Accelerating drug discovery with deep neural networks

literature review

Tobias Sikosek
Senior Data Scientist
In Silico Unit
(Heidelberg)
Renewed focus on multi-layer (=deep) artificial neural networks with improved algorithms, more data, and more compute power (GPUs)

Breakthroughs in image and language recognition
Drug Discovery in a nutshell

Preclinical drug discovery

Disease ➔ Intra-cellular Pathways (genes relevant for disease) ➔ Drug Target (protein) ➔ Small molecules (compounds / drugs) To modulate target activity ➔ Clinical trial

Optimization cycle:
Test
Refine

➢ Increase on-target activity
➢ Reduce off-target activity / toxicity / side-effects
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

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 intrinsic **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.

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)

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.

---

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

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

**AP-1 reporter assays**

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)

PROTEIN STRUCTURE
Scaffold to support and position active site

ACTIVE SITE

BINDING SITES
Bind and orient substrate(s)

CATALYTIC SITE
Reduce chemical activation energy

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 $\rightarrow$ **fixed-length input**:
  - Break sequences into artificial chunks
    - Problem: often protein needs to be studied in its entirety
  - Choose input size $\leq$ 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“

![Original Sequence](image)

\[
\begin{align*}
(1) \quad & M \\
(2) \quad & A \\
(3) \quad & FSAEDVLKEYDRRRRMEAL.. \\
\end{align*}
\]

Splittings

\[
\begin{align*}
1) \quad & MAF, SAE, DVL, KEY, DRR, RRM, .. \\
2) \quad & AFS, AED, VLK, EYD, RRR, RME, .. \\
3) \quad & FSA, EDV, LKE, YDR, RRR, MEA, .. \\
\end{align*}
\]

**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.

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)

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


Constrained folding simulation \(\rightarrow\) 3D structure
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).

Overlap between top L/2 predicted contacts (in red or green) and the native contactmap (in grey) for CAMEO target 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).

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.

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

Protein structures
Amino acid substitutions: 3D atomic coordinates; Conv Net

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

**a** Deep Learning Framework - 3DCNN

**b** FEATURE Softmax Classifier

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 $j$.

Protein structures
Amino acid substitutions: 3D atomic coordinates; Conv Net

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Variant</th>
<th>Wild type PDB ID</th>
<th>Wild True</th>
<th>Wild Predicted</th>
<th>Mutant PDB ID</th>
<th>Mutant True</th>
<th>Mutant Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Destabilizing</td>
<td>M102K</td>
<td>2LZM</td>
<td>MET</td>
<td>MET</td>
<td>1L54</td>
<td>LYS</td>
<td>MET</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>L99G</td>
<td>2LZM</td>
<td>LEU</td>
<td>LEU</td>
<td>1QUD</td>
<td>GLY</td>
<td>LEU</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>V149S</td>
<td>2LZM</td>
<td>VAL</td>
<td>VAL</td>
<td>1G08</td>
<td>SER</td>
<td>ILE</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>V149C</td>
<td>2LZM</td>
<td>VAL</td>
<td>VAL</td>
<td>1G07</td>
<td>CYS</td>
<td>VAL</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>V149G</td>
<td>2LZM</td>
<td>VAL</td>
<td>VAL</td>
<td>1G0P</td>
<td>GLY</td>
<td>VAL</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>T157I</td>
<td>2LZM</td>
<td>THR</td>
<td>THR</td>
<td>1L10</td>
<td>ILE</td>
<td>THR</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>G158D</td>
<td>2LZM</td>
<td>GLY</td>
<td>GLY</td>
<td>1L16</td>
<td>ASP</td>
<td>GLY</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>R96H</td>
<td>2LZM</td>
<td>ARG</td>
<td>ARG</td>
<td>1L34</td>
<td>HIS</td>
<td>ARG</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>D92N</td>
<td>2LZM</td>
<td>ASP</td>
<td>ASP</td>
<td>1L55</td>
<td>ASN</td>
<td>ASP</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>I3P</td>
<td>2LZM</td>
<td>ILE</td>
<td>ILE</td>
<td>1L97</td>
<td>PRO</td>
<td>ILE</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>V87M</td>
<td>2LZM</td>
<td>VAL</td>
<td>VAL</td>
<td>1C03</td>
<td>MET</td>
<td>PHE</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>R96N</td>
<td>2LZM</td>
<td>ARG</td>
<td>ARG</td>
<td>3CDT</td>
<td>ASN</td>
<td>ARG</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>R96D</td>
<td>2LZM</td>
<td>ARG</td>
<td>ARG</td>
<td>3C8Q</td>
<td>ASP</td>
<td>ARG</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>R96W</td>
<td>2LZM</td>
<td>ARG</td>
<td>ARG</td>
<td>3F15</td>
<td>TRP</td>
<td>PHE</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>R96Y</td>
<td>2LZM</td>
<td>ARG</td>
<td>ARG</td>
<td>3C80</td>
<td>TYR</td>
<td>ARG</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>M102L</td>
<td>2LZM</td>
<td>MET</td>
<td>MET</td>
<td>1L77</td>
<td>LEU</td>
<td>ILE</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>M106K</td>
<td>2LZM</td>
<td>MET</td>
<td>MET</td>
<td>231L</td>
<td>LYS</td>
<td>MET</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>M120K</td>
<td>2LZM</td>
<td>GLU</td>
<td>GLU</td>
<td>232L</td>
<td>LYS</td>
<td>TYR</td>
</tr>
<tr>
<td>Destabilizing</td>
<td>I3Y</td>
<td>2LZM</td>
<td>ILE</td>
<td>ILE</td>
<td>1L18</td>
<td>TYR</td>
<td>TYR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mutant</th>
<th>G77A</th>
<th>2LZM</th>
<th>GLY</th>
<th>ALA</th>
<th>1L24</th>
<th>ALA</th>
<th>ARG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>ALA</td>
<td>PRO</td>
<td>1L24</td>
<td>PRO</td>
<td>PRO</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>ALA</td>
<td>ASP</td>
<td>125L</td>
<td>THR</td>
<td>ASP</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>THR</td>
<td>GLU</td>
<td>130L</td>
<td>SER</td>
<td>ASP</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>THR</td>
<td>GLU</td>
<td>131L</td>
<td>SER</td>
<td>ALA</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>VAL</td>
<td>VAL</td>
<td>1C06</td>
<td>MET</td>
<td>ILE</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>SER</td>
<td>SER</td>
<td>1L61</td>
<td>ASN</td>
<td>ASP</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>SER</td>
<td>SER</td>
<td>1L19</td>
<td>ASP</td>
<td>ASP</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>ASN</td>
<td>GLU</td>
<td>1L20</td>
<td>ASP</td>
<td>GLU</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>THR</td>
<td>GLU</td>
<td>1L59</td>
<td>ASN</td>
<td>ALA</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>THR</td>
<td>GLU</td>
<td>1L52</td>
<td>ASP</td>
<td>GLU</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>ASN</td>
<td>ASP</td>
<td>1L57</td>
<td>ASP</td>
<td>ASP</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>MET</td>
<td>LEU</td>
<td>1P46</td>
<td>ILE</td>
<td>VAL</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>MET</td>
<td>GLU</td>
<td>1P8Y</td>
<td>TYR</td>
<td>TYR</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>VAL</td>
<td>VAL</td>
<td>1G0Q</td>
<td>ILE</td>
<td>ILE</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>THR</td>
<td>VAL</td>
<td>100L</td>
<td>VAL</td>
<td>VAL</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>GLU</td>
<td>ILE</td>
<td>1QUG</td>
<td>VAL</td>
<td>VAL</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>SER</td>
<td>ALA</td>
<td>1TLA</td>
<td>PHE</td>
<td>PHE</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>MET</td>
<td>LEU</td>
<td>234L</td>
<td>LEU</td>
<td>LEU</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>ILE</td>
<td>ILE</td>
<td>1L17</td>
<td>VAL</td>
<td>ILE</td>
</tr>
<tr>
<td>Neutral</td>
<td>ASP</td>
<td>2LZM</td>
<td>MET</td>
<td>GLU</td>
<td>233L</td>
<td>LEU</td>
<td>LEU</td>
</tr>
</tbody>
</table>

Small molecules

Representing small drug-like molecules for machine learning
Small molecules

Conventional representations

- Molecular structure graph

  ![Molecular Structure Graph of Aspirin](https://www.ebi.ac.uk/chembl/compound/inspect/CHEMBL25)

- 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

![SMILES](image1) → 2D Drawing → “Grid” Image → Deep Neural Network → Prediction

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

<table>
<thead>
<tr>
<th></th>
<th>Train AUC</th>
<th>Validation AUC</th>
<th>Test AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>nr-ahr</td>
<td>0.825</td>
<td>+/-</td>
<td>0.018</td>
</tr>
<tr>
<td>nr-ar</td>
<td>0.843</td>
<td>+/-</td>
<td>0.010</td>
</tr>
<tr>
<td>nr-ar-lbd</td>
<td>0.887</td>
<td>+/-</td>
<td>0.034</td>
</tr>
<tr>
<td>nr-aromatase</td>
<td>0.801</td>
<td>+/-</td>
<td>0.010</td>
</tr>
<tr>
<td>nr-er</td>
<td>0.747</td>
<td>+/-</td>
<td>0.020</td>
</tr>
<tr>
<td>nr-er-lbd</td>
<td>0.824</td>
<td>+/-</td>
<td>0.029</td>
</tr>
<tr>
<td>nr-ppar-gamma</td>
<td>0.791</td>
<td>+/-</td>
<td>0.038</td>
</tr>
<tr>
<td>sr-are</td>
<td>0.724</td>
<td>+/-</td>
<td>0.009</td>
</tr>
<tr>
<td>sr-atad55</td>
<td>0.841</td>
<td>+/-</td>
<td>0.022</td>
</tr>
<tr>
<td>sr-hse</td>
<td>0.776</td>
<td>+/-</td>
<td>0.032</td>
</tr>
<tr>
<td>sr-mmp</td>
<td>0.791</td>
<td>+/-</td>
<td>0.020</td>
</tr>
<tr>
<td>sr-p53</td>
<td>0.844</td>
<td>+/-</td>
<td>0.034</td>
</tr>
<tr>
<td><strong>Tox21</strong></td>
<td><strong>0.808</strong></td>
<td><strong>0.044</strong></td>
<td><strong>0.760</strong></td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Step $t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_t$</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>$x_t$</td>
<td>(1,0,0)</td>
<td>(0,1,0)</td>
<td>(1,0,0)</td>
<td>...</td>
<td>(1,0,0)</td>
<td>(0,1,0)</td>
</tr>
<tr>
<td>$h_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_6$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$h_9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_t$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_{t+1}$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example input molecules with SMILES strings

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

### Valid SMILES emerge over training time

<table>
<thead>
<tr>
<th>Batch</th>
<th>Generated Example</th>
<th>valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0c.BK5i?ur+7oAFc7L3=3F8B5e=n)CS6RCTAR((0VCp1CApb)</td>
<td>no</td>
</tr>
<tr>
<td>1000</td>
<td>0F=CCC20CCC(C2)C1CNC2CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC</td>
<td>no</td>
</tr>
<tr>
<td>2000</td>
<td>0=O(N)C(=O)N(c10ccc1DC)c2cccccc2DC</td>
<td>yes</td>
</tr>
<tr>
<td>3000</td>
<td>0=C1C=2N(c3cc(c30c2CC1)CCCc4cn(c5c(Cl)cccc54)cC</td>
<td>yes</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>#</th>
<th>$pIC_{50}$</th>
<th>Train.</th>
<th>Test</th>
<th>Gen. mols.</th>
<th>Reprod.</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&gt; 8</td>
<td>1239</td>
<td>1240</td>
<td>128,256</td>
<td>28%</td>
<td>66.9</td>
</tr>
<tr>
<td>2</td>
<td>&gt; 8</td>
<td>100</td>
<td>1240</td>
<td>93,721</td>
<td>7%</td>
<td>19.0</td>
</tr>
<tr>
<td>3</td>
<td>&gt; 9</td>
<td>100</td>
<td>1022</td>
<td>91,034</td>
<td>11%</td>
<td>35.7</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Entry</th>
<th>pMIC</th>
<th>Train.</th>
<th>Test</th>
<th>Gen. mols.</th>
<th>Reprod.</th>
<th>EOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&gt; 3</td>
<td>1000</td>
<td>6051</td>
<td>51,052</td>
<td>14%</td>
<td>155.9</td>
</tr>
<tr>
<td>2</td>
<td>&gt; 3</td>
<td>50</td>
<td>7001</td>
<td>70,891</td>
<td>2.5%</td>
<td>21.6</td>
</tr>
<tr>
<td>$3^a$</td>
<td>&gt; 3</td>
<td>50</td>
<td>7001</td>
<td>85,755</td>
<td>1.8%</td>
<td>6.3</td>
</tr>
<tr>
<td>$4^b$</td>
<td>&gt; 3</td>
<td>50</td>
<td>7001</td>
<td>285</td>
<td>0%</td>
<td>—</td>
</tr>
<tr>
<td>$5^c$</td>
<td>&gt; 3</td>
<td>0</td>
<td>7001</td>
<td>60,988</td>
<td>6%</td>
<td>59.6</td>
</tr>
</tbody>
</table>

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

---

Small molecules
Generating novel compounds: Adversarial autoencoders

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

Small molecules
Generating novel compounds: Adversarial autoencoders

(a) Variational autoencoder (VAE)  (b) Adversarial autoencoder (AAE)

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)

Drug-target interactions
DL-based scoring function of binding

Input: 3D grid 24x24x24
34 atom type channels
Density distribution around atom center

<table>
<thead>
<tr>
<th>Type</th>
<th>Ligand</th>
<th>Receptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>AliphaticCarbonXSHydrophobe</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>AliphaticCarbonXSNonHydrophobe</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>AromaticCarbonXSHydrophobe</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>AromaticCarbonXSNonHydrophobe</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bromine</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Calcium</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Chlorine</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Fluorine</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Iodine</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Iron</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Magnesium</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>NitrogenXSAcceptor</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>NitrogenXSDonor</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>NitrogenXSDonorAcceptor</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Oxygen</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>OxygenXSAcceptor</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>OxygenXSDonorAcceptor</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Sulfur</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>SulfurAcceptor</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Zinc</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

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)

Drug-target interactions
DL-based scoring function of binding

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

Example: CNN wins

Example: CNN loses

CNN loses to Vina at same-target predictions

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 \(\rightarrow\) free energy (strength of drug binding)

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)

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)

<table>
<thead>
<tr>
<th>Split</th>
<th>ACNN</th>
<th>GRID-RF</th>
<th>GRID-NN</th>
<th>GCNN</th>
<th>ECFP-RF</th>
<th>ECFP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>.912</td>
<td>.448</td>
<td>.969</td>
<td>.336</td>
<td>.963</td>
<td>.058</td>
</tr>
<tr>
<td>Stratified</td>
<td>.939</td>
<td>.116</td>
<td>.969</td>
<td>.132</td>
<td>.963</td>
<td>.165</td>
</tr>
<tr>
<td>Scaffold</td>
<td>.911</td>
<td>.043</td>
<td>.965</td>
<td>.109</td>
<td>.953</td>
<td>.067</td>
</tr>
<tr>
<td>Temporal</td>
<td>.923</td>
<td>.251</td>
<td>.972</td>
<td>.287</td>
<td>.957</td>
<td>.245</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Split</th>
<th>ACNN</th>
<th>GRID-RF</th>
<th>GRID-NN</th>
<th>GCNN</th>
<th>ECFP-RF</th>
<th>ECFP-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.325</td>
<td>0.774</td>
<td>0.385</td>
<td>0.741</td>
<td>0.230</td>
<td>0.877</td>
</tr>
<tr>
<td>Stratified</td>
<td>0.282</td>
<td>0.997</td>
<td>0.339</td>
<td>0.990</td>
<td>0.205</td>
<td>0.813</td>
</tr>
<tr>
<td>Scaffold</td>
<td>0.410</td>
<td>0.993</td>
<td>0.338</td>
<td>1.397</td>
<td>0.211</td>
<td>1.630</td>
</tr>
<tr>
<td>Temporal</td>
<td>0.363</td>
<td>0.974</td>
<td>0.368</td>
<td>0.860</td>
<td>0.237</td>
<td>0.809</td>
</tr>
</tbody>
</table>

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


Uses DeepChem...
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/
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.

Drug vs Biology

DeepTox: Toxicity Prediction using Deep Learning

<table>
<thead>
<tr>
<th>our method</th>
<th>AVG</th>
<th>NR</th>
<th>SR</th>
<th>AHr</th>
<th>AR</th>
<th>AHR-LBD</th>
<th>ARE</th>
<th>Aromatase</th>
<th>ATAD5</th>
<th>ER</th>
<th>ER-LBD</th>
<th>HSE</th>
<th>MMP</th>
<th>p53</th>
<th>PPAR-γ</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMAZIZ</td>
<td>0.838</td>
<td>0.816</td>
<td>0.854</td>
<td>0.913</td>
<td>0.770</td>
<td>0.846</td>
<td>0.805</td>
<td>0.819</td>
<td>0.828</td>
<td>0.810</td>
<td>0.814</td>
<td>0.865</td>
<td>0.942</td>
<td>0.862</td>
<td>0.861</td>
</tr>
<tr>
<td>dmhab</td>
<td>0.824</td>
<td>0.811</td>
<td>0.850</td>
<td>0.781</td>
<td>0.828</td>
<td>0.819</td>
<td>0.768</td>
<td>0.833</td>
<td>0.800</td>
<td>0.766</td>
<td>0.772</td>
<td>0.856</td>
<td>0.945</td>
<td>0.860</td>
<td>0.831</td>
</tr>
<tr>
<td>T</td>
<td>0.823</td>
<td>0.798</td>
<td>0.842</td>
<td>0.913</td>
<td>0.676</td>
<td>0.848</td>
<td>0.801</td>
<td>0.825</td>
<td>0.814</td>
<td>0.784</td>
<td>0.805</td>
<td>0.811</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>microsomes</td>
<td>0.810</td>
<td>0.785</td>
<td>0.814</td>
<td>0.901</td>
<td>-</td>
<td>-</td>
<td>0.804</td>
<td>-</td>
<td>0.812</td>
<td>0.785</td>
<td>0.627</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>flipesPL</td>
<td>0.798</td>
<td>0.765</td>
<td>0.817</td>
<td>0.893</td>
<td>0.736</td>
<td>0.743</td>
<td>0.758</td>
<td>0.776</td>
<td>-</td>
<td>0.771</td>
<td>-</td>
<td>-</td>
<td>0.769</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Charte</td>
<td>0.785</td>
<td>0.750</td>
<td>0.811</td>
<td>0.896</td>
<td>0.688</td>
<td>0.789</td>
<td>0.739</td>
<td>0.781</td>
<td>0.751</td>
<td>0.707</td>
<td>0.798</td>
<td>0.856</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RCC</td>
<td>0.772</td>
<td>0.751</td>
<td>0.781</td>
<td>0.872</td>
<td>0.763</td>
<td>0.747</td>
<td>0.761</td>
<td>0.792</td>
<td>0.673</td>
<td>0.781</td>
<td>0.762</td>
<td>0.755</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>frozenarm</td>
<td>0.771</td>
<td>0.759</td>
<td>0.768</td>
<td>0.865</td>
<td>0.744</td>
<td>0.722</td>
<td>0.700</td>
<td>0.740</td>
<td>0.728</td>
<td>0.745</td>
<td>0.790</td>
<td>0.755</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ToxFit</td>
<td>0.763</td>
<td>0.753</td>
<td>0.756</td>
<td>0.862</td>
<td>0.744</td>
<td>0.757</td>
<td>0.697</td>
<td>0.738</td>
<td>0.729</td>
<td>0.729</td>
<td>0.752</td>
<td>0.688</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CQL</td>
<td>0.759</td>
<td>0.720</td>
<td>0.791</td>
<td>0.866</td>
<td>0.742</td>
<td>0.566</td>
<td>0.747</td>
<td>0.749</td>
<td>0.737</td>
<td>0.759</td>
<td>0.727</td>
<td>0.771</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SuperTox</td>
<td>0.743</td>
<td>0.682</td>
<td>0.768</td>
<td>0.854</td>
<td>-</td>
<td>0.560</td>
<td>0.711</td>
<td>0.742</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>kibutz</td>
<td>0.741</td>
<td>0.731</td>
<td>0.731</td>
<td>0.865</td>
<td>0.750</td>
<td>0.694</td>
<td>0.708</td>
<td>0.729</td>
<td>0.737</td>
<td>0.757</td>
<td>0.779</td>
<td>0.586</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MML</td>
<td>0.734</td>
<td>0.700</td>
<td>0.753</td>
<td>0.871</td>
<td>0.693</td>
<td>0.660</td>
<td>0.701</td>
<td>0.709</td>
<td>0.749</td>
<td>0.750</td>
<td>0.710</td>
<td>0.644</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>NCI</td>
<td>0.717</td>
<td>0.651</td>
<td>0.791</td>
<td>0.812</td>
<td>0.628</td>
<td>0.592</td>
<td>0.783</td>
<td>0.698</td>
<td>0.714</td>
<td>0.483</td>
<td>0.703</td>
<td>0.656</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VIF</td>
<td>0.708</td>
<td>0.702</td>
<td>0.692</td>
<td>0.827</td>
<td>0.797</td>
<td>0.610</td>
<td>0.636</td>
<td>0.671</td>
<td>0.666</td>
<td>0.732</td>
<td>0.735</td>
<td>0.724</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Toxic Avg</td>
<td>0.644</td>
<td>0.659</td>
<td>0.697</td>
<td>0.715</td>
<td>0.721</td>
<td>0.611</td>
<td>0.633</td>
<td>0.671</td>
<td>0.593</td>
<td>0.646</td>
<td>0.640</td>
<td>0.466</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Swamidassin</td>
<td>0.576</td>
<td>0.596</td>
<td>0.593</td>
<td>0.353</td>
<td>0.571</td>
<td>0.748</td>
<td>0.372</td>
<td>0.274</td>
<td>0.391</td>
<td>0.680</td>
<td>0.738</td>
<td>0.713</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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)

Drug vs Biology

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

A

B

E

Top-3 predicted FDA approval

Phenindamine  Carboxamine  Chlorcyclizine

3.0610  2.5025  2.6309

C

D

FTT = failed toxic clinical trial

openFDA (New Predictions)

Prediction

Toxic

Safe

F

Top-3 predicted failure

Docetaxel  Bortezomib  Rosiglitazone

-1.7760  -6.1882  -4.7182

Drug vs Biology

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

– PrOCTOR score predicts frequency of side effects

Review articles

Deep Learning in Bio/Chem/Med/Pharma

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