Using AI to accelerate drug development

Generate insights for disease understanding, target identification, and lead optimization

Subhasree Nag, PhD
Regional Solutions Consultant
04.05.2021
Introduction and background

2. How is Artificial Intelligence (AI) being used in drug discovery?

3. How can Clarivate Solutions help?

4. What are some of the barriers to success and what can we do to overcome them?

5. Key takeaways
Introduction
Costs of developing a drug

Number of years it takes to bring a drug to market: 10-15

Estimated time costs associated with developing a drug: $1.2B

Total cost of developing a drug from the lab, through the clinic, and to market: $3B

Approval rate that has decreased by half since 2003: 12%

Sources:
1. Tufts Center for the Study of Drug Development
2. The Centre for Medicines Research
What is AI?

Definition of artificial intelligence

Artificial Intelligence (AI):
• “Science and engineering of making intelligent machines”

Machine Learning (ML):
• “An artificial intelligence technique that can be used to design and train software algorithms to learn from and act on data”
Traditional VS AI-based drug discovery methods

**TRADITIONAL**

- Target-driven
- Work well for easily druggable targets that have a well-defined structure and whose interactions inside the cell are understood in detail
- Extremely limited due to the complex nature of cellular interactions & limited knowledge of intricate cellular pathways

**AI-BASED**

- Data-driven
- Complex algorithms and machine learning can extract meaningful information from a large dataset
- Identify compounds that could bind to 'undruggable targets', i.e., proteins whose structures are not defined

---

A predictive set of compounds can be easily identified with AI in a relatively small amount of time and at a quarter of the cost of traditional methods.
Chemical data and compound libraries

- Quantitative structure activity relationship (QSAR)
- Absorption, distribution, metabolism, excretion and toxicity (ADMETox)
- Ligand-protein docking: Target 3-D structures
- PrOCTOR: predicting clinical trial failures
“OMICs” data
Transcriptome, Proteome and Metabolome: understanding the interactome

Neighborhood analysis
• Understanding target liability by association

Adverse outcome Pathways (AoPs)
• Integrative data analysis to understand mechanism of toxicity

Source: Bioinfogate: OFF-X (Safety maps)

Toxicogenomics

- Influence of genome on adverse events
  - Role of gene variants, polymorphisms
  - Role of epigenetics (e.g., DNA methylation, histone acetylation, RNA silencing) in adverse outcomes

- Examples
  - Anti-virals and drug hypersensitivity (e.g., Abacavir/HLA-B*5701)
  - Anthracyclines and cardiotoxicity. (e.g., ABCC1 G671 V)
  - Corticosteroids and Osteonecrosis (e.g., TYMS, enhancer tandem repeat)

Coupling AI with network biology to enable disease understanding and target ID (MetaCore™)
Networks are how biology works

- Disease mechanism understanding
- Target ID

Network by Martin Grandjean
Network Biology in a nutshell

<table>
<thead>
<tr>
<th>Interaction networks</th>
<th>Association networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node A</td>
<td>Gene A</td>
</tr>
<tr>
<td>Node B</td>
<td>Gene B</td>
</tr>
<tr>
<td>Edge</td>
<td>Co-expression</td>
</tr>
<tr>
<td>Protein A</td>
<td>Protein B</td>
</tr>
<tr>
<td>Protein B</td>
<td>Protein A</td>
</tr>
<tr>
<td>Physical interaction</td>
<td>Targeting drug sharing</td>
</tr>
<tr>
<td>Protein</td>
<td>Gene</td>
</tr>
<tr>
<td>Gene</td>
<td>Drug A</td>
</tr>
<tr>
<td>Protein</td>
<td>Drug B</td>
</tr>
<tr>
<td>Transcriptional regulation</td>
<td>Protein target sharing</td>
</tr>
<tr>
<td>Substrate</td>
<td>Product</td>
</tr>
<tr>
<td>Product</td>
<td>Gene A</td>
</tr>
<tr>
<td>Biochemical reaction</td>
<td>Gene B</td>
</tr>
<tr>
<td>Gene</td>
<td>Genetic disorder sharing</td>
</tr>
<tr>
<td>Protein</td>
<td>Ligand/small molecule</td>
</tr>
<tr>
<td>Physical binding</td>
<td>Disorder A</td>
</tr>
<tr>
<td>Disorder B</td>
<td>Disease gene sharing</td>
</tr>
</tbody>
</table>

- Network represents the presence of objects or entities in a system as “nodes”, and the relationships or interactions among the nodes are called “edges”

- nodes → biological molecules such as genes, proteins, and ligands, or even larger entities such as cells or individual humans.

- Edges → physical interactions or contacts between biological molecules, biochemical processes between substrates and products, genetic interactions between genes, and in some cases, interactions between cells or individual organisms.

Figure taken from Jinawath, N. et al. J Transl Med, 2016.
Coupling biological networks with deep neural networks to enable disease understanding and target ID

Node embeddings from random walks in (1) structural graph; and (2) attribute graph

 training set of known targets

GWAS hits

KInases

DEGs

Novel predicted targets
Key challenges

- Garbage in – garbage out
  - The need for high-quality networks
  - And large high-quality training sets

- Knowledge bias
  - It’s hard to predict completely unknown from the known

- Model interpretation
  - Opening the “black box”
Triple Negative Breast Cancer

ER−, PR−, HER2−
High proliferation, poor differentiation, basal marker (cytokeratin 5/6) positive
Shorter overall and disease-free survival
Greatest challenge: No effective therapy and Lack of therapeutic targets

Hudis and Gianni, The Oncologist, 2011
Active drug development in TNBC

- 456 drug and biologics under development
- 164 in clinical trials
- 43 launched

Source: Cortellis Drug Discovery Intelligence
Active drug development

- Vast majority of the TNBC drugs in clinical trials have no specific mechanisms.

- Drugs launched for treating advanced cancer types are in Phase II/III trials for TNBC.

- Trilaciclib hydrochloride, a small-molecule inhibitor of cyclin-dependent kinases 4 and 6 (CDK4/6) developed by G1 Therapeutics, is in Phase III trials.

Source: Cortellis Drug Discovery Intelligence
Among 640 novel therapeutics of Phase 3 clinical trials (1998-2008), 344 (54%) failed in clinical development, 230 (36%) were approved by the US Food and Drug Administration (FDA), and 66 (10%) were approved in other countries but not by the FDA. Most products failed due to inadequate efficacy (n = 195; 57%), while 59 (17%) failed because of safety concerns and 74 (22%) failed due to commercial reasons.
How do we leverage AI to understand MoA and identify new targets in TN breast cancer?
Understanding biology is critical for target ID

- Networks can uniquely identify potential effectors that target distinct host nodes
- Determine how mutations can affect the interaction
- Delineate which targets have high therapeutic potential

Genetic interaction network from the microarray gene expression profile to identify the molecular mechanisms involved in breast cancer liver metastases
Understanding biology is critical for target ID

- DEG analyses in high-throughput techniques, such as RNA-sequencing (RNA-Seq), are performed by comparing gene expression profiles between two different conditions.
- The results define a set of genes with high expressions in cancer and low expressions in normal tissue through statistical tests.
- However, in some cases, non-differentially expressed genes can contribute to disease dysfunction through clusters of co-expressed genes.
- These genes may manifest their functions through interaction networks with other differentially expressed genes.

Chen et al., Aging 2021
Ideal literature curation workflow

Manual curation ensures the high quality

Define project
Define curation template, inclusion/exclusion criteria and prioritization strategy

Construct search strings
Find relevant articles for review

Review and prioritize abstracts
Manual review and prioritize based on inclusion/exclusion criteria

Acquire data and articles
Experience in biomedical literature monitoring

Annotate and curate articles
Controlled vocabularies and public database IDs

QC and format for delivery
Knowledge in development of biological databases
How is a database like this constructed?

A solution for interactome reconstruction, data management and integration

Literature curation and database construction

Curator
- Annotates
- Enriches data
- Quality control

Articles and data
- Metabolite-host interactions
- Microbe-microbe interactions
- And more

Administrator
- Design
- Development
- Maintenance

Proprietary data sources

User interface

User query

Summary statistics

Interaction networks
- Table of interactions
- Access to related articles
- And more

Database of Breast Cancer Gene/Protein Interactions

Public data sources

Articles and data

• Metabolite-host interactions
• Microbe-microbe interactions
• And more
The interaction database will help leveraging AI

Novel predicted targets
Using Curated Data to better inform Drug Discovery decisions (Cortellis™)
Have confidence and make data-driven decisions

<table>
<thead>
<tr>
<th>622K+ drugs and biologics</th>
<th>46K+ genes and targets records</th>
<th>169K+ experimental models</th>
</tr>
</thead>
<tbody>
<tr>
<td>44K+ biomarkers</td>
<td>2.3M+ biomarker uses</td>
<td>2.9M+ pharmacological data points</td>
</tr>
<tr>
<td>485K+ patents</td>
<td>2.8M+ literature records</td>
<td>Used by 90% of the top 20 pharma companies</td>
</tr>
</tbody>
</table>
Target validation using CDDI

• Quickly retrieve information on drug targets with detailed information on their roles, lifecycle phases and references to easily validate and prioritize potential targets

• Make more informed go/no-go decisions based on previous and active research

• Link to all related information directly in CDDI to make it easier to connect-the-dots
See the various routes of synthesis
Can be used to find non-infringing routes of synthesis
Which synthetic route needs minimal work-up?
Cheapest route of synthesis
Chirality in synthesis? Does reaction yield correct enantiomer?
Toxic reagents used?
Unstable intermediates?
Use Case 1: Lab working on a particular target

1. Is the target druggable?
2. How crowded is the area? What are the different disease pathologies where this target plays a role?
3. Which chemical structures have inhibitory activity against the target?
4. What are the animal models that have been used?
5. What are the assays used for testing target inhibition?
6. What is the latest information available from conferences?

Goal
   a. Understand the role of the target in the disease
   b. What are the pharmacological approaches being used for modulating the target?
   c. Identify an effective pharmacophore
   d. Generate an SAR table
   d. Identify if any lead compound has progressed to human trials
Use Case 1: Research Group working on MDM2

Quick Search MDM2

Go to Targets and Pathways

Choose Target
Targetscape-Validation of the Target
a) Go to experimental pharmacology records
b) Filter by value range (IC50 in M)
c) Select 1X10E-9 IC50
d) Generate SAR table
Identify Pharmacophore

Experimental Pharmacology Search Results

<table>
<thead>
<tr>
<th>Drug Name &amp; Structure</th>
<th>Mechanism of Action</th>
<th>Material</th>
<th>Method</th>
<th>Value</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>7955000</td>
<td>MDM2 (hom) Inhibitors</td>
<td>Recombinant human protein</td>
<td>Fluorescent polarization assay</td>
<td>$&lt; 100 \mu M$</td>
<td>Fig. 4</td>
</tr>
<tr>
<td>C-15</td>
<td></td>
<td></td>
<td></td>
<td>$&lt; 3.00 \mu M$</td>
<td>Fig. 2</td>
</tr>
<tr>
<td>ML-1661</td>
<td>MDM2 (hom) Inhibitors</td>
<td>Recombinant human protein</td>
<td>Fluorescent polarization assay</td>
<td>$&lt; 100 \mu M$</td>
<td>Fig. 4</td>
</tr>
<tr>
<td>ML-219</td>
<td>Adipocytes Inducers</td>
<td>MDM2 (hom) Inhibitors</td>
<td>Recombinant human protein</td>
<td>$&lt; 1.00 \mu M$</td>
<td>Fig. 2</td>
</tr>
<tr>
<td>ML-516-54</td>
<td>MDM2 (hom) Inhibitors</td>
<td>Adipocytes Inducers</td>
<td>Fluorescent polarization assay</td>
<td>$&lt; 1.00 \mu M$</td>
<td>Fig. 4</td>
</tr>
</tbody>
</table>
• Filter by molecular mechanisms
• Choose mechanism of action of interest
• Generate a more focused SAR table
• Look at PK/Met table to see if human studies done.
• Finally, do literature search <Source=AACR on latest research in MDM2>
Generate SAR for a specific pharmacophore against a specific target

Find out the Candidate that has progressed to human studies
Use Case 2: Lab working on a particular disease

Typical Questions

1. What therapeutic approaches and targets are being investigated in atherosclerosis?
2. Which chemical structures have inhibitory activity against PCSK9?
3. What are the animal models that have been used?

Goal

a. Identify novel targets
b. What are the pharmacological approaches being used for modulating the target?
c. Generate an SAR table
Use Case 2: Lab working on a particular disease

Target of interest: PC9
### Comparing Activities: SAR Table

#### Experimental Activity: Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9) inhibition, IN VITRO

**Pharmacological Activity: Kexin (PC SK9), inhibition**

<table>
<thead>
<tr>
<th>Drug Name &amp; Structure</th>
<th>Mechanism of Action</th>
<th>Material</th>
<th>Method</th>
<th>Value</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>566410</td>
<td>Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9)</td>
<td>Human liver cancer cells transduced with human PCSK9</td>
<td>Chemiluminescent assay</td>
<td>0.660 μM</td>
<td>Pat. 2</td>
</tr>
<tr>
<td>566440</td>
<td>Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9)</td>
<td>Human liver cancer cells transduced with human PCSK9</td>
<td>Chemiluminescent assay</td>
<td>0.520 μM</td>
<td>Pat. 2</td>
</tr>
<tr>
<td>566442</td>
<td>Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9)</td>
<td>Human liver cancer cells transduced with human PCSK9</td>
<td>Chemiluminescent assay</td>
<td>0.440 μM</td>
<td>Pat. 2</td>
</tr>
<tr>
<td>566443</td>
<td>Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9)</td>
<td>Human liver cancer cells transduced with human PCSK9</td>
<td>Chemiluminescent assay</td>
<td>0.680 μM</td>
<td>Pat. 2</td>
</tr>
<tr>
<td>566444</td>
<td>Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9)</td>
<td>Human liver cancer cells transduced with human PCSK9</td>
<td>Chemiluminescent assay</td>
<td>1.06 μM</td>
<td>Pat. 2</td>
</tr>
<tr>
<td>566445</td>
<td>Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9)</td>
<td>Human liver cancer cells transduced with human PCSK9</td>
<td>Chemiluminescent assay</td>
<td>1.56 μM</td>
<td>Pat. 2</td>
</tr>
</tbody>
</table>
Use Case 3: Find the toxicity of a class of compounds

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Mechanism of Action</th>
<th>Experimental Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>953460</td>
<td>Phosphatidylinositol-3-Kinase (PI3K) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>745383</td>
<td>Phosphatidylinositol-3-Kinase gamma (PI3Kgamma) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>775625</td>
<td>Phosphatidylinositol-3-Kinase delta (PI3Kdelta) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>774330</td>
<td>Phosphatidylinositol-3-Kinase alpha (PI3Kalpha) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>756167</td>
<td>Phosphatidylinositol-3-Kinase delta (PI3Kdelta) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>735149</td>
<td>Phosphatidylinositol-3-Kinase beta (PI3Kbeta) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>725459</td>
<td>Protein Kinase B (PKB/Akt) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>720505</td>
<td>Phosphatidylinositol-3-Kinase (PI3K) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>756167</td>
<td>Phosphatidylinositol-3-Kinase delta (PI3Kdelta) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>735149</td>
<td>Phosphatidylinositol-3-Kinase beta (PI3Kbeta) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
<tr>
<td>720505</td>
<td>Phosphatidylinositol-3-Kinase gamma (PI3Kgamma) Inhibitors</td>
<td>Experimental Pharmacology</td>
</tr>
</tbody>
</table>

Experimental Activity:
- Toxicity/Adverse Events
- Potassium Voltage-Gated Channel Subfamily M Number 2 (KCNQ2) Inhibitors
- In Vitro

Experimental Pharmacology:
- Mechanism of Action:

<table>
<thead>
<tr>
<th>Drug Name</th>
<th>Mechanism of Action</th>
<th>Material</th>
<th>Dose</th>
<th>Value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>953460</td>
<td>Phosphatidylinositol-3-Kinase (PI3K) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>745383</td>
<td>Phosphatidylinositol-3-Kinase gamma (PI3Kgamma) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>775625</td>
<td>Phosphatidylinositol-3-Kinase delta (PI3Kdelta) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>774330</td>
<td>Phosphatidylinositol-3-Kinase alpha (PI3Kalpha) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>756167</td>
<td>Phosphatidylinositol-3-Kinase delta (PI3Kdelta) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>735149</td>
<td>Phosphatidylinositol-3-Kinase beta (PI3Kbeta) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>720505</td>
<td>Protein Kinase B (PKB/Akt) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>725459</td>
<td>Phosphatidylinositol-3-Kinase gamma (PI3Kgamma) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>756167</td>
<td>Phosphatidylinositol-3-Kinase delta (PI3Kdelta) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>735149</td>
<td>Phosphatidylinositol-3-Kinase beta (PI3Kbeta) Inhibitors</td>
<td>Experimental Pharmacology</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Searching for Collaborators
Challenges and Strategies
Challenges and strategies

**Challenge: Data quality**
- QSAR training sets: rely on experimental data such as solubility, pharmacology data
- Machine learning: quality, training set collected and trained by a user
- Quantity: datasets needed for training run in tens of thousands
- Annotation: what conditions was testing performed in? Definition of toxicity?
- Biomarkers: reliable and validated

**Strategy: Databases**
- Public and subscription based databases
- Quality of curation, sources
Challenges and strategies

Challenge: Lack of reproducibility
- Lack of publishing guidelines

Strategies: Guidelines for reporting
- Computational infrastructure
- Training sets and performance
- Implemented code
- Average run time
AI is promising significant advances in the data-rich biomedical field.

Biological networks are different from common AI inputs but approaches have emerged to feed biological networks into AI techniques.

Manual curation remains important for creating high-quality biological networks and training sets for AI.

Time will show how much of transformation versus incremental progress AI will bring into pharma R&D.
Further Reading

4. https://www.biorxiv.org/content/early/2018/01/07/243998
5. https://www.nature.com/articles/nmeth.3940
6. http://msb.embopress.org/content/13/4/924
7. https://www.nature.com/articles/nmeth.4397
10. https://www.nature.com/articles/nmeth.4169
12. https://elifesciences.org/articles/25754
13. https://genome.cshlp.org/content/28/3/334
17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577280/#:~:text=AI%20can%20be%20used%20effectively,and%20uncertainty%20of%20the%20data.