidass	0.576	0.596	0.593	0.353	0.571	0.748	0.372	0.274	0.391	0.680	0.738	0.711			



Mayr, A., Klambauer, G., Unterthiner, T. and Hochreiter, S. (2016) 'DeepTox: Toxicity Prediction using Deep Learning', *Frontiers in Environmental Science*, 3(February). doi: 10.3389/fenvs.2015.00080.

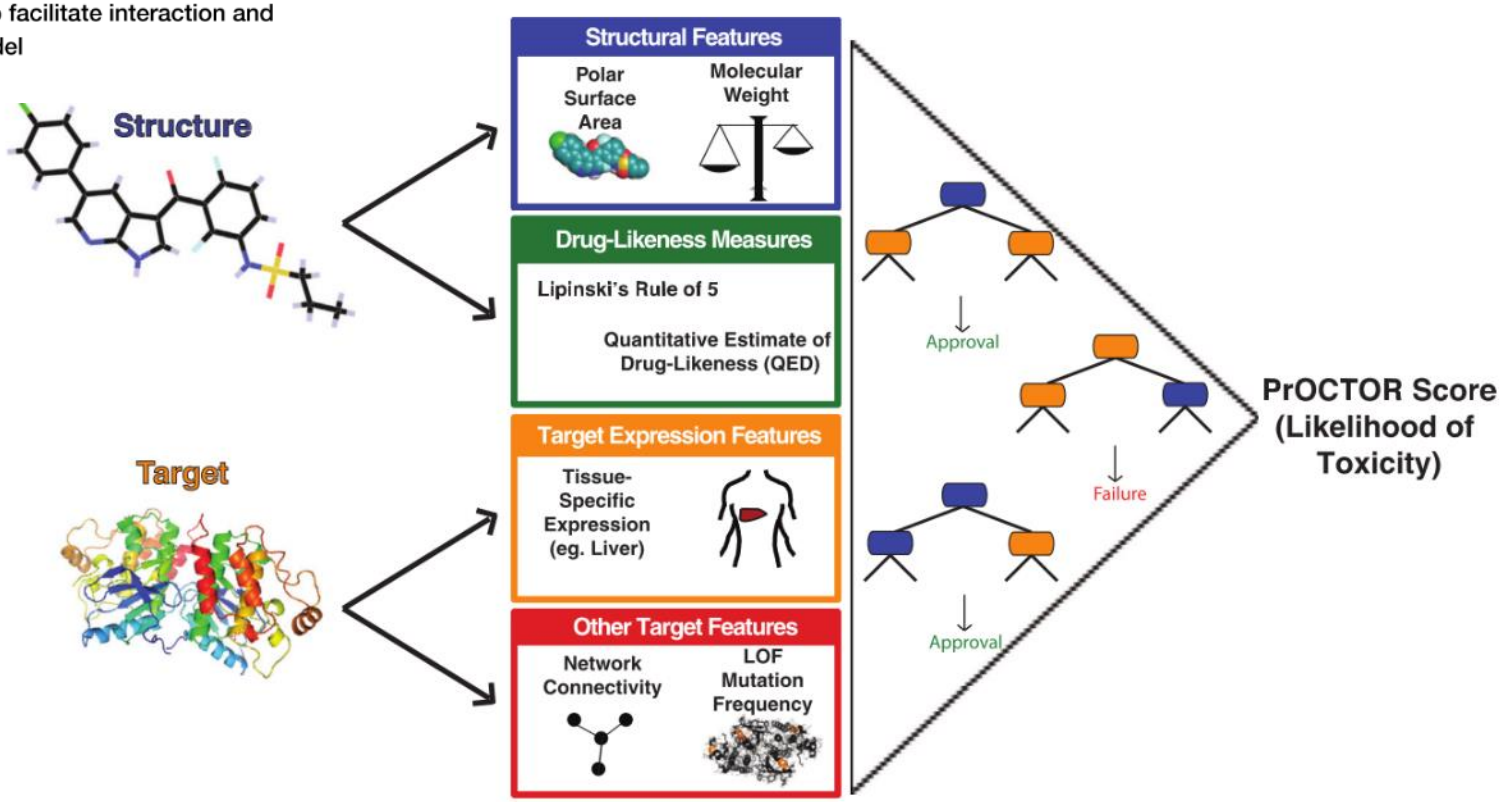
Drug vs Biology

Predicting tox based on molecular features, trained on clinical trial outcomes

Highlights

- Computational approach predicts the likelihood of clinical trial toxicity
- Identification of molecule and target properties associated with clinical toxicity
- Development of a tool to facilitate interaction and interpretation of the model

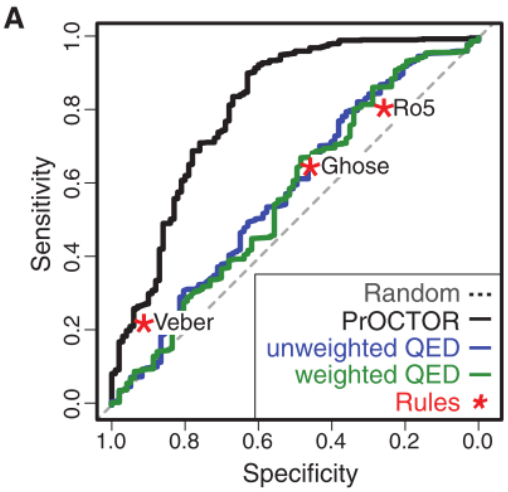
(Not a DNN! Random Forest)

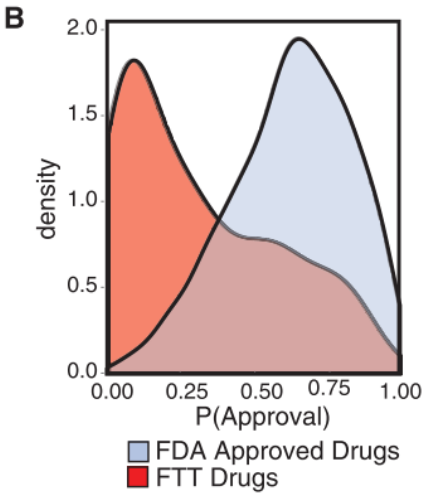


Gayvert, K. M., Madhukar, N. S. and Elemento, O. (2016) 'A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials', *Cell Chemical Biology*. Elsevier Ltd, pp. 1–8. doi: 10.1016/j.chembiol.2016.07.023.

Drug vs Biology

Predicting tox based on molecular features, trained on clinical trial outcomes

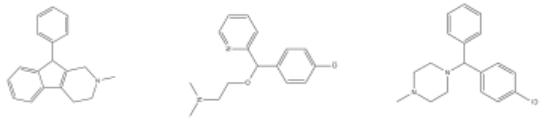




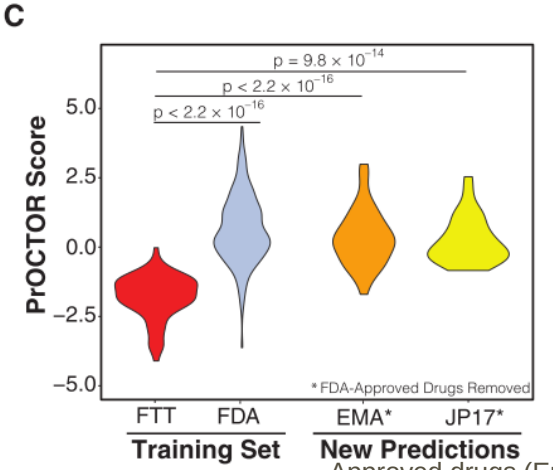
FTT = failed toxic clinical trial

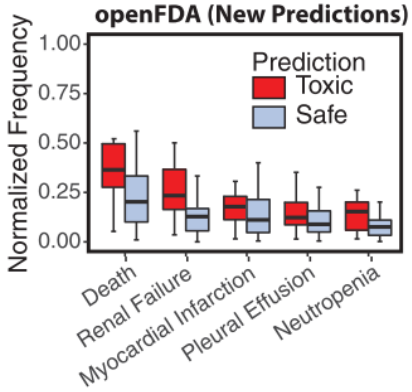
E

Top-3 predicted FDA approval



	Phenindamine	Carbinoxamine	Chlorcyclizine
PrOCTOR Score	3.0610	2.5025	2.6309

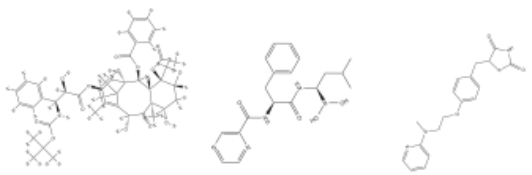




Approved drugs (Europe, Japan)

F

Top-3 predicted failure



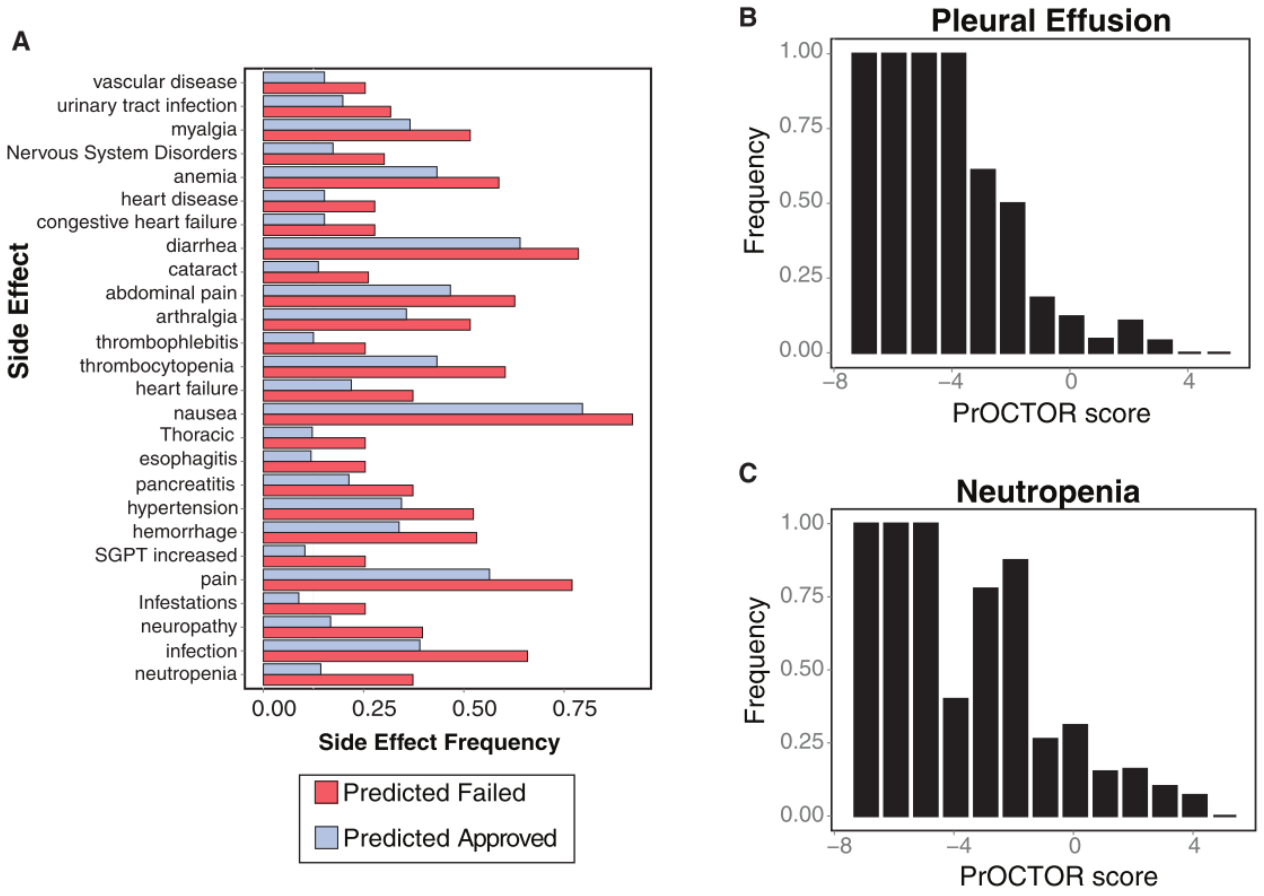
	Docetaxel	Bortezomib	Rosiglitazone
	-1.7760	-6.1882	-4.7182

Gayvert, K. M., Madhukar, N. S. and Elemento, O. (2016) 'A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials', *Cell Chemical Biology*. Elsevier Ltd, pp. 1–8. doi: 10.1016/j.chembiol.2016.07.023.

Drug vs Biology

Predicting tox based on molecular features, trained on clinical trial outcomes

– PrOCTOR score predicts frequency of side effects



Gayvert, K. M., Madhukar, N. S. and Elemento, O. (2016) 'A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials', *Cell Chemical Biology*. Elsevier Ltd, pp. 1–8. doi: 10.1016/j.chembiol.2016.07.023.

Review articles

Deep Learning in Bio/Chem/Med/Pharma

- Mamoshina, P., Vieira, A., Putin, E. and Zhavoronkov, A. (2016) 'Applications of Deep Learning in Biomedicine', *Mol Pharm*, 13(5), pp. 1445–1454. doi: 10.1021/acs.molpharmaceut.5b00982.
- Min, S., Lee, B. and Yoon, S. (2016) 'Deep Learning in Bioinformatics'. doi: 10.1093/bib/bbw068.
- Angermueller, C., Pärnamaa, T., Parts, L. and Stegle, O. (2016) 'Deep learning for computational biology', *Molecular Systems Biology*, 12(7), p. 878. doi: 10.15252/msb.20156651.
- Gawehn, E., Hiss, J. A. and Schneider, G. (2016) 'Deep Learning in Drug Discovery', *Molecular Informatics*, 35(1), pp. 3–14. doi: 10.1002/minf.201501008.
- Baskin, I. I., Winkler, D. and Tetko, I. V. (2016) 'A renaissance of neural networks in drug discovery.', *Expert opinion on drug discovery*. Taylor & Francis, 441(June), p. 17460441.2016.1201262. doi: 10.1080/17460441.2016.1201262.
- Goh, G. B., Hodas, N. O. and Vishnu, A. (2017) 'Deep learning for computational chemistry', *Journal of Computational Chemistry*, 38(16), pp. 1291–1307. doi: 10.1002/jcc.24764.
- Chen, Y., Li, Y., Narayan, R., Subramanian, A. and Xie, X. (2016) 'Gene expression inference with deep learning', *Bioinformatics*, 32(12), pp. 1832–1839. doi: 10.1093/bioinformatics/btw074.
- Hodos, R. A., Kidd, B. A., Shameer, K., Readhead, B. P. and Dudley, J. T. (2016) 'In silico methods for drug repurposing and pharmacology', *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(3), pp. 186–210. doi: 10.1002/wsbm.1337.
- Shen, D., Wu, G. and Suk, H.-I. (2017) 'Deep Learning in Medical Image Analysis', *Annual Review of Biomedical Engineering*. Annual Reviews , 19(1), pp. 221–248. doi: 10.1146/annurev-bioeng-071516-044442.

Thanks

heidelberg.ai

Dec. 12th, 6pm: [Paper Discussion] Dynamic Routing between Capsules