Artificial Intelligence for Medicine and Pharma

Sašo Džeroski
Jozef Stefan Institute (JSI), Ljubljana, Slovenia
Jozef Stefan International Postgraduate School (JS IPS)
Medicine is the science and practice of establishing
• diagnosis,
• prognosis,
• treatment, and
• prevention
of disease.

Encompasses a variety of health care practices to
maintain and restore health by prevention/treatment.
Health care

Health care is the maintenance or improvement of health via

• prevention,
• diagnosis,
• treatment,
• recovery, or
• cure

of disease, illness, injury, and other physical and mental impairments in people.
Drug discovery

• Drug discovery is the process by which new candidate medications are discovered.

• Compound screening: Applying compounds from a library (to a cellular assay) to determine their effect

• Genomic screening: Turning on/off of genes, to determine the effect, potential targets for compounds
Drug repositioning/repurposing

• Drug repositioning (aka drug repurposing) involves the investigation of existing drugs for new therapeutic purposes.

• The most famous/successful example of drug repurposing: sildenafil, originally used to treat pulmonary arterial hypertension
A taxonomy of AI approaches

- Knowledge representation
- Knowledge engineering
- Reasoning & Planning
- Natural language processing
- Computer vision
- Machine learning (from big data)
  - Neural networks
    - Deep neural networks (DNNs)
  - Learning understandable/ explainable models
    - Trees & tree ensembles
    - Rules & rule ensembles
DNNs for Image-based Diagnosis: Classification of skin cancer

- DNN pretrained on ImageNet
- Fine tuned on 129450 images of skin lesions
- 757 training classes defined according to a novel taxonomy of skin disease
• A subset of the top of the tree-structured taxonomy of skin disease.
• A set of testing images (photos & dermoscopy images)
Performance:
Comparison to Dermatologists

a
Carcinoma: 135 images

Melanoma: 130 images

Melanoma: 111 dermoscopy images

b
Carcinoma: 707 images

Melanoma: 225 images

Melanoma: 1,010 dermoscopy images

sensitivity = \frac{\text{true positive}}{\text{positive}}

specificity = \frac{\text{true negative}}{\text{negative}}
Diabetes causes blindness

• Fastest growing cause of blindness as

• A significant proportion of the population (5-10%) is diabetic

• Should be checked/ screened annually for diabetic retinopathy

• There is shortage of personnel to check/ grade images

• Grading is highly technical
Diagnosing Diabetic Retinopathy via Retinal Fundus Images

How DR is Diagnosed: Retinal Fundus Images

Healthy

Diseased

No DR  Mild DR  Moderate DR  Severe DR  Proliferative DR

Hemorrhages
Additional problems with grading
Training a DNN for diagnosing DR

Adapt deep neural network to read fundus images

Conv Network - 26 layers

Labeling tool
54 ophthalmologists
130k images

No DR
Mild DR
Moderate DR
Severe DR
Proliferative DR
Diagnostic performance as compared to ophthalmologists

F-score

0.95  Algorithm
0.91  Ophthalmologist (median)

“The study by Gulshan and colleagues truly represents the brave new world in medicine.”

Dr. Andrew Beam, Dr. Isaac Kohane
Harvard Medical School

“Google just published this paper in JAMA (impact factor 37) [...] It actually lives up to the hype.”

Dr. Luke Oakden-Rayner
University of Adelaide
Understandable ML for therapy

• Indicating change of therapy for Parkinson’s patients
• From patient’s symptoms at visit, predict whether physician will change each of three groups of meds: Levodopa, dopamine agonists, MAO-B inhibitors
Drug discovery/repurposing

• Perform compound screening with a relatively small compound library to collect data
• From the collected data, learn a predictive QSAR model that relates compound structure to activity

• Apply the learned model to perform virtual compound screening on a large set of compounds
• Find candidate compounds for new drugs
Real compound screening: Collecting data

• Testing compounds from libraries on cellular assays
Learning QSAR models for Virtual Compound Screening

### Labeled data

<table>
<thead>
<tr>
<th>Example</th>
<th>Descriptive space</th>
<th>Target space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>1 TRUE</td>
<td>0.49 0.69</td>
</tr>
<tr>
<td>Example 2</td>
<td>2 FALSE</td>
<td>0.08 0.07</td>
</tr>
<tr>
<td>Example 3</td>
<td>1 FALSE</td>
<td>0.08 0.07</td>
</tr>
<tr>
<td>Example 4</td>
<td>2 TRUE</td>
<td>0.49 0.69</td>
</tr>
</tbody>
</table>

### Unlabeled data

| Example N+1 | 1 TRUE | 0.86 0.35 |
| Example N+2 | 2 FALSE | 0.09 0.05 |
| Example N+3 | 4 FALSE | 0.07 0.01 |
| Example N+4 | 2 TRUE  | 0.91 0.78 |
| Example N+5 | 2 TRUE  | 0.42 0.69 |

Kills cancer cells? MCF7 HeLa

Learn Supervised model

Predict

Drug candidates
Virtual compound screening

• Descriptive variables refer to compound structure
  • Functional groups
  • Fingerprints
  • Bulk properties

• May also describe the compound in terms of the proteins it targets (e.g. from PubChem)
  • Their functional annotations
  • Pathways they are involved in
  • Proteins that the targets interact with (and/or their functional annotations, pathways they are involved in)

• Target variables describe compound activity and toxicity
Host-targeted Drugs for MTB (Tuberculosis) and STM (Salmonella)

- Library of compounds
  - LOPAC library - Library Of Pharmacologically Active Compounds
    - 1260 compounds
  - Well-characterized compounds, many already applied in clinical practice for a range of conditions

- Flow cytometry (FACS) - measured reduction in bacterial load
  - MelJuSo cells infected with Mycobacterium tuberculosis at MOI 10 – Mtb
  - HeLa cells infected with Salmonella typhimurium at MOI 10 - Stm
MTB&STM: Host-targeted Drugs

• Given SDF files, find PubChemID

• PubChem repository
  • Retrieve the proteins that were found to be active in bio-assays with human cells

• Dataset
  • 964 compounds were found active on human protein targets
  • 711 distinct protein targets were identified

• Each compound is described with
  • the respective protein targets
  • functional annotations of the respective protein targets
  • functional annotations of both the respective protein targets and the proteins they interact with
Model excerpts

• MTR models (trees, ensembles) were built that predict the effect of a compound on
  • Bacterial load
  • Host cell

• Example rule from a tree:
  IF compound targets the protein AAL 06595
  THEN bacterial load effect = -5.269 &
    host cell effect = 0.0475

• Functional profiles of targeted proteins
  IF a protein with function GO0002637 (regulation of immunoglobulin production) is targeted THEN ...
MTB&STM: Host-targeted Drugs

The Data Analysis Workflow

1. **Compound information** → **PubChem compounds** (Link to PubChem IDs)
2. **PubChem compounds** → **BioActivity profiles** (Retrieve BioAssay info)
3. **BioActivity profiles** → **Processed compounds** (Filter human targets)

4. **Training compounds** → **Predictive model** (Data mining algorithm)
5. **Predictive model** → **Predicted activity** (Apply pred. model)
6. **Predicted activity** → **Candidate compounds** (Reliability estimation)
7. **Candidate compounds** → **Select commercially available compounds**

Select commercially available compounds
MTB&STM: Host-targeted Drugs

Results

- Greatly increased proportions of hit compounds
  - 5 out of 9 (55.6%) for Mtb and
  - LOPAC primary screen (90 out of 1260 (7.1%) for Mtb

- The *in silico* predictive model successfully identified active compounds *de novo*
Analyzing data from High-contents Screens

- Compounds described by fingerprints
- Generated by open-source chemoinformatics SW library RDkit
- The FCFP2 fingerprints were used (1024 features)
- Also considered profiles of targeted proteins
- These are the attributes

- Assays photographed under the microscope
- Features extracted from images
- These are then the targets
HTS: Modulating fibroblast to myofibroblast transition

cardiac fibroblasts from α-SMA-RFP/
Coll α1(I)-EGFP mice

FDA-approved drugs (640 compounds)

cell fixation, image acquisition and elaboration

SMA intensity

Collagen intensity
Hits in the HTS screen

Fumagillone (fu)
Flubendazole (flu)
Mebendazol (me)
Rapamycin (rap)
Dextromethorphan (de)

Clodronate disodium (clo)
Apomorphine (apo)
Retinoic acid (ra)
Haloperidol (halo)
Dexamethasone (dex)

TGFβ

αSMA Hoechst

αSMA intensity (fold change)

- - + + + + + + TGFβ
- - + + + + + + Dexamethasone
- - + + + + + + Haloperidol
- - + + + + + + Apomorphine
- - + + + + + + Retinoic acid
- - + + + + + + Clodronate disodium
Reducing fibrosis in myocardial infarction

• High content screen using a library of 640 FDA approved drugs (ENZO)
• Identify drugs to reduce fibrosis in myocardial infarction
• Screen used murine cardiac fibroblasts which differentiate into myofibroblasts in culture, expressing increased alpha SMA-RFP and collagen-alpha1-EGFP

• Targets: Intensity of
  • alphaSMA
  • Collagen

• Attributes
  • Fingerprints
New candidate drugs to help recovery after heart attack

- SMILE strings used in Chemmine to identify substances with structural similarity to non commercial compounds with high predicted values

- Three related compounds identified which are described in literature to have an anti-fibrotic effect
  - Melatonin * and Indomethacin *
  - Acyclovir

- Four related compounds identified which were not previously described to have an anti-fibrotic effect
  - Dopamine
  - Amiodarone * and Progesterone *
  - Zanamivir
AI in Medicine and Pharma

• Many different tasks to use AI for, from pharma, medicine, healthcare
• Many different AI methods to use, e.g., also decision support systems to avoid hospital infections

• Important issues unique to uses in medicine/healthcare
  • Explainability
  • Regulating the use of AI in medicine (FDA approvals)