Cheminformatics in Drug Discovery, an Industrial Perspective

Hongming Chen, Thierry Kogej, Ola Engkvist
Hit Discovery, Discovery Sciences, Astrazeneca R&D Gothenburg
We need to become better, faster & cheaper
Cheminformatics @ AstraZeneca

- HTS work-up
- Library design
- Virtual screening
- Machine learning & AI
High Throughput Screening

From Millions to just a few

Low cost/compound

Primary Screening (Millions)

Confirmation (tons of thousands)

Dose Response (a few thousand)

Trace (Hundreds)

~0-4 Chemical Series

High cost/compound

Slide modified from Mark Wigglesworth, AZ, with permission
HTS Analysis: Clustering analysis

Early days

- Heavily dependent on computational chemistry resources
- Linux, scripts, static workflows, data in flat files
- Cutting, pasting and reformatting between applications
- Difficult to revisit or take over an analysis from a colleague
- Time-consuming

iHAT: An Spotfire add-in for HTS Analysis

- Leverage the powerful visualization function of Spotfire
- Annotation of compounds with in-house experimental and predicted data
- Data integration from multiple sources
- Clustering of compounds
- Visualization and manipulation of cluster tree
- NN search
iHAT: Clustering and Reclustering
Library design @ AstraZeneca

- Diversity library is generally out of fashion
- Focused library fit for specific project need
- DNA encoded libraries become popular, but analysis is challenging, >60M to 8B library sizes

Currently, use classical library design method to reduce to 50M preferred AZ library size
Definition of VS

• Virtual screening refers to any in-silico techniques used to search large compound databases (available chemicals or virtual libraries) to select a smaller number for biological testing

• Virtual screening can be used to
  − Select compounds for screening from in-house databases
  − Choose compounds to purchase from external suppliers
  − Select compounds from virtual libraries to be synthesized

• The technique applied depends on the amount of information available about the particular disease target and the desired outcome
VS methods

3D Structure of Target

Unknown

Lots of actives and inactives known

Ligand-based methods

Actives known

aligned conformation

Co-crystallized ligand conf.

structure-based methods

Protein structure

2D ligand

• Substructure search;
  • Fingerprint-based similarity search

• Expand SAR
  • Improve affinity

3D ligand

• Shape similarity search
  • Pharmacophore mapping
  • Core replacement

• Scaffold hopping
  • Filter, improve hit rate

Binding pocket

• Docking and scoring
  • Pharmacophore mapping
  • Shape similarity

• Filter, improve hit rate

QSAR models

Known
• Identification of sPLA2X inhibitors using ligand and structure based virtual screening

IC50 (NMR) 20uM
Virtual screening platform @ AZ

Theoretical chemical space

(AZ-Virtual space $10^{15}$
- Diverse: > 4000 libraries
- Synthesizable: historic AZ-reactions
- Searchable in subsets of $10^{8-11}$)

Known existing compounds $10^8$

AZ $10^6$

Iterative virtual screening workflow

- VS chemical space $10^{15}$
- Sub-sets $10^{8-11}$
- Primary VS $10^{8-11}$
- Virt. hits $10^6$
- Secondary VS $10^6$
- 20-30 virt. lib.s
- 1-3 libs./50-100 cpd. ordered

Ligand

Structure, Ligand, Properties, Novelty

Reactions, reagents

Computational strategy

$10^{15}$

3D similarity search
Structure based workup

Representative subset (as large as possible)

Virtual Synthesis

Few libraries are enriched specifically to the query project

$10^5$

Selection of libraries (3D similarity and structure based workup)
AI & Machine Learning Today
Context, Definition & Advances

Source: http://www.geeksforgeeks.org/artificial-intelligence-an-introduction/

Artificial Intelligence (AI)
- deep learning
- predictive analytics
- translation
- machine learning
- classification & clustering
- natural language processing (NLP)
- information extraction
- speech to text
- text to speech
- speech
- expert systems
- planning, scheduling & optimization
- robotics
- vision
- image recognition
- machine vision

Source: http://hduongtrong.github.io
The rise of deep learning in drug discovery

- Deep learning technologies have been adopted in drug discovery
- Various forms of NN have been applied so far
De novo molecular generation with deep learning has developed very rapidly.
Deep learning @ AstraZeneca: Vision

• Creating an integrate AI platform to impact drug discovery projects

Deep learning @ AZ: De Novo Molecular Augmented Design Platform (REINVENT)

Generation of novel chemical space

AZ+ Pub

Prior model

Reinforcement learning to generate project relevant compounds

Agent model

Iterations

Desirability function $\sum IC50$, LogP, Novelty etc.

Iterations of design and compound synthesis
Deep learning at AstraZeneca: Reaction informatics

• First steps, building:
  – World-class Reaction Knowledge Base
  – On our work (past collaboration with M. Segler)

---

**Predictive chemical reaction models**

- AZ Reaction Connect
  ~20mill reactions

---

**Reaxys®**

- MedChem ELN
- AstraZeneca
- PharmSci ELN
- Reaxys
- Patent data

---

**AZ ChemistryConnect**

- iLab
  - Design
  - Make
  - Test
  - Analyst

---

**Planning chemical syntheses with deep neural networks and symbolic AI**

To plan the synthesis of small organic molecules, chemists use retrosynthetic, a problem-solving technique in which larger molecules are recursively transformed into functional high-energy precursors. Computer-aided retrosynthesis would be invaluable but at present it is slow and provides results of uncertain quality. Here we use Reaxys (a chemical search tool) and evaluate a new approach, called Reaction Networks. In the traditional reaction design stage, chemical decomposition is performed to break down the chemical structure into substructures. This can be done by chemical reactions or by systematic schemes. These deep neural networks were trained on essentially all reactions ever published in organic chemistry. The approach has been shown to be accurate, faster than systematic schemes, and generates a wider range of molecules than the traditional computer-aided search method, which is based on canonical rules and hand-designed heuristics. In a double-blind Aït All test, chemists on average synthesized test compounds generated randomly to be evaluated in reported literature texts.
Becoming FASTER with AI
Through unsupervised learning for hit identification

Deep CNN autoencoder ➔ Manifold Learning (t-SNE) ➔ Deep CNN classifier
Becoming CHEAPER with ML/AI
By only conducting experiments when needed

- Extract features (signature descriptor)
- Build/Apply ML model (Mondrian TCP on SVM)
- p-value (class 😃)

Don’t synthesize

Apply ML model (Mondrian TCP on SVM)

p-value (class 😃)

Don’t test
Becoming FASTER and CHEAPER with AI

AI augmented *de novo* molecule design

Molecular de-novo design through deep reinforcement learning

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist and Hongming Chen

RNN + Reinforcement learning
AZ’s first DMTA automation platform

- First prototype built during 2017
- All DMTA steps fully integrated
- Suited for 100s of uninterrupted DMTA cycles. ML/AI module is integrated.
- Cycle times of ca. 2h
- Successfully applied in ongoing research project
Conclusions

• Cheminformatics is widely applied in Pharmaceutical industry

• Cheminformatics includes various aspects across different disciplines

• Adoption of machine learning and AI technologies will help Cheminformatics to better fit current and future research needs
Acknowledgement

- Marcus Olivercrona
- Thomas Blaschke
- Isabella Feierberg
- Christophe Grebner
- Erik Malmerberg
- Christian Tyrchan
- Garry Pairaudeau
- Clive Green
- Lars Carlsson
- Peter Varkonyi
- Michael Kossenjans

- EU Fundings
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, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK, T: +44(0)203 749 5000, www.astrazeneca.com