The convergence of artificial intelligence and chemistry for drug discovery

Clive Green
ELRIG Drug Discovery, Liverpool, UK

October, 2017
89% failure rate for pre-clinical research

AstraZeneca internal measure for the period 2015-17
Putin: Leader in artificial intelligence will rule world
Evolution in AI

Artificial Intelligence
A computerized system that exhibits behaviour that is commonly thought of as requiring intelligence

Machine Learning
A statistical process that starts with a body of data and tries to derive a rule or procedure that explains the data or can predict future data

Deep Learning
A subfield of ML which uses structures loosely inspired by the human brain, consisting of a set of units (or "neurons")
Rich history of predicting molecular properties

**Molecular Property Prediction**

Hammet equation – relating reaction rates and equilibrium constants for reactions of benzoic acid derivatives with meta- and para-substituents to each other using a substituent constant and a reaction constant

**QSAR**

Hansch equation – computer-assisted identification and quantification of physicochemical properties of a drug on biological activity

**Learning Algorithms**

Modelling of complex and high dimensional data sets

1930’s

1960’s

2000’s

An explosion of measurements fuelled by technologies that make assays easier
Property prediction enables high value lab testing

Lab testing in areas where we can’t predict

Making accurate predictions with confidence
Our scientific interest

What to make next?

De novo design

How to make it?

Retrosynthesis
De novo design methods

Structure-based approaches

Virtual library search
Can we conduct molecular de novo design by learning from structures already designed and synthesized by human experts?

Database of high-level Go games

Pre-trained Network

Play itself

Reinforcement Learning

AlphaGo

Database of chemical designs

Pre-trained Network

Scoring function

Reinforcement Learning

De Novo Design Network
Learning from doing

Often use pre-trained model as a starting point

- Design molecule
  - Active?
    - Good ADME?
    - Synthetically accessible?
- Make more like this?
- Make something else instead?
Apply concepts from natural language processing to SMILES

• Conditional probability distributions

• \( P(green \mid is, grass, The) \)

• \( P(O \mid =, C, C) \)
Tokenisation of SMILES

Graph:

\[
\text{Cl} \quad \text{Cl}c1c[nH]cn1
\]

SMILES:

<table>
<thead>
<tr>
<th>Cl</th>
<th>C</th>
<th>c</th>
<th>l</th>
<th>c</th>
<th>nH</th>
<th>c</th>
<th>n</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

One-hot encoding:
Training the initial Network

- Restrain structures to **10 to 50** heavy atoms
- Restrain atoms to **H, B, C, N, O, F, P, S, Cl, Br, I**
- **1.5 million SMILES** from ChEMBL

Molecular de-novo design through deep reinforcement learning; Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen; Journal of Cheminformatics, 2017, 9:48
https://doi.org/10.1186/s13321-017-0235-x
Molecular de-novo design through deep reinforcement learning; Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen; Journal of Cheminformatics, 2017, 9:48
https://doi.org/10.1186/s13321-017-0235-x
How a model thinks while generating a molecule
Observations and performance

Initially, the distribution is an approx. of the distribution of first tokens for SMILES in the training set.

Once the aromatic "n" has been sampled, the model has come to expect a ring opening (i.e. a number).

Once opened, the aromatic atoms become probable, until 5-6 steps later when the model thinks it is time to close the ring.

94% of generated sequences are valid SMILES and 90% are novel. Most common error is not closing an opened ring or branch.

Molecular de-novo design through deep reinforcement learning: Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen; Journal of Cheminformatics, 2017, 9:48
https://doi.org/10.1186/s13321-017-0235-x
Randomly selected structures generated by the model
Identifying a unique molecule

- Chose Celecoxib

- Scoring function based on similarity

- After 1000 training steps the model generates **only Celecoxib when all structures with similarity to Celecoxib >0.5 are removed from the training set**

- What about generating structures that are moderately similar to the query structure?
How a model thinks while generating structural analogues
Generating structural analogues
Generating structures with predicted biological activity

- Dopamine receptor type 2 (DRD2)
- Using activity model (SVM model) as the scoring function
- Removed all actives from ChEMBL
- The model generates structures of which >95% are predicted to be active
- The model recovered 7% of test actives – it has learnt to generate "novel" structures that have been seen neither by the DRD2 activity model nor the pre-trained network, and are experimentally confirmed actives

Molecular de-novo design through deep reinforcement learning; Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen; Journal of Cheminformatics, 2017, 9:48
https://doi.org/10.1186/s13321-017-0235-x
Structures designed by the model to target DRD2

Recovered test actives

Randomly selected
A chemists view of automated retrosynthesis…

- **Creative** process is treasured by chemists
- Proposed routes often contain what chemists immediately recognize as **chemically unreasonable** steps – a lack of chemical intelligence
- **Time-consuming** task and humans are not infallible

- We don’t need automated tools
- Retrosynthesis is the **chemists job**
- AI can’t help with retrosynthesis
Search tree representations

- Number of **applicable transformations** per step (branching factor) and the **search area** can be **vast** if solutions are explored exhaustively.
- **Ranking** alternatives is **difficult**
  - Chemists tend to disagree about how easy it is to make a molecule
  - Value of a position depends highly on the availability of suitable precursors

---

Learning to plan chemical synthesis; Marwin H. S. Segler, Mike Preuss and Mark P. Waller; arXiv, 2017
arXiv:1708.04202
The model

Key features

- Guided into promising directions by proposing a restricted number of possible, automatically extracted transformations

- Predicts whether the corresponding reactions are actually feasible

- Networks were trained on 12.4 million reactions from the Reaxys database
Experimental performance

- Model finds a 6-step synthesis route to the intermediate drug candidate autonomously in 5.4 seconds.
- The route is identical to that originally published in 2015.
- Chemists did not significantly prefer literature routes over routes found by the model.
- Unsolved: natural product synthesis and stereochemistry.
For the future…

- SMILES from ChEMBL
- Pre-trained network
- Scoring function
- Augmented likelihood
- De-novo design network
- SMILES strings

Initiate network → De-novo design network → Generate sequences → SMILES strings

- Consortium / collaborative data access
- Bioactivity
  - Synthetic accessibility
  - ADME
  - Toxicity

Update network likelihood
Acknowledgments

- Ola Engkvist
- Lars Carlsson
- Thierry Kogej
- Marcus Olivecrona
- Claus Bendtsen
- Garry Pairaudeau
- Thomas Blaschke
- Hongming Chen
Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 2 Kingdom Street, London, W2 6BD, UK, T: +44(0)20 7604 8000, F: +44 (0)20 7604 8151, www.astrazeneca.com