Applications of machine learning and artificial intelligence to designing chemicals and materials with the desired properties

Alexander Tropsha
UNC Eshelman School of Pharmacy
Outline

• Brief notes on machine learning/QSAR
• Materials Informatics and Materials Design
• Design, development and application of the **Reinforcement Learning for Structural Evolution (ReLeaSE)***
• Summary and future work: QSAR without borders

Machine Learning Framework

\[ y = f(x) \]

- **Training:** given a *training set* of labeled examples \( \{(x_1, y_1), ..., (x_N, y_N)\} \), estimate the prediction function \( f \) by minimizing the prediction error on the training set.
- **Testing:** apply \( f \) to a never before seen *test example* \( x \) and output the predicted value \( y = f(x) \).
The newly-appointed President-Elect of the Royal Society of Chemistry today forecast the impact of advances in modelling and computational informatics on chemistry.

Next RSC president predicts that in 15 years no chemist will do bench experiments without computer-modelling them first

Jul 17, 2013

Professor Dominic Tildesley, who will become president in 2014, said: "The speed and development of computers is now so rapid, and the advances in modelling and informatics are so dramatic that in 15 years’ time, no chemist will be doing any experiments at the bench without trying to model them first."

Professor Tildesley is a world-leading expert in large-scale computational modelling and...
Automated Retrosynthesis (Chematica)
The growing appreciation of molecular modeling and informatics

A brave new world of robot chemists and 'synthesiser farms' awaits

**Wanted: synthetic chemists (humans need not apply)**

24 JANUARY 2018

Automation could free chemists from tedious lab work – if they're ready to think differently about research
Promise of dramatic acceleration of drug discovery

GSK Has Developed A New Analytics Platform That Can Reduce The Time It Takes To Analyze Clinical Data From Months To Clicks
Source: Thomas Macaulay, CIO UK
March 12, 2018

The platform uses large-scale data analytics to drive better decisions about the drug discovery pipeline, by allowing the pharmaceuticals giant to test the potential for new drugs before it begins clinical trials.
Deloitte Insight: Over 100,000 legal roles to be automated

Added on the 16th Mar 2016 at 10:28 am

Over 100,000 jobs in the legal sector have a high chance of being automated in the next twenty years, according to extensive new analysis by Deloitte.

The Deloitte Insight report, which predicts “profound reforms” across the legal profession within the next 10 years, finds that 39% of jobs (114,000) in the legal sector stand to be automated in the longer term as the profession feels the impact of more “radical changes.”
The ultimate dream of a computational chemist
The chief utility of computational models: Annotation of new compounds

10^6 – 10^9 molecules

- CHEMICAL STRUCTURES
- CHEMICAL DESCRIPTORS
- PREDICTIVE QSAR MODELS
- PROPERTY/ACTIVITY

CONFIRMED ACTIVES (TOXIC)

CONFIRMED INACTIVES (NON-TOXIC)

CHEMICAL DATABASE

VIRTUAL SCREENING

QSAR MAGIC
QSAR Modeling Workflow: the importance of rigorous validation

5-fold External Validation
courtesy of L. Zhang

Datasets

Modeling set

External set

Experimental confirmation

Virtual screening (with AD threshold)

Evaluation of external performance

An ensemble of QSAR Models

Internal validation

Model selection

Combi-QSAR modeling

Modeling methods

K-Nearest Neighbors (kNN)
Random Forest (RF)
Support Vector Machines (SVM)

Descriptors

Dragon
MOE


Fully implemented on CHEMBENCH.MML.UNC.EDU
Material Science and the Rise of Materials Informatics

- Explosive growth of materials data, both experimental databases and computational repositories.
  - **Structural data**: 160,000 entries in the Inorganic Crystal Structure Database (ICSD)
  - **Experimental data**: Numerous commercial and open experimental databases NIST, MatWeb, MatBase etc.
  - **Computational data**: Huge databases such as AFLOWLIB, Materials Project, and Harvard Clean Energy
  - Chemical space of possible materials is HUGE \(~10^{100}\) candidates [Nat. Chem. 7, 274-275 (2015)]

- Materials Genome Initiative or MGI (US Govt): Need for new high performance materials
Closing the gap: materials structure-property relationships

Material Informatics/MQSAR Workflow

Experimental structures

Data parsing
Error checking

Descriptor generation

Geometry optimization
Band structure & Property calculations

Aflowlib database

Properties DBs
Commercial/Open DBs
Literature & reference data

Error check
Duplicate removal
File conversion
Unit conversion

Data Integration

Band structures,
DOS, Symmetry,
Geometry, Unit cell

Descriptor generation

Fingerprints, clustering, and modeling

New Materials
With Desired
Predicted
Properties

Experimental validation

Predictive QSAR modeling

Data driven discovery

Material Map (B-Fingerprints)

>15000 materials from ICSD
DFT PBE calculations from aflowlib.org

Cluster C: metallic comp. with non-metallic atoms
Cluster B: bimetals, polymetals
Cluster D: small band gap comp., semiconductors
Cluster A: insulators, ceramics, complex oxides

Orphans

Band gap, eV

Systematic representation of materials using fragment descriptors

A. Crystal Structure

B. Voronoi tessellation and neighbors search

C. Infinite periodic graph construction and property labeling (EA, IP, En, Rcov, etc)

D. Decomposition to fragments

Nodes (atoms)

Edges (bonds, vDW contacts)

Path fragments of length N, N = 2, 3, ...

Circular fragments (polyhedrons)

All models are trained based on DFT-computed properties (VASP s/w from U. Vienna)

Prediction of Electronic Properties

Classification accuracy 95%
ROC Curve (AUC) 0.98

Learning approach for all models:
Gradient Boosting Decision Trees (GBT)

Prediction of Thermomechanical Properties

($E_{BG}$ - band gap energy; $B_{VRH}$ - bulk modulus; $G_{VRH}$ - shear modulus; $\theta_D$ - Debye temperature; $C_P$ - heat capacity at constant pressure; $C_V$ - heat capacity at constant volume; $\alpha_V$ - thermal expansion coefficient)

<table>
<thead>
<tr>
<th>property</th>
<th>RMSE</th>
<th>MAE</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{BG}$</td>
<td>0.51 (eV)</td>
<td>0.35 (eV)</td>
<td>0.90</td>
</tr>
<tr>
<td>$B_{VRH}$</td>
<td>14.25 (GPa)</td>
<td>8.68 (GPa)</td>
<td>0.97</td>
</tr>
<tr>
<td>$G_{VRH}$</td>
<td>18.43 (GPa)</td>
<td>10.62 (GPa)</td>
<td>0.88</td>
</tr>
<tr>
<td>$\theta_D$</td>
<td>56.97 (K)</td>
<td>35.86 (K)</td>
<td>0.95</td>
</tr>
<tr>
<td>$C_P$</td>
<td>2.31 ($k_B$/cell)</td>
<td>0.84 ($k_B$/cell)</td>
<td>0.99</td>
</tr>
<tr>
<td>$C_V$</td>
<td>2.01 ($k_B$/cell)</td>
<td>0.70 ($k_B$/cell)</td>
<td>0.99</td>
</tr>
<tr>
<td>$\alpha_V$</td>
<td>$1.47 \times 10^{-5}$ (K)$^{-1}$</td>
<td>$5.69 \times 10^{-6}$ (K)$^{-1}$</td>
<td>0.91</td>
</tr>
</tbody>
</table>

**TABLE I.** Statistical summary of the five-fold cross-validated predictions for the seven regression models (Figure 3).
Summary of Materials Informatics: Methods

- Fast, accurate general purpose machine learning methods for material’s property prediction. **Millisounds on laptop vs. days on HPC cluster**
- Universal applicability to different materials: currently covered **85 elements** (H – Pu, without noble gases, Tc, Fr, Ra). All types of crystal lattices and symmetries.
  - Most competing approaches are **specific to one prototype/family of materials or single property**
- Works for other properties: elastic, thermoelectric, etc.
- Possible to gain *some* chemically/physically interpretable insight into “black box” model.
- Possible to derive materials design rules
- User friendly web app and RESTful API (http://aflow.org/aflow-ml/)
Photocathode materials
Evaluated as DSSCs

Dye-sensitized solar cells (DSSCs)
Design of alternate photocathodes
A materials informatics approach

(AFLOWLIB)

Moot, Isayev, Tropsha, Cahoon, Materials Discovery, 2016, 6, 9-16
PbTiO$_3$ was identified as very similar to NiO AND It is has a dielectric constant >100


Design of alternate photocathodes

A materials informatics approach

PbTiO$_3$ was identified as very similar to NiO

AND

It is has a dielectric constant >100

Moot, Isayev, Tropsha, Cahoon, Materials Discovery, 2016, 6, 9-16
PbTiO$_3$ was identified as very similar to NiO in terms of electronic properties despite different crystal structures.
PbTiO3 is identified as a new photocathode material.

Successful experimental validation

Record fill factors of >50

First fully aqueous DSSC device

Currently, device performance is low; possible improvement by designing a new dye
The eternal philosophical question:
Which came first?

R or R²?
The eternal question: Which came first?

In the beginning was the Word…

And the Word was… **embedded**

(freely adopted from the Gospel of John)

“You should know a word by the company it keeps”
J.R.Firth 1957

British linguist; formulated the notion of the “context-dependent nature of meaning”
Learning semantic context with Word2Vec

Source Text

The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog.

The quick brown fox jumps over the lazy dog.

Training Samples

C=# words of context

Can be used to learn:

**CBOW:**
• \( Pr(\text{word}_k|\text{words}_\text{context}) \)

**Skip-Gram:**
• \( Pr(\text{words}_\text{context}|\text{word}_k) \)


Word2Vec Images courtesy of Chris McCormick:
Word embedding and similarity in the semantic space
Aspirin, also known as acetylsalicylic acid, is a medication used to treat pain, fever, and inflammation.

Penicillic acid is a mycotoxin that is produced by Aspergillus flavus and Penicillium roqueforti mold.

SMILES are words that uniquely describe sentence-molecules!
ReLeaSE* design principles: learning and exploiting structural linguistics of SMILES notation

- SMILES notations reflect rules of Chemistry
- SMILES notation embeds linguistic rules
- Neural nets could learn both of the above types of rules
- This knowledge can be transformed into the generation of new SMILES corresponding to novel chemically feasible molecules (generative model)
- One can build QSAR models based solely on SMILES notation (predictive model)
- QSAR models can be used as a reward function for reinforcement learning to bias the design of novel libraries

Design of the ReLeaSE* method
(Reinforcement Learning for Structural Evolution)

Elements of the thought cycle (molecules->models->molecules):

- Generate chemically feasible SMILES
- Develop SMILES-based QSAR model
- Employ QSAR model to bias library generation
- Produce new SMILES

Generative model: training mode

1.5M molecules from ChEMBL

RNN: \( \{W_1^0, \ldots, W_2^0\} \) → c1ccc(O)cc1

Did the training converge?

YES

RNN: \( \{W_1^*, \ldots, W_2^*\} \)

NO

\[ \langle \text{START} \rangle c1ccc(0)ccl\langle \text{END} \rangle \]

\[ c1cc)(F)ccl\langle \text{END} \rangle \]

RNN: \( \{W_1, \ldots, W_2\} \)

Softmax loss

+ loss
Generative model: training mode

- Training continues until convergence
- Every SMILES from ChEMBL is used as training example ~ 3-5 times
Are we making legitimate Smiles?

PubChem/ChEMBL

Smiles strings

AI learning system

ChemAxon

95% Valid Chemically-feasible molecules
Smile-ification of QSAR!

Quantitative Smiles – Activity Relationships

<table>
<thead>
<tr>
<th>Compounds</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>O=C(C)Oc1cccc1C(=O)O</td>
<td>0.531</td>
</tr>
<tr>
<td>CCOc1cc(C)ccc1OCC=CF</td>
<td>1.299</td>
</tr>
<tr>
<td>COc1cccccc1OCCO</td>
<td>0.946</td>
</tr>
<tr>
<td>CC(N)Sc1ccc(Cl)nc1</td>
<td>-0.218</td>
</tr>
<tr>
<td>COC(=O)NCc1cccccc1Cl</td>
<td>0.017</td>
</tr>
</tbody>
</table>
QSAR modeling using Smiles strings only*

Property prediction

Neural Network

\[
\text{CN2C(=O)N(C)C(=O)C1=C2N=CN1C}
\]

*LogP data for ~16K molecules from PHYSPROP (srcinc.com), Toxcast Dashboard (https://comptox.epa.gov/dashboard), and others.

5CV RF model with DRAGON7 Descriptors
5CV NN model with SMILES directly

\[
\begin{align*}
\text{RMSE: } & 0.57 & 0.53 \\
\text{MAE: } & 0.37 & 0.35 \\
R^2_{\text{ext}}: & 0.90 & 0.91
\end{align*}
\]
Reinforcement learning for chemical design

Generative model

RNN:\n\{W_1, \ldots, W_n\}

Predictive model

Property

RNN = \{\sigma, W_h, W_x\}

c1cccccc1

Fc1ccc2c(Nc3ccc(F)c(F)c3)ncnc2c1
Reinforcement learning for chemical design

Generative model

$0c(cc1cc2)ccc1cc2N$

$RNN: \{W_1, \ldots, W_n\}$

$<\text{START}>$

Predictive model

Property

$RNN = \{\sigma, W_h, W_x\}$

c1ccccc1

$Fc1ccc2c(Nc3ccc(F)c(F)c3)ncnc2c1$
Reinforcement learning for chemical design

Generative model

\[ 0c(cc1cc2)ccc1cc2N \]

\[ RNN: \{W_1, ..., W_n\} \]

\(<\text{START}>\)

Predictive model

Property

\[ RNN = \{\sigma, W_h, W_x\} \]

clcccccl

\(\text{ACTIVE!}\)
Reinforcement learning for chemical design

Generative model

Predictive model

RNN: \( \{W_1, ..., W_n\} \)

\( <\text{START}> \)

Property

RNN = \( \{\sigma, W_h, W_x\} \)

c1ccccc1
Reinforcement learning for chemical design
Reinforcement learning for chemical design

Generative model

Predictive model

\[
RNN = \{\sigma, W_h, W_x\}
\]
Reinforcement learning for chemical design

Generative model

RNN:
\{W_1, ..., W_n\}

Predictive model

Property

RNN = \{\sigma, W_h, W_x\}

FC(F) COc1ccc2c(Nc3ccc(Cl)c(Cl)c3)nncnc2c1

0c(cc1cc2)ccc1cc2N

<START>
Reinforcement learning for chemical design
Reinforcement learning for chemical design

Generative model

Predictive model

INACTIVE!
Reinforcement learning for chemical design
Reinforcement learning for chemical design
Reinforcement learning for chemical design
Technical details

- Models were trained on Nvidia Titan X and Titan V GPUs
- Training the generative model on ChEMBL took ~ 25 days
- Training of predictive models took ~ 2 hours
- Biasing the generative model with reinforcement learning for one property ~ 1 day
- Generative model produces 1000 compounds per minute
Results: Synthetic accessibility score* of the designed libraries

PoC: Structural Bias

A: increase in number of substituents

B: increase in number of benzene rings

Reward increase
Results: Biasing target properties in the designed libraries

- Minimized
- Maximized
- Baseline

Melting temperature ($T_m$), °C

Number of substituents

*
Results: Biasing target properties in the designed libraries

Minimized
Maximized
Baseline

Optimized
Baseline

JAK2 Inhibition (pIC50)
Partition coefficient (logP)
Target predictions for generated compounds using SEA*

<table>
<thead>
<tr>
<th>Query</th>
<th>Target Key</th>
<th>Target Name</th>
<th>Description</th>
<th>P-Value</th>
<th>MaxTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPM_HUMAN+5</td>
<td>NPM1</td>
<td>Nucleophosmin</td>
<td></td>
<td>3.118e-74</td>
<td>0.49</td>
</tr>
<tr>
<td>CCNH_HUMAN+5</td>
<td>CCNH</td>
<td>Cyclin-H</td>
<td></td>
<td>2.571e-32</td>
<td>0.38</td>
</tr>
<tr>
<td>PAK1_HUMAN+5</td>
<td>PAK1</td>
<td>Serine/threonine-protein kinase PAK 1</td>
<td></td>
<td>5.277e-24</td>
<td>0.39</td>
</tr>
<tr>
<td>ALK_HUMAN+5</td>
<td>ALK</td>
<td>ALK tyrosine kinase receptor</td>
<td></td>
<td>3.714e-23</td>
<td>0.54</td>
</tr>
<tr>
<td>JAK2_HUMAN+5</td>
<td>JAK2</td>
<td>Tyrosine-protein kinase JAK2</td>
<td></td>
<td>1.136e-21</td>
<td>0.61</td>
</tr>
<tr>
<td>INSR_HUMAN+5</td>
<td>INSR</td>
<td>Insulin receptor</td>
<td></td>
<td>2.36e-17</td>
<td>0.54</td>
</tr>
<tr>
<td>CCNB1_HUMAN+5</td>
<td>CCNB1</td>
<td>G2/mitotic-specific cyclin-B1</td>
<td></td>
<td>2.22e-16</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Target predictions for generated compounds using SEA*

<table>
<thead>
<tr>
<th>Query</th>
<th>Target Key</th>
<th>Target Name</th>
<th>Description</th>
<th>P-Value</th>
<th>MaxTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR_HUMAN+5</td>
<td>EGFR</td>
<td></td>
<td>Epidermal growth factor receptor</td>
<td>8.688e-244</td>
<td>0.61</td>
</tr>
<tr>
<td>ERBB2_HUMAN+5</td>
<td>ERBB2</td>
<td></td>
<td>Receptor tyrosine-protein kinase erbB-2</td>
<td>8.544e-169</td>
<td>0.55</td>
</tr>
<tr>
<td>ERBB2_RAT+5</td>
<td>Erbb2</td>
<td></td>
<td>Receptor tyrosine-protein kinase erbB-2</td>
<td>5.893e-87</td>
<td>0.42</td>
</tr>
<tr>
<td>VGFR2_HUMAN+5</td>
<td>KDR</td>
<td></td>
<td>Vascular endothelial growth factor receptor 2</td>
<td>6.294e-65</td>
<td>0.58</td>
</tr>
<tr>
<td>ERBB4_HUMAN+5</td>
<td>ERBB4</td>
<td></td>
<td>Receptor tyrosine-protein kinase erbB-4</td>
<td>1.354e-64</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Results: analysis of similarity

Distribution of Tanimoto similarity to the nearest neighbor in training dataset for compounds predicted to be active for EGFR by consensus of QSAR models:
Model visualization for putative JAK2 inhibitors (projection using t-SNE)

ZINC19982368  
\( pIC_{50} = 8.64 \)

ZINC4699992  
\( pIC_{50} = 8.23 \)

ZINC66347860  
\( pIC_{50} = 3.31 \)

ZINC2876515  
\( pIC_{50} = 8.39 \)

ZINC3549031  
\( pIC_{50} = 3.76 \)

ZINC19982368  
\( pIC_{50} = 3.76 \)

ZINC2876515  
\( pIC_{50} = 0.63 \)

ZINC3549031  
\( pIC_{50} = 10.37 \)
Summary

• AI methods coupled with SMILES representation (only!) afford biased library generation
• The system naturally embeds reinforcement learning to produce novel structures with the desired property
• The system can be tuned to bias libraries towards specific property ranges
• Next phase is experimental validation of hits
# Summary of recent AI-based studies on chemical library design

<table>
<thead>
<tr>
<th>Molecular representations</th>
<th>Generative models</th>
<th>Method of biasing generated compounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Fingerprints</td>
<td>• Autoencoders</td>
<td>• None</td>
</tr>
<tr>
<td>• SMILES</td>
<td>• Generative adversarial models</td>
<td>• Latent space optimization</td>
</tr>
<tr>
<td>• Graphs</td>
<td>• Recurrent neural networks</td>
<td>• Fine-tuning on small subset of molecules with the desired property</td>
</tr>
<tr>
<td></td>
<td>• Convolutional neural networks</td>
<td>• Reinforcement Learning</td>
</tr>
</tbody>
</table>

MML UNC.EDU
An example of experimental validation of AI-based models*

- First training on large dataset
- Then fine-tuning on small subset of active compounds
- “These observations corroborate the ability of the generative AI model to produce novel chemical entities within the training data domain”.

---

**Table 1. In vitro activity of designs 1–5 on RXRs and PPARs (EC$_{50}$ values ± SEM [µM]; n = 2 (when inactive) or 4 (when active) independent experiments in duplicates; inactive, no statistically significant reporter transactivation at a compound concentration of 30 µM).**

<table>
<thead>
<tr>
<th>Compound no.</th>
<th>RXRα</th>
<th>RXRβ</th>
<th>RXRγ</th>
<th>PPARα</th>
<th>PPARγ</th>
<th>PPARδ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.13 ± 0.01</td>
<td>1.1 ± 0.3</td>
<td>0.06 ± 0.02</td>
<td>inactive</td>
<td>2.3 ± 0.2</td>
<td>inactive</td>
</tr>
<tr>
<td>2</td>
<td>13.0 ± 0.1</td>
<td>9 ± 2</td>
<td>8.0 ± 0.7</td>
<td>inactive</td>
<td>2.8 ± 0.3</td>
<td>inactive</td>
</tr>
<tr>
<td>3</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>10.1 ± 0.3</td>
<td>inactive</td>
</tr>
<tr>
<td>4</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>9 ± 3</td>
<td>14 ± 2</td>
</tr>
<tr>
<td>5</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
<td>inactive</td>
</tr>
<tr>
<td>reference agonists$^a$</td>
<td>0.033 ± 0.002</td>
<td>0.024 ± 0.004</td>
<td>0.025 ± 0.002</td>
<td>0.006 ± 0.002</td>
<td>0.6 ± 0.1</td>
<td>0.5 ± 0.1</td>
</tr>
</tbody>
</table>

$^a$ Reference agonists, literature data: bexarotene$^{17}$ for RXRs, GW7647$^{18}$ for PPARα, pioglitazone$^{19}$ for PPARγ, L165,041$^{19}$ for PPARδ

Many virtues of Cheminformatics

- Drug discovery
- Toxicity prediction
- Nano-technology
- Student Performance
- Text and Social Media mining
- Patient Outcomes
- Materials Science
- Drug Delivery
Acknowledgements

Principal Investigator
Alexander Tropsha

Research Professors
Alexander Golbraikh
Olexander Isayev
Eugene Muratov

Postdoctoral Fellows
Vinicius Alves
Stephen Capuzzi
Joyce Borba

Graduate students
Sherif Faraq
Kyle Bowers
Maria Popova
Andrew Thieme

• Duke University
  – Stefano Curtarolo
  – Corey Oses

• UNC Chemistry
  – Jim Cahoon
  – Taylor Moot
  – Aaron Taggart

MAJOR FUNDING

NIH
- 1U01CA207160
- R01-GM114015
- 5U54CA198999
- 1OT3TR002020

ONR
- N00014-16-1-2311