Materials Discovery with Artificial Intelligence

2019 03. 21.

Hyo Sug Lee & Youn-Suk Choi
Contents

- Trend in Computer-Aided Materials Discovery
- High-Throughput Computational Screening & Exhaustive Enumeration
- Deep-Learning-based Evolutionary Design
- Deep-Learning-based Inverse Design
- Efficacy of Computer-Aided Materials Discovery
Trend in Computer-Aided Materials Discovery

- For accelerated materials discovery

1. Conventional
   - Trial-and-Error (high cost)
   - Iterative experiments

2. 1st Gen.
   - Simulation (low throughput)
   - Pre-validation

3. 2nd Gen.
   - Virtual screening (low hit-rate)
   - High throughput

4. 3rd Gen.
   - Targeted design (high hit-rate)
   - Right solutions with minimum effort

- Rationalization
- Efficiency
- Intelligence
Trend in Computer-Aided Materials Discovery

- Prediction of materials property based on machine learning
  - Build-up of *Materials vs. Property DB* → *Materials Informatics*

- Cheminformatics
- Introduction stage of machine learning
- Development stage

**QSAR***(62, Hansch&Fujita)***

**ANN****(71, UC Irvine)**

**Graph Kernels***(05 @ UC Irvine)**

**Bayesian Modeling***(09 @ MIT)**

**SMILES***(87, Weininger)**

**Kernel methods**

**Deep Learning** *(16 @ Stanford)*

**(18 @ Harvard)**

*QSAR: Quantitative Structure-Activity Relationship
**ANN: Artificial Neural Network
***SMILES: Simplified Molecular-input-line Systems

**SMILES**: CC(C)NCC(O)COC1=CC(CC2=CC=CC=C2)=C(CC(N)=O)C=C1

**Fingerprint**: 011100011111101010010100100000101010001001010...

**Descriptor Vector**

**SMILES**: CC(C)NCC(O)COC1=CC(CC2=CC=CC=C2)=C(CC(N)=O)C=C1

**Fingerprint**: 011100011111101010010100100000101010001001010...

**Descriptor** → **Training**

**Analysis**

**Deep Thought's Answer is 42**
Trend in Computer-Aided Materials Discovery

- Materials design based on machine learning
  - Inverse QSAR $\rightarrow$ Inverse Design

- Inverse QSAR (Late 80’s–)
- Exhaustive Generation (’12 @ Tokyo)
- Inverse Design (’16 @ SAIT)
- Genetic Algorithms (’92 @ Purdue)
- SMILES Autoencoder (’16 @ Harvard)
- GAN’ for molecules (’17 @ Harvard)

Focus on autonomous molecular generation

*GAN: Generative Adversarial Network
Trend in Computer-Aided Materials Discovery

- In-silico technologies for materials discovery
High-Throughput Computational Screening & Exhaustive Enumeration

“Landscape of phosphorescent light-emitting energies of homoleptic Ir(III)-complexes predicted by a graph-based enumeration and deep learning”, GI01.02.02, 2018 MRS fall meeting
High-Throughput Computational Screening

- Property prediction with high-performance computing for large-scale exploration of materials candidates
High-Throughput Computational Screening

- ML (Machine Learning)-assisted HTCS for higher efficiency

Seed Fragments → Combination → Candidate Pool

(1) Simulation + ML

(2) Prioritizing calculation based on active learning

Database → Verification → Target Materials

large amounts of candidates
High-Throughput Computational Screening

- Exhaustive enumeration based on graph-theory
  - “Graphs”
    - Mathematical structures used to model pairwise relations between objects.
    - Made up of nodes and edges.
    - In chemistry, graph is used to model molecules, where nodes represent atoms and edges represent bonds.

※ Exhaustive enumeration:
Systematical enumeration of all possible molecules for optimal solution search
High-Throughput Computational Screening

- Complete list of non-isomorphic graphs

![Graphs with IDs and edge counts](http://www.cadaec.net/graphpics.htm)
High-Throughput Computational Screening

- Landscape of phosphorescent light-emitting energies of homoleptic Ir(III)-complex core structures
  - Ir(III)-complexes
    - Widely used as phosphorescent OLED dopants.
    - Figuring out the full landscape of emission color is important for discovering high-performing molecules in target color regions.

Organic Electronics, 63, 244–249 (2018)
High-Throughput Computational Screening

- **Approach**
  - Consider the nodes in graph as rings and edges as ring-connections.
  - Limited the total number rings between 3 and 5.
  - Exclude non-planar type (5-21) and invalid structures as dopant.
  
  \[ \rightarrow \text{Only 11 graphs are valid among the total 29 graphs}. \]
High-Throughput Computational Screening

- **Enumeration**
  - For 5- and 6-membered rings.
  - Substitute some carbons of each molecule with nitrogen atoms (max. five).
    - Total 9,919,469 (~10M) core structures

1. **Graphs**
   - 3-1
   - 4-2, 4-3, 4-5
   - 5-2, 5-3, 5-5, 5-6
   - 5-9, 5-15, 5-16

2. **Skeletons**
   - total 405 EA

3. **Set Iridium positions**

4. **Substitute some carbon atoms with nitrogen atoms**
High-Throughput Computational Screening

- **Property prediction**
  - Trained a deep-neural-network model with simulated $T_1$ data
    - Input: ECFP (Extended Connectivity FingerPrints) of molecular structures
    - Outputs: $T_1$ energy (phosphorescent light-emitting wavelength)

With 80k training data, the average prediction error was less than 0.1 eV

$$\frac{80k}{10M} = 0.8\%$$

By simulating the properties of only 0.8% molecules, we can fully scan the chemical space of 10M!
High-Throughput Computational Screening

- **Results**
  - Distribution of $T_1$ values
  - Blue-color emitting materials are rare compared with red and green

![Graph showing distribution of $T_1$ values with blue, green, and red bars representing different percentages. Blue (0.4%), Green (4.3%), Red (18.4%)].
Conclusions

- In materials discovery, deep-learning-based HTCS is a good alternative to conventional trial-and-error type approach.
- Moreover, exhaustive enumeration makes it possible to systematically explore the whole chemical space.
- With the proposed exhaustive enumeration method based on graph theory and deep learning, the whole landscape of 10M phosphorescent Ir-dopants could be scanned with just 0.8% computational cost compared with the pure simulation-based approach.
Deep-Learning-based Evolutionary Design

“Evolutionary design of organic molecules based on deep learning and genetic algorithm”, COMP, ACS fall 2018 National Meeting
Evolutionary Design

- A generic population-based metaheuristic optimization technique
- Uses bio-inspired operators to reach near-optimal solutions; mutation, crossover, and selection in case of genetic algorithm

[Diagram showing the process of evolutionary design with steps such as initial population, calculate fitness, selection, mutation, crossover, and new population.]

[Image of a fitness landscape with a graph showing average fitness over generations.]
Deep-Learning-Based Evolutionary Design

### Proposed approach

<table>
<thead>
<tr>
<th></th>
<th>Conventional</th>
<th>Proposed</th>
<th>Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular Descriptor</td>
<td>Graph or ASCII string</td>
<td>Bit string (ECFP)</td>
<td>• Prevent heuristic bias</td>
</tr>
<tr>
<td>Molecular Evolution</td>
<td>Heuristic</td>
<td>Random</td>
<td>• Secure chemical validity</td>
</tr>
<tr>
<td>Fitness Evaluation</td>
<td>Simple assessment</td>
<td>DNN</td>
<td>• Versatile evaluation is possible</td>
</tr>
</tbody>
</table>

*ECFP (Extended Connectivity FingerPrint)*

DNN (Deep Neural Network), RNN (Recurrent Neural Network)

SMILES (Simplified Molecular-Input Line-Entry System)
Deep Learning-Based Evolutionary Design

- Deep learning models
  - [DNN] 3 hidden layers, 500 hidden units in each layer
  - [RNN] 3 hidden layers, 500 long short-term memory units

\[
y_1 = 'CCC' \\
y_2 = 'CCC' \\
y_3 = 'CC(' \\
\ldots \\
y_t = ')' = O' \\
\]

\[
y = ('CCC', 'CCC', 'CC(', \ldots, \ldots)' = O') \rightarrow 'CCCC(N)=O'
\]

*ECFP (dimension=5,000, neighbor size=6)
Deep Learning-Based Evolutionary Design

- Validation test
  - Design target: change the $S_1$ (light-absorbing wavelength) of seed molecules
  - Training data: M.W. 200~600 g/mol from PubChem (10,000~50,000 molecules)

<table>
<thead>
<tr>
<th>No. of training data</th>
<th>Prediction accuracy of DNN*1 (R, MAE)</th>
<th>Success rate of decoding*2 (RNN)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_1$</td>
<td>HOMO</td>
</tr>
<tr>
<td>① 50,000</td>
<td>0.973, 0.198</td>
<td>0.945, 0.172</td>
</tr>
<tr>
<td>② 30,000</td>
<td>0.930, 0.228</td>
<td>0.934, 0.191</td>
</tr>
<tr>
<td>③ 10,000</td>
<td>0.913, 0.278</td>
<td>0.885, 0.244</td>
</tr>
</tbody>
</table>

※1. No. of test data=No. of training data/10
※2. Chemical validity is evaluated with RDKit, No. of test data=5,000
Deep Learning-Based Evolutionary Design

- Evolution toward the increase and decrease of $S_1$ (eV)
  - Seed: randomly selected 50 molecules (3.8<$S_1<$4.2)
  - Number of training data = 10k, 30k, 50k

![Diagram showing the change in $S_1$ over generations with different training data sizes.](image)

- $S_1$ distribution in the training data (50k)
  - 4.0 eV
Deep Learning-Based Evolutionary Design

- Evolution under the constraint of HOMO and LUMO (eV)
  - Seed: randomly selected 50 molecules (3.8 < S₁ < 4.2)
  - Number of training data = 50k
  - Constraint: -7.0 < HOMO < -5.0, LUMO < 0.0

<table>
<thead>
<tr>
<th></th>
<th>Increase of S₁</th>
<th>Decrease of S₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>LUMO</td>
<td>①</td>
<td>③</td>
</tr>
<tr>
<td></td>
<td>②</td>
<td>④</td>
</tr>
<tr>
<td>HOMO</td>
<td>①</td>
<td>③</td>
</tr>
<tr>
<td></td>
<td>②</td>
<td>④</td>
</tr>
</tbody>
</table>

HOMO & LUMO distributions in the training data (50k)
Deep Learning-Based Evolutionary Design

- Examples of evolved molecules (No. of training data = 50k)

- Constraint (eV)
  - -7.0 < HOMO < -5.0
  - LUMO < 0.0
Conclusions

- A fully data-driven evolutionary molecular design based on deep-learning models (DNN & RNN) was proposed and automatically evolved seed molecules toward target without any pre-defined chemical rules.

- Unlike HTCS, the closed-loop evolutionary workflow guided by deep-learning automatically derived target molecules and found rational design paths by elucidating the relationship between structural features and their effect on the molecular properties.
Deep-Learning-based Inverse Design

npj Comput. Mater., 4, 67, 2018
Deep-Learning-Based Inverse Design

- Paradigm shift of ML in computer-aided materials discovery

<table>
<thead>
<tr>
<th>Passive Role</th>
<th>Active Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficient screening based on property prediction</td>
<td>Propose candidates via automated design</td>
</tr>
<tr>
<td>Highly depends on explicit knowledge of chemists</td>
<td>Provides implicit knowledge from data</td>
</tr>
</tbody>
</table>

Candidate pool → Predict materials properties → Screening → Target Properties

Generating candidate pool via database

Artificial Intelligence for Materials Design

Propose target materials
Deep-Learning-Based Inverse Design

- Implementation of inverse-design model

- **Input**
- **encoder**
- **decoder**
- **Output**

Hidden Factor (fixed-length vector)

Molecular descriptor (\(x\); ECFP format)

- **e(·)**: encoding function
- **f(·)**: property prediction function
- **d(·)**: decoding function
- **z**: encoded vector of molecular descriptor

Molecular property (\(t\))

- **f(z)**

Molecular structure identifier (\(y\); SMILES format)

Hidden Factor (fixed-length vector)

- **z = e(x)**

DNN

RNN
Deep-Learning-Based Inverse Design

- Inverse design of light-absorbing organic molecules (1/2)
  - Training DB
    - 50k molecules sampled from PubChem (M.W. 200~600)
    - DFT calculations for $S_1$

**Distribution of $\lambda_{\text{max}}$ of the inverse-designed molecules**

*Simulation values for the 500 molecules in each target

<table>
<thead>
<tr>
<th>Target $\lambda_{\text{max}}$</th>
<th>Hit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>200–300 nm</td>
<td>82.6%</td>
</tr>
<tr>
<td>300–400 nm</td>
<td>64.8%</td>
</tr>
<tr>
<td>400–500 nm</td>
<td>45.6%</td>
</tr>
</tbody>
</table>

※ About 10% of the designed molecules were found in PubChem even though those were not included in the randomly selected training library.
Deep-Learning-Based Inverse Design

- Inverse design of light-absorbing organic molecules (2/2)

a. Antraquinone derivative ($\lambda_{\text{max}}=433.4$ nm)

b. Azobenzene derivative ($\lambda_{\text{max}}=527.5$ nm)

c. Isoidoline derivative ($\lambda_{\text{max}}=434.4$ nm)

d. Squaraine derivative ($\lambda_{\text{max}}=503.5$ nm)

Examples of inverse-designed molecules which share the moieties with well-known dye materials
Deep-Learning-Based Inverse Design

- Inverse design of hosts for blue phosphorescent OLED (1/3)
  - Target: $T_1 \geq 3.00$ eV
  - Training DB
    - In-house library of 6,000 molecules by combinatorial enumeration (with nine linker (L) and fifty-seven terminal fragments (R) which are frequently employed in OLED hosts; symmetric R-L-R & R-R type enumeration).
    - Property labeling with DFT calculations.

The distribution of simulated $T_1$ (eV) energy levels for the generated 3,205 molecules
- a. mean=2.94, std=0.15
- b. mean=3.02, std=0.10
- c. mean=2.92, std=0.13

The fractions of the hosts that satisfied the target ($T_1 \geq 3.00$ eV)
- 36.2% for a
- 58.7% for b
- 26.9% for c (3,497 molecules)
Deep-Learning-Based Inverse Design

- Inverse design of hosts for blue phosphorescent OLED (2/3)

Examples of inverse-designed host materials

<table>
<thead>
<tr>
<th>Experiment (eV)</th>
<th>HOMO (eV)</th>
<th>LUMO (eV)</th>
<th>$S_1$ (eV)</th>
<th>$T_1$ (eV)</th>
<th>$\Delta E_{ST}$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>-5.98</td>
<td>-2.43</td>
<td>3.56</td>
<td>3.06</td>
<td>0.55</td>
</tr>
<tr>
<td>b1</td>
<td>-5.96</td>
<td>-2.14</td>
<td>3.64</td>
<td>2.93</td>
<td>1.01</td>
</tr>
<tr>
<td>c1</td>
<td>-6.07</td>
<td>-2.65</td>
<td>3.38</td>
<td>2.97</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Asymmetric molecules with the given fragments in the training library

Symmetric molecules where the new fragments were introduced

Asymmetric molecules where the new fragments were introduced
The connection rules of the inverse-designed molecules

L-(R₁,R₂,R₃) ★
(4)
R-L-R
(3,010)
R-L'R
(3,010)
R₁-L'R
(1,931)
R₁-L'asym-R₁
(636)
R₁-L'asym-R₂
(443)
R₁-L_sym -R₁
(403)
R₁-L_sym -R₂
(1,528)
R₁-Lasym-R₁
(1,079)
R₁-Lasym-R₂
(443)

Total host molecules
(3,205)

Linker
Terminal₁
Terminal₂
Terminal₃

L: Linker fragment
R: Terminal fragment
L_sym: Symmetric linker
L_asym: Asymmetric linker

Inverse design of hosts for blue phosphorescent OLED (3/3)
Conclusions

- A fully data-driven inverse design method successfully extracted the latent materials design rules and proposed target molecular structures without any external intervention.

- The inverse design model successfully proposed new candidates by modifying the assemble rules and creating new fragments.
Efficacy of Computer-Aided Materials Discovery

Simulation-based Screening

- HTCS for pre-defined chemical space

1st trial: 1M Candidates
  - QC simulations take 1.5 years
  - Fail to find the target structure

2nd trial: 1M Candidates
  - QC simulations take 1.5 years
  - Fail to find the target structure

3rd trial: 1M Candidates
  - QC simulations take 1.5 years
  - Succeed to find the right structure
  - Total TAT took 4.5 years

Inverse Design

[Step1] Building the training dataset
  - Needs only QC sim. for 50k molecules
  - (27 days)

[Step2] Deep learning model training with GPU (3 days)

[Step3] QC simulations for the proposed molecules (1 day)

Total TAT takes 1 month

more than 50X speed up (4.5 years vs. 1 month)

“The inverse design learns by itself the molecular design rules inherent in the libraries and can reduce the effort of researchers and total time to reach the goal”

- QC simulation tool: turbomole
  - Total computational resources=10,000 CPU
  - In case of 10 CPU computing per molecule, the simulation requires about 13 hrs.
Prospects for AI-based Materials Development

**Design**

- Target Properties
- Artificial Intelligence for Materials Design
- Propose target materials

**Simulation**

- Energy
- Electronic Properties
- Training Data: DFT simulation (<100 atoms)
- Neural Network (MD potential)
- Meso-scale simulation (~10^4 atoms)

**Analytical Chemistry**

- Database
- Queuing'system'

**Synthesis**

- Training
- Design
- Analysis
- Synthesis
Thank you (Q&A)

ysuk.choi@samsung.com