NVIDIA's GPU Technology Conference (GTC) 2019
San Jose, CA
Mar 18–21, 2019
[S9967]

Materials Discovery with Artificial Intelligence

2019 03. 21.

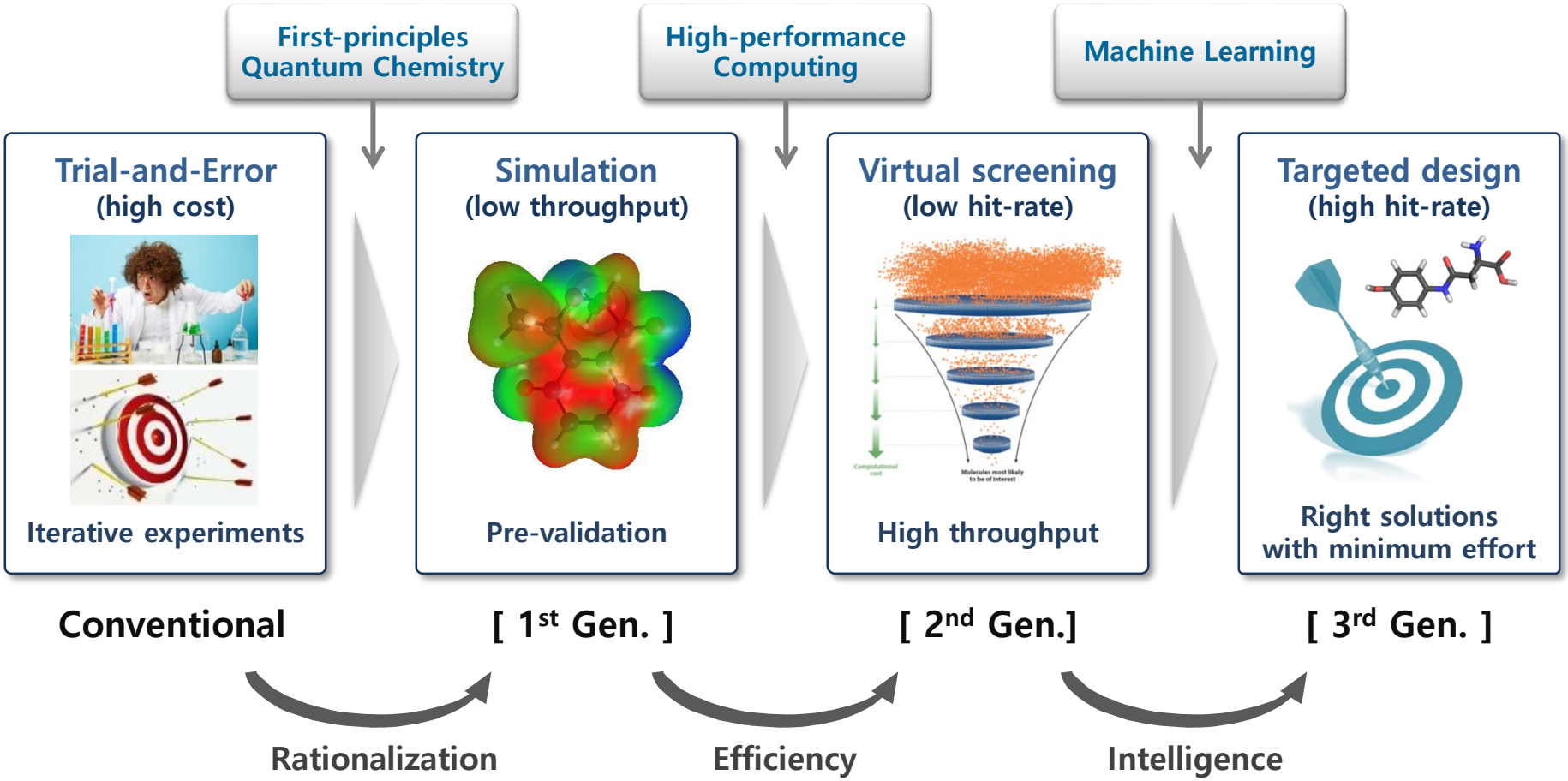
Hyo Sug Lee & Youn-Suk Choi

SAMSUNG ADVANCED INSTITUTE OF TECHNOLOGY

Contents

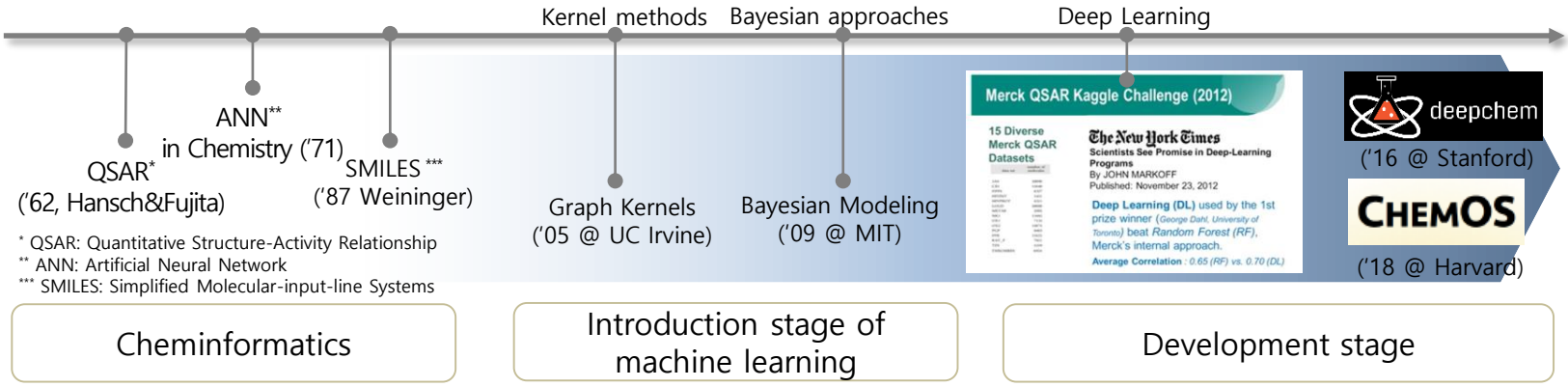
Trend in Computer-Aided Materials Discovery

- **For accelerated materials discovery**

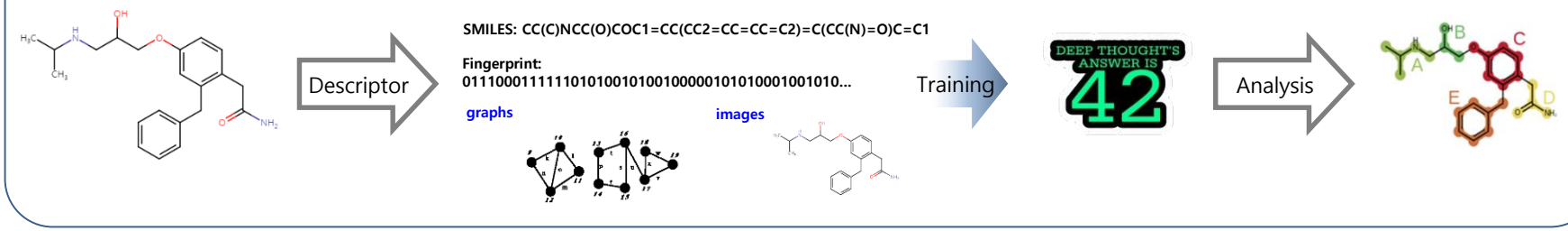


Trend in Computer-Aided Materials Discovery

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

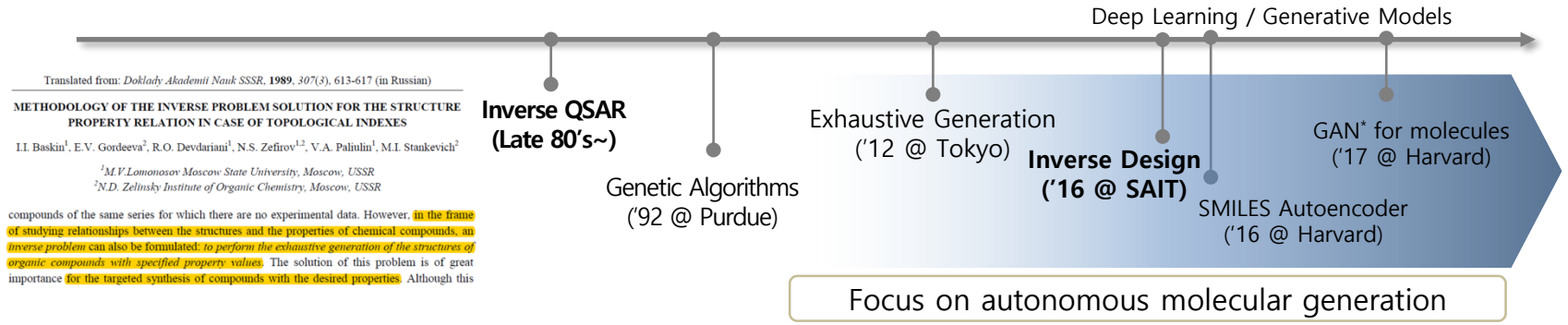


Process of Machine Learning @ Materials Research

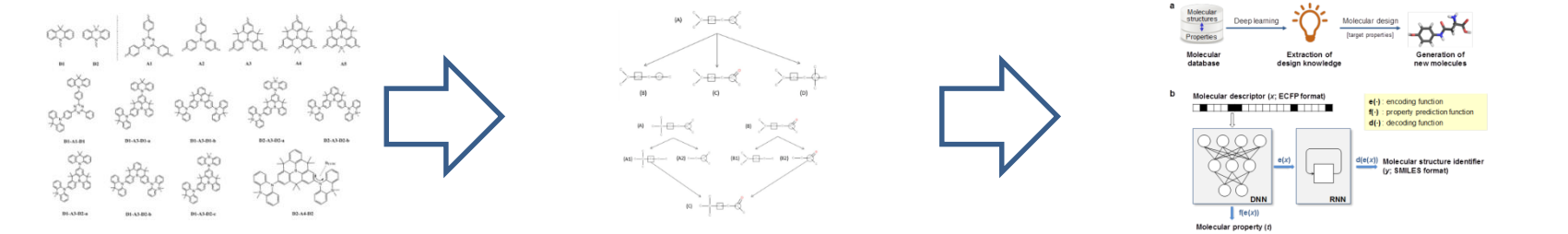


Trend in Computer-Aided Materials Discovery

- **Materials design based on machine learning**
  - *Inverse QSAR → Inverse Design*

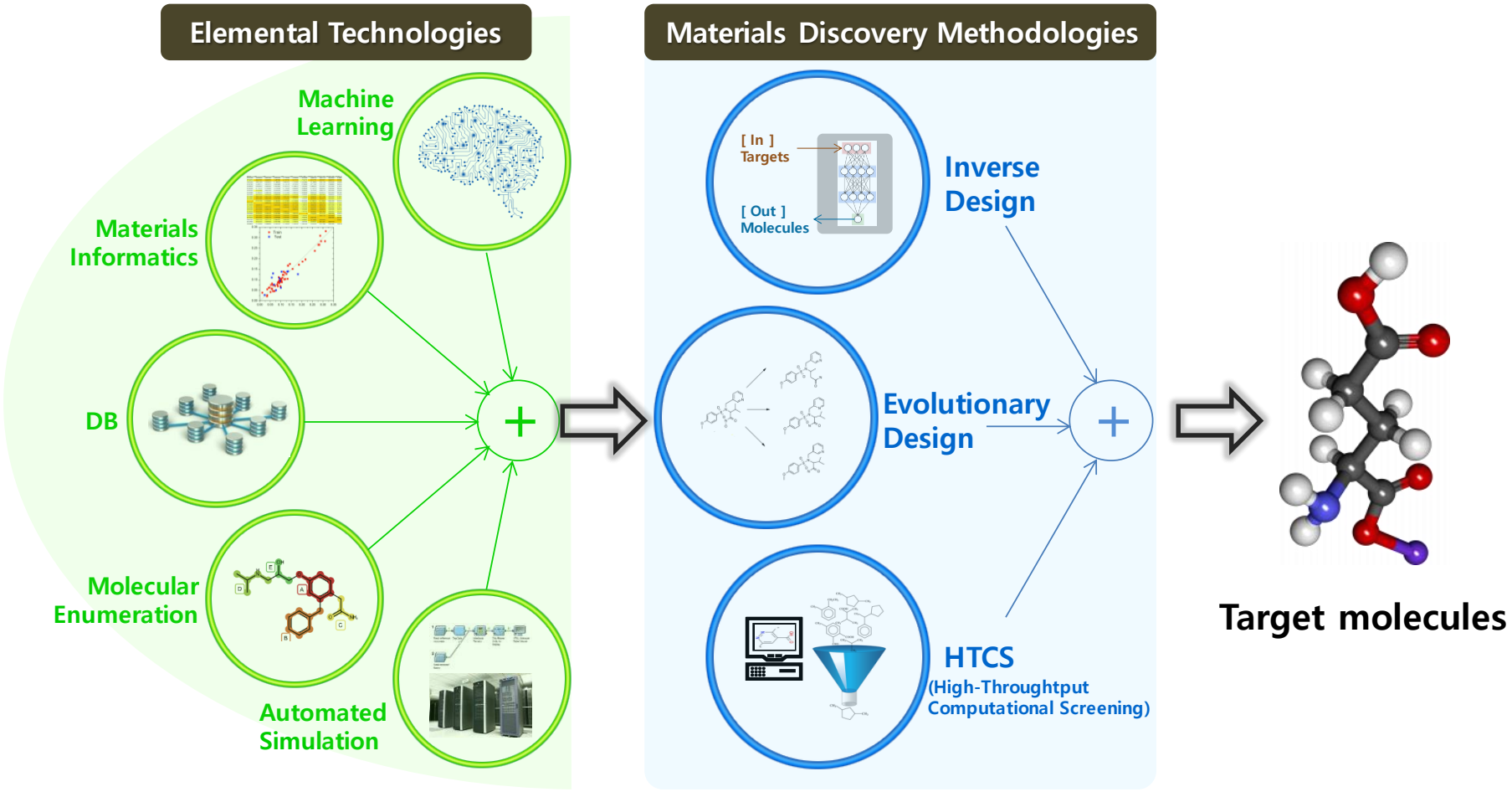


*GAN: Generative Adversarial Network



Trend in Computer-Aided Materials Discovery

- **In-silico technologies for materials discovery**

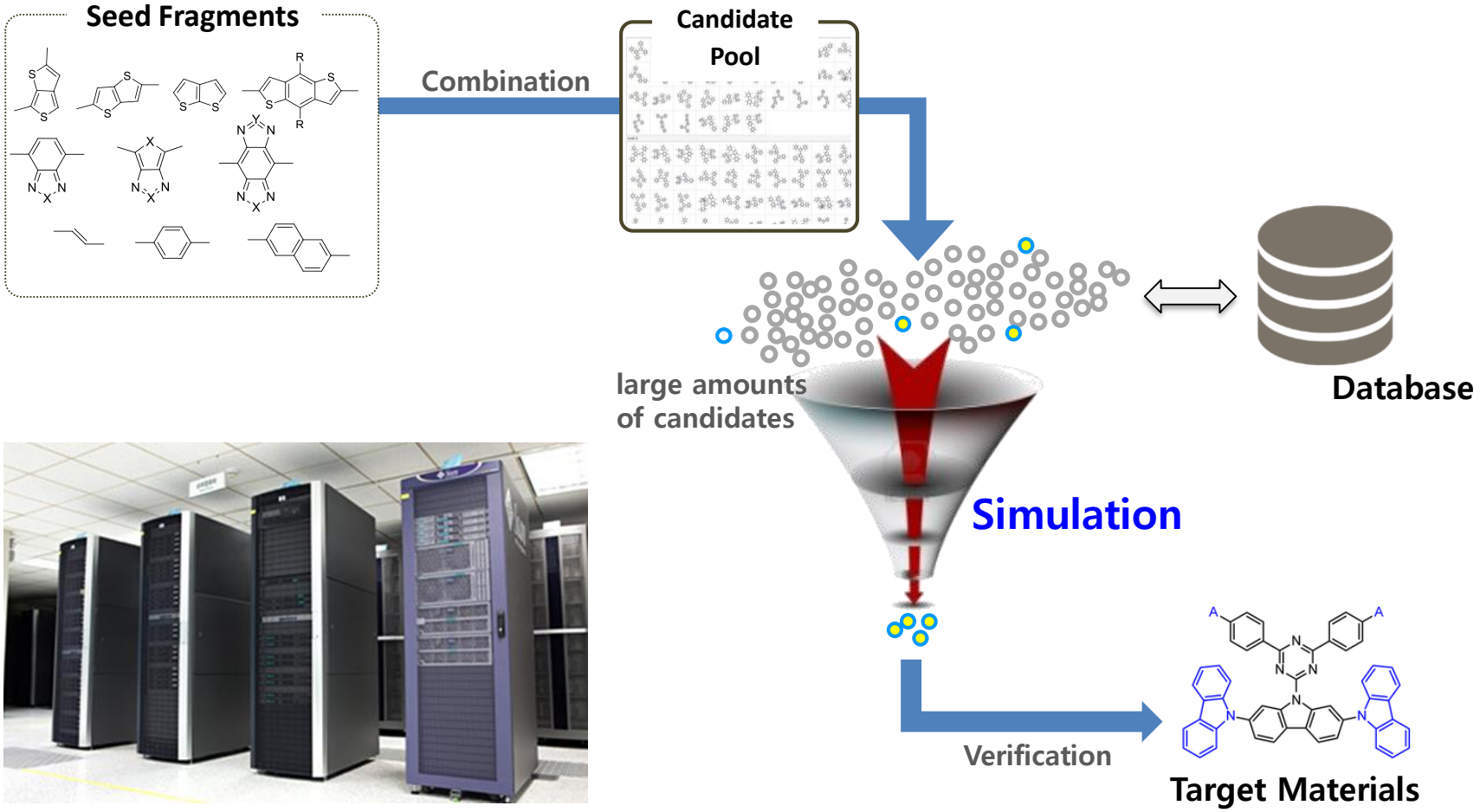


High-Throughput Computational Screening & Exhaustive Enumeration

"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*

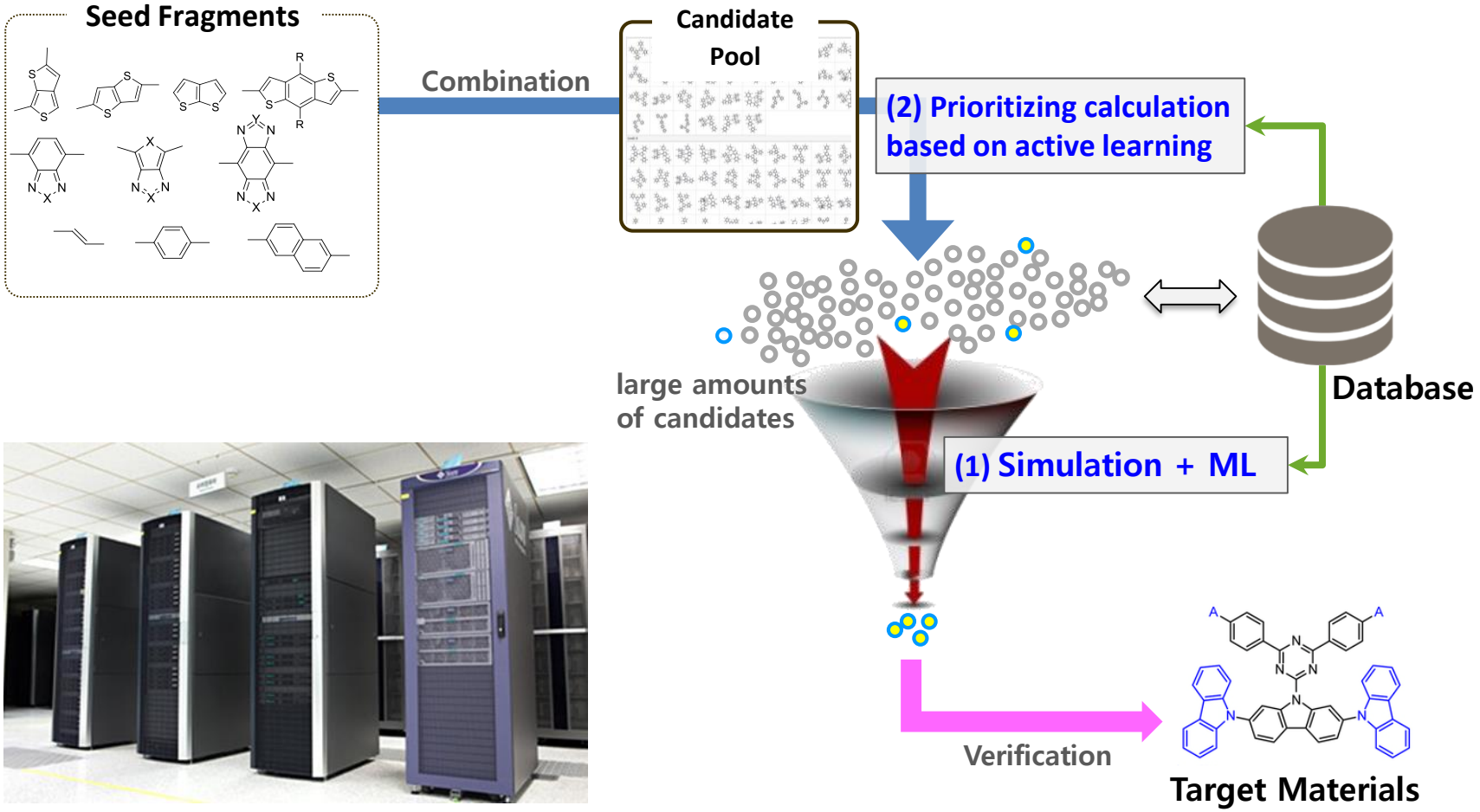
High-Throughput Computational Screening

- **Property prediction with high-performance computing for large-scale exploration of materials candidates**



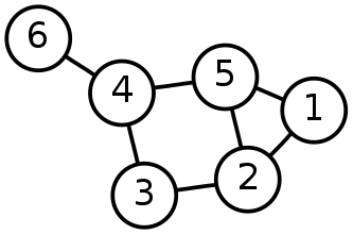
High-Throughput Computational Screening

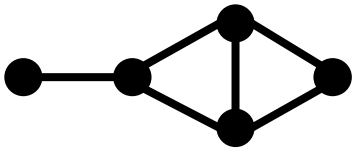
- ML (Machine Learning)-assisted HTCS for higher efficiency



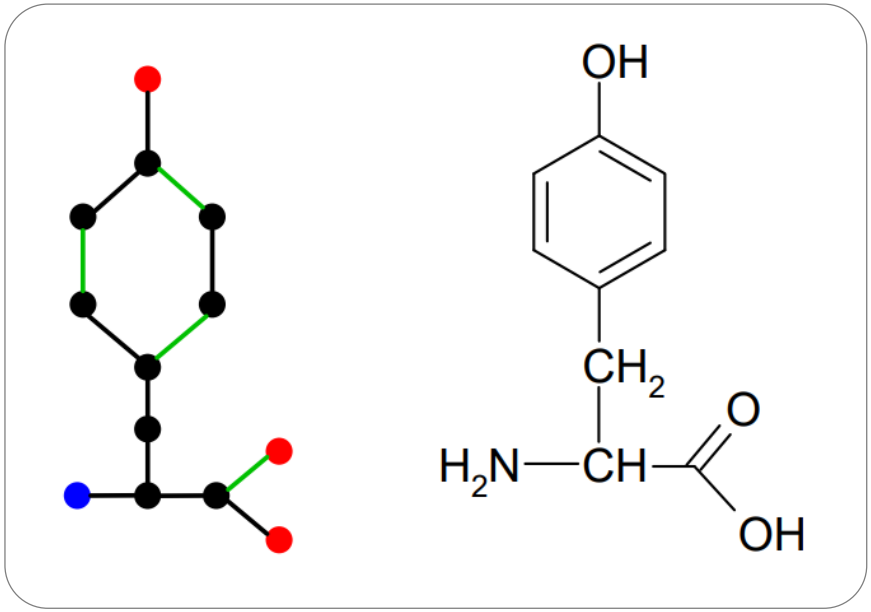
High-Throughput Computational Screening

- **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.



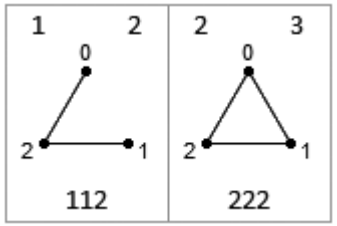


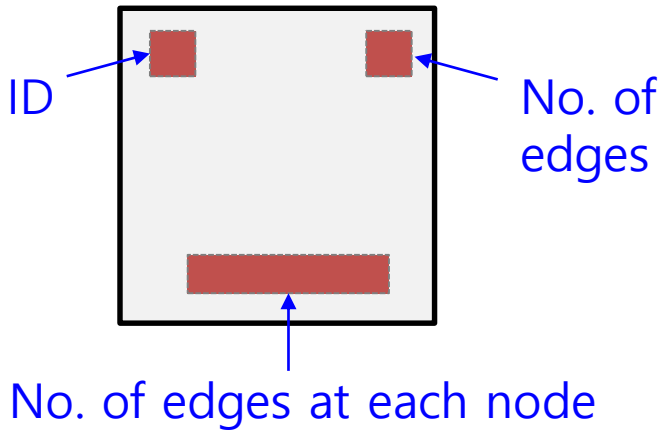
※ **Exhaustive enumeration**:
Systematical enumeration of all possible molecules for optimal solution search

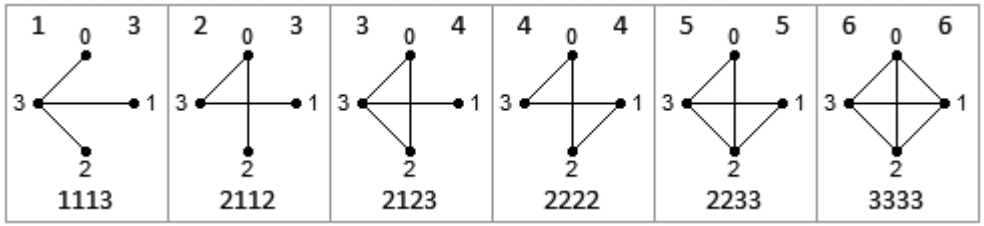


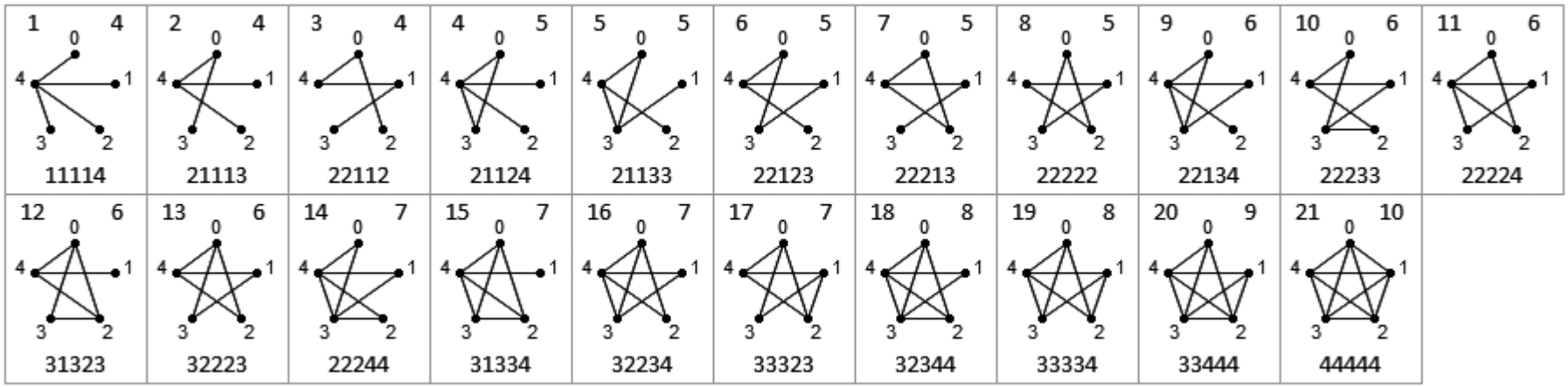
High-Throughput Computational Screening

- **Complete list of non-isomorphic graphs**





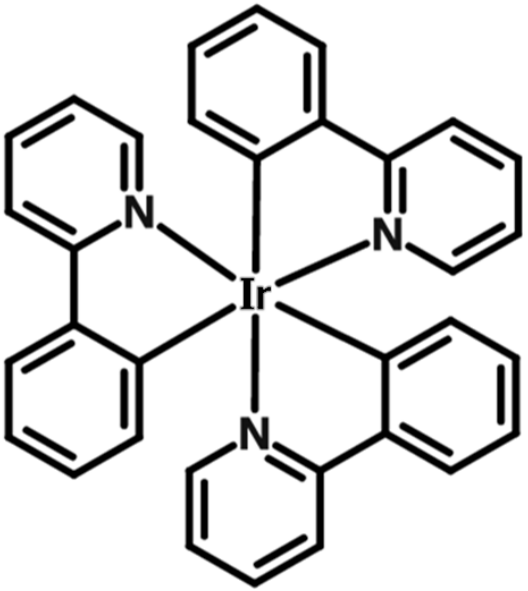


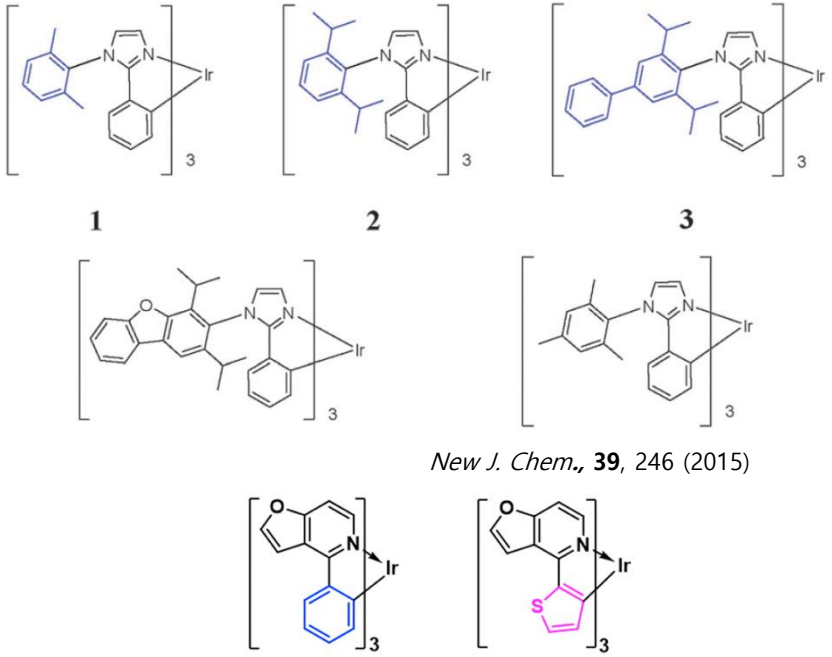


http://www.cadaeic.net/graphpics.htm

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.





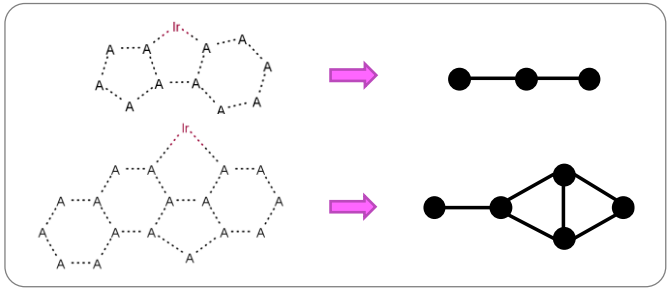
New J. Chem., **39**, 246 (2015)

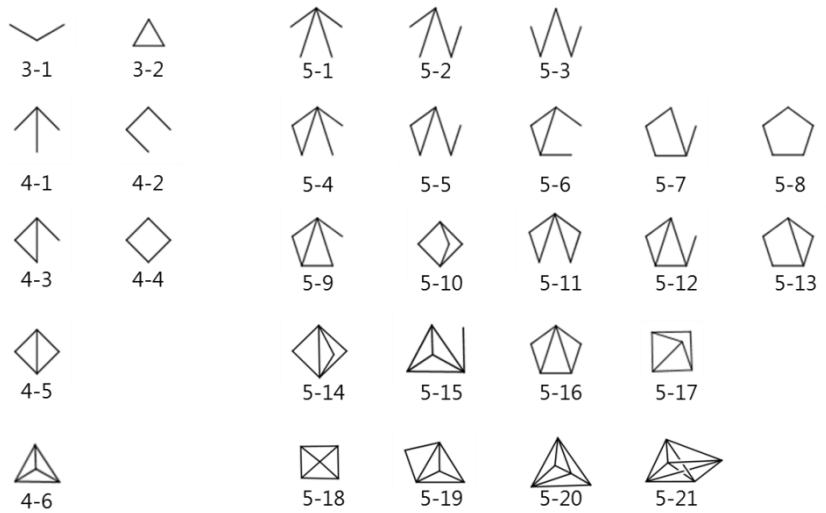
ACS Appl. Mater. Interfaces, **10**, 1888–1896 (2018)
Organic Electronics, **63**, 244–249 (2018)

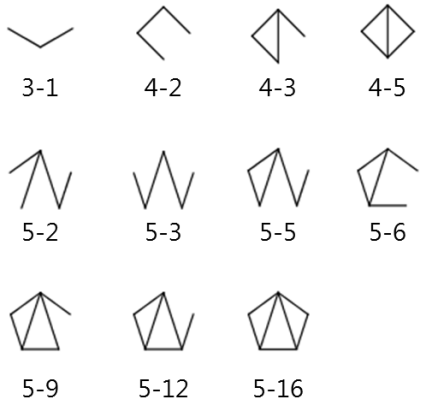
High-Throughput Computational Screening

- **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.

    → Only 11 graphs are valid among the total 29 graphs.





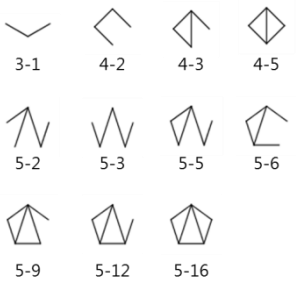


High-Throughput Computational Screening

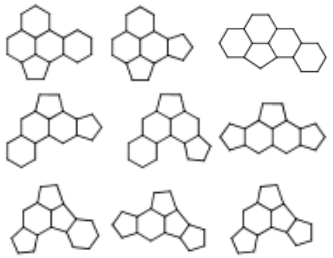
- **Enumeration**
  - For 5- and 6-membered rings.
  - Substitute some carbons of each molecule with nitrogen atoms (max. five).

    → Total 9,919,469 (~10M) core structures

1. Graphs

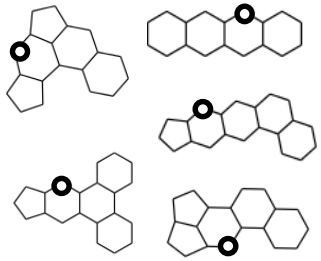


2. Skeletons



total 405 EA

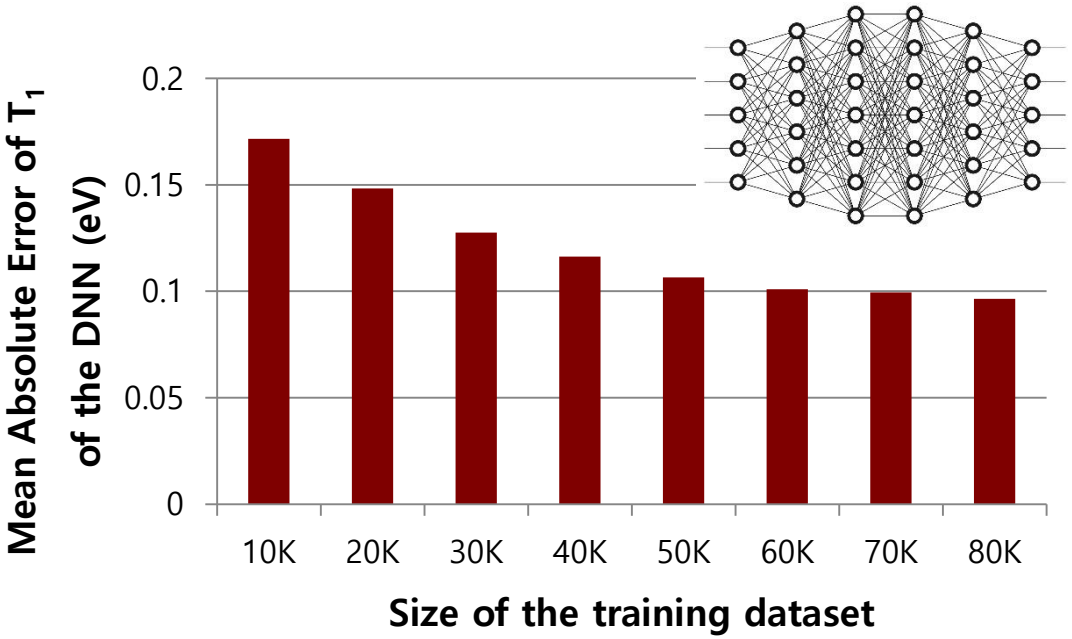
3. Set Iridium positions



⇨ **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)



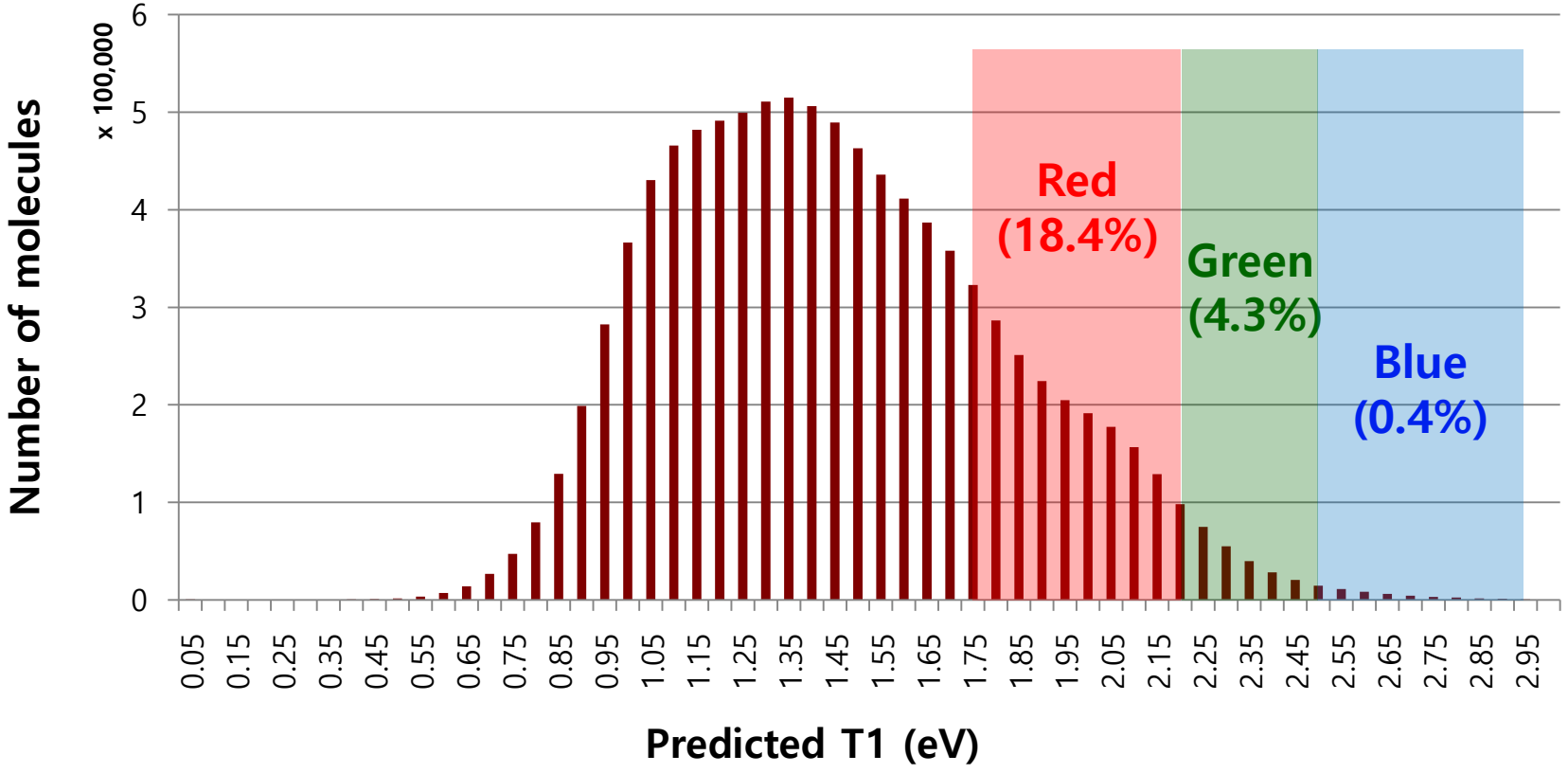
With 80k training data, the average prediction error was less than 0.1 eV

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

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



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.

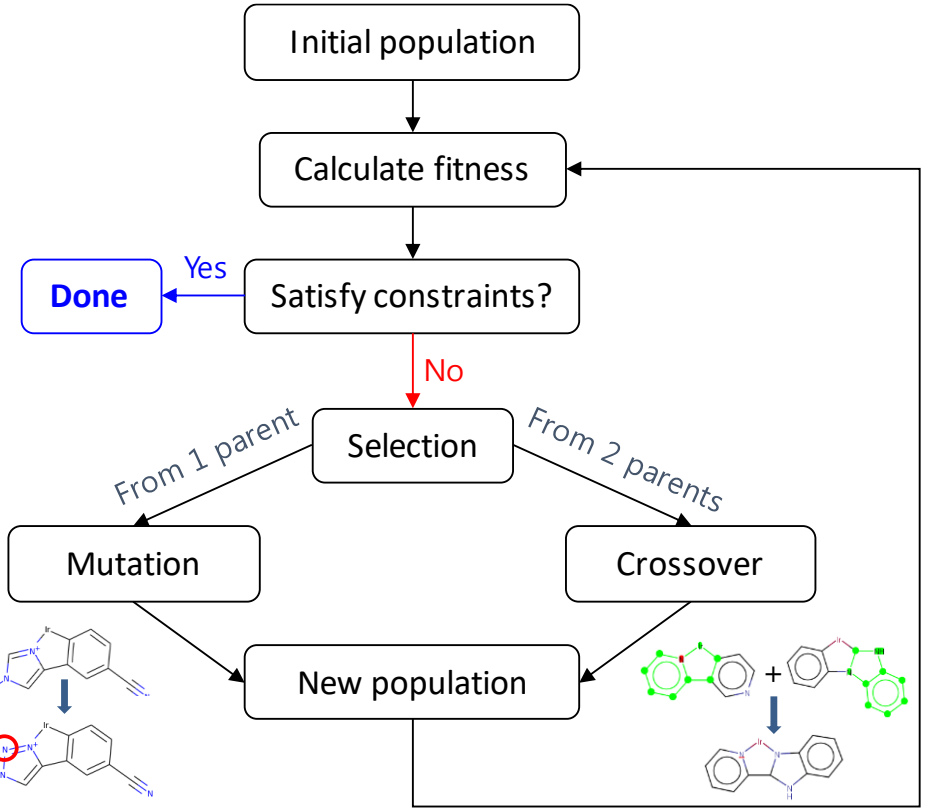
Deep-Learning-based Evolutionary Design

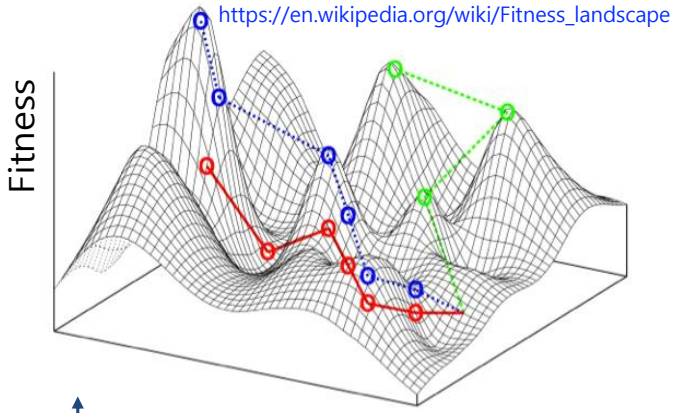
"Evolutionary design of organic molecules based on deep learning and genetic algorithm", COMP, *ACS fall 2018 National Meeting*

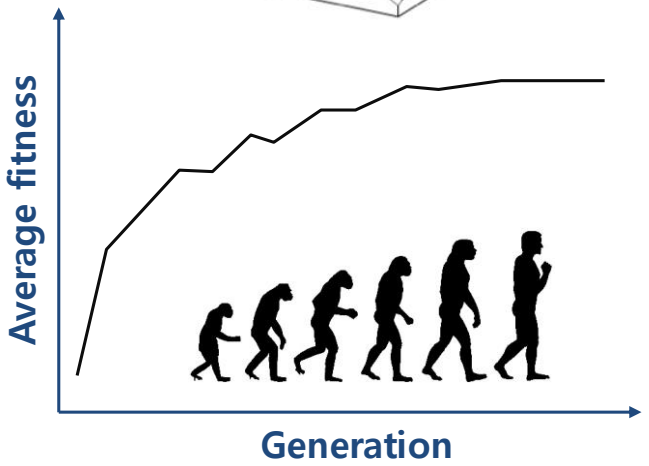
Evolutionary Design

- **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





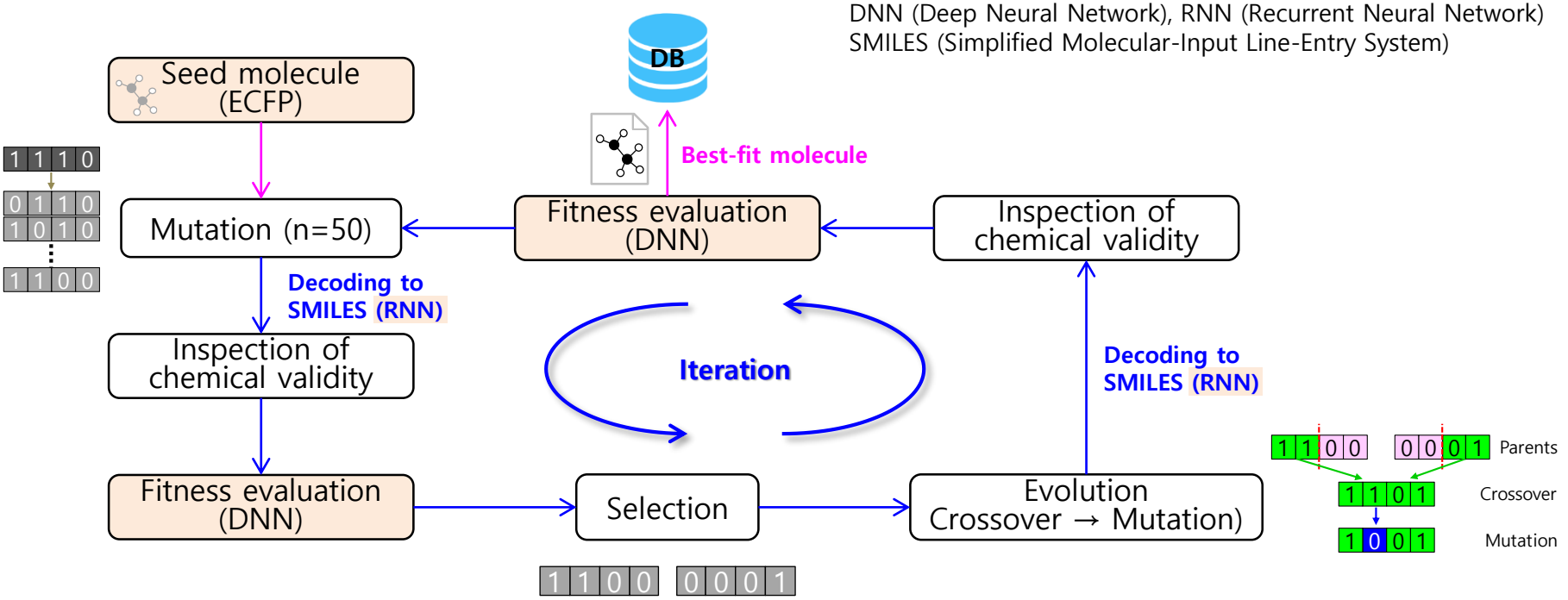


Deep-Learning-Based Evolutionary Design

▪ Proposed approach

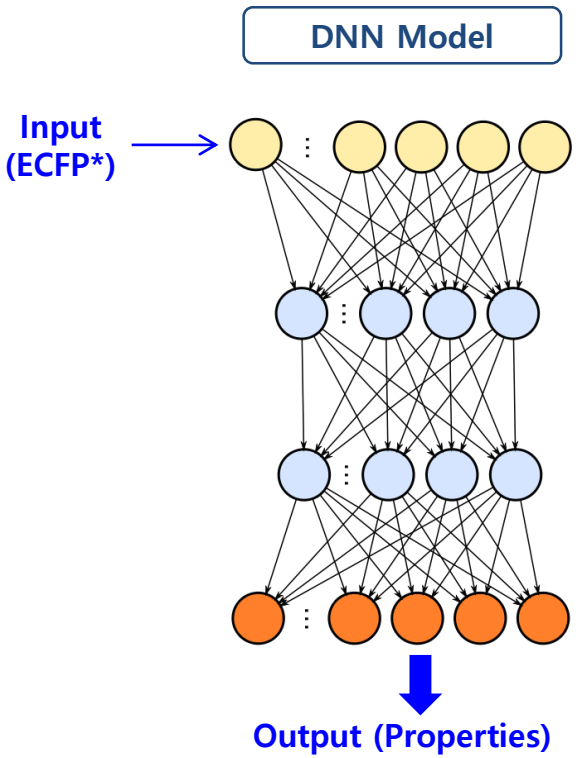
	Conventional	Proposed	Expectations
Molecular Descriptor	Graph or ASCII string	Bit string (ECFP)	• Prevent heuristic bias • Secure chemical validity
Molecular Evolution	Heuristic	Random	
Fitness Evaluation	Simple assessment	DNN	• Versatile evaluation is possible

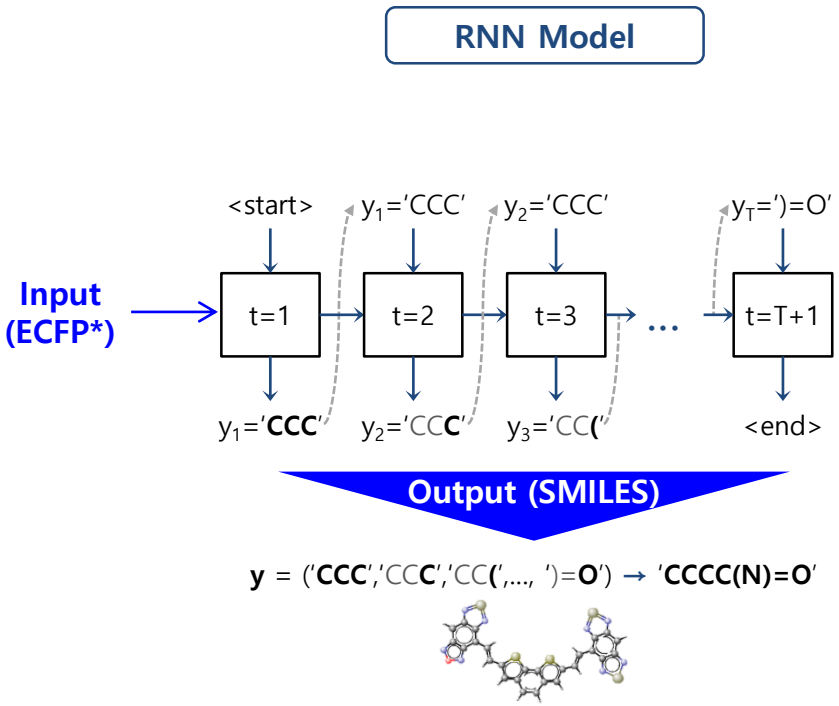
*ECFP (Extended Connectivity FingerPrint)
DNN (Deep Neural Network), RNN (Recurrent Neural Network)
SMILES (Simplified Molecular-Input Line-Entry System)



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**





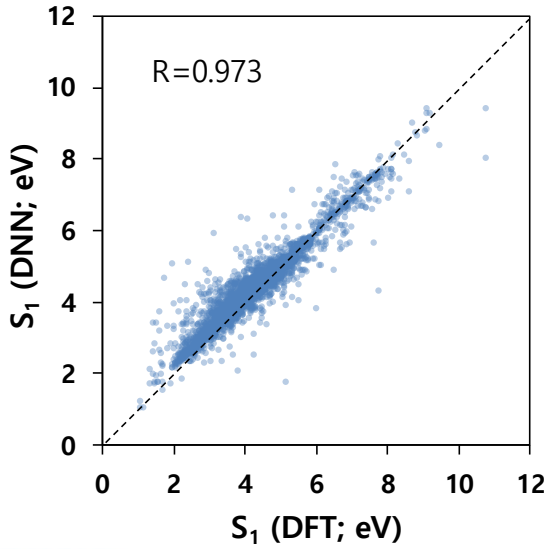
*ECFP (dimension=5,000, neighbor size=6)

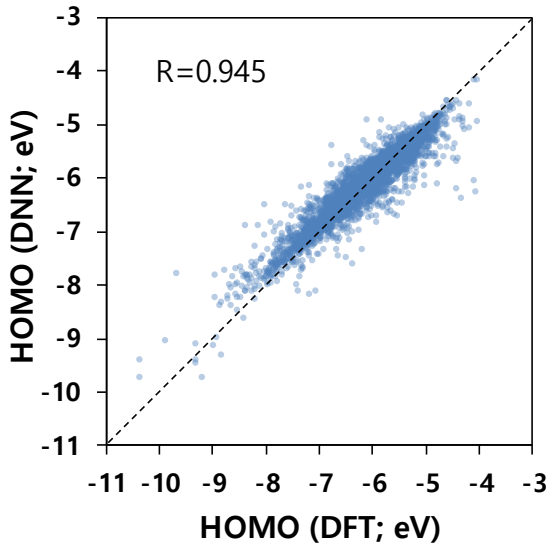
Deep Learning-Based Evolutionary Design

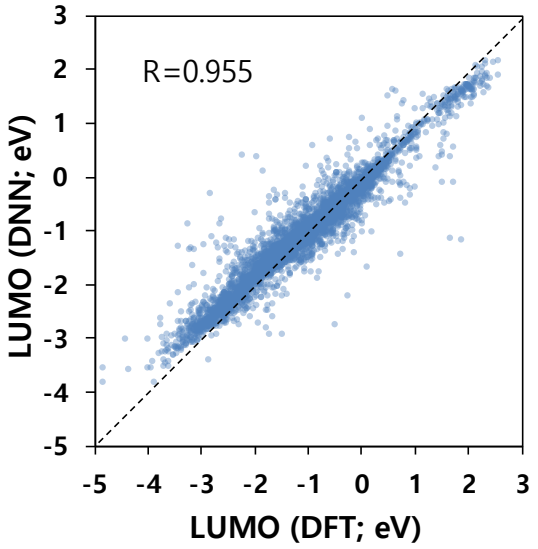
- **Validation test**
  - **Design target: change the S_1 (light-absorbing wavelength) of seed molecules**
  - **Training data: M.W. 200~600 g/mol from PubChem (10,000~50,000 molecules)**

No. of training data	Prediction accuracy of DNN※1 (R, MAE)			Success rate of decoding※2 (RNN)
	S_1	HOMO	LUMO	
① 50,000	0.973, 0.198	0.945, 0.172	0.955, 0.209	86.7%
② 30,000	0.930, 0.228	0.934, 0.191	0.945, 0.224	85.3%
③ 10,000	0.913, 0.278	0.885, 0.244	0.917, 0.287	83.2%

※1. No. of test data=No. of training data/10
※2. Chemical validity is evaluated with RDKit, No. of test data=5,000

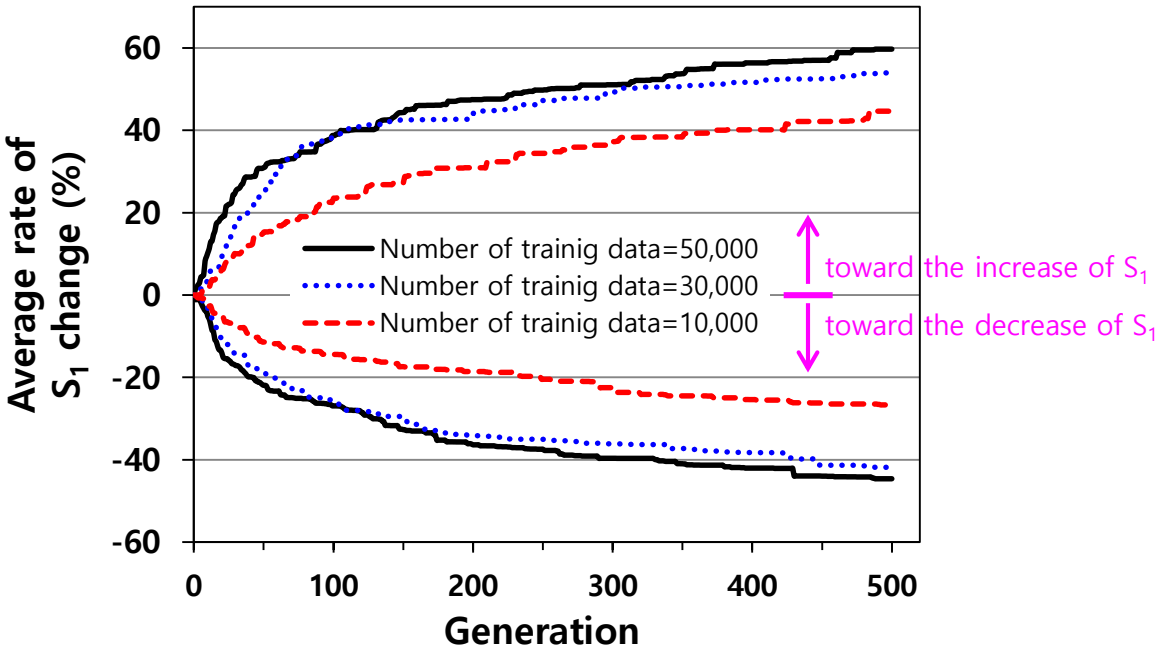


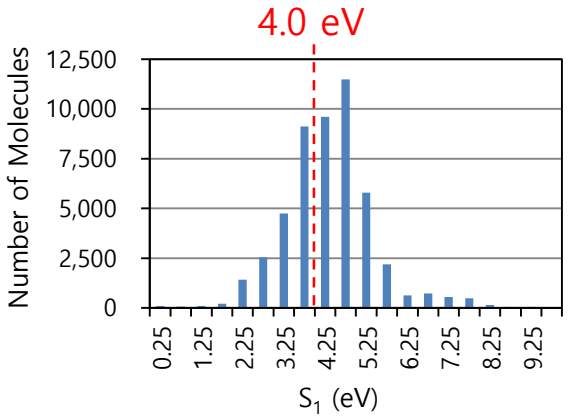




Deep Learning-Based Evolutionary Design

- **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**

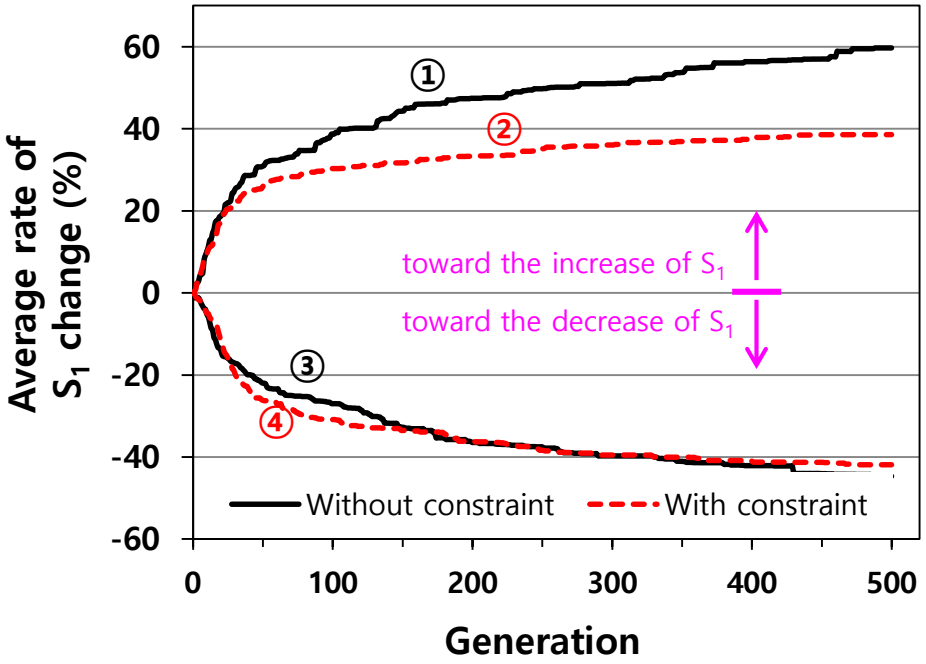


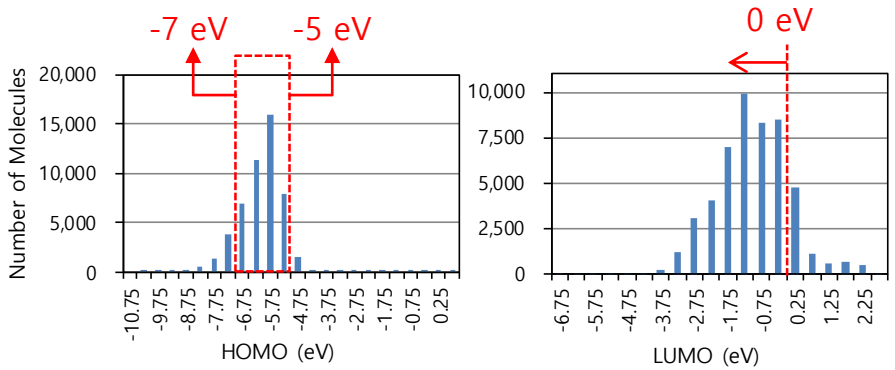


S_1 distribution in the training data (50k)

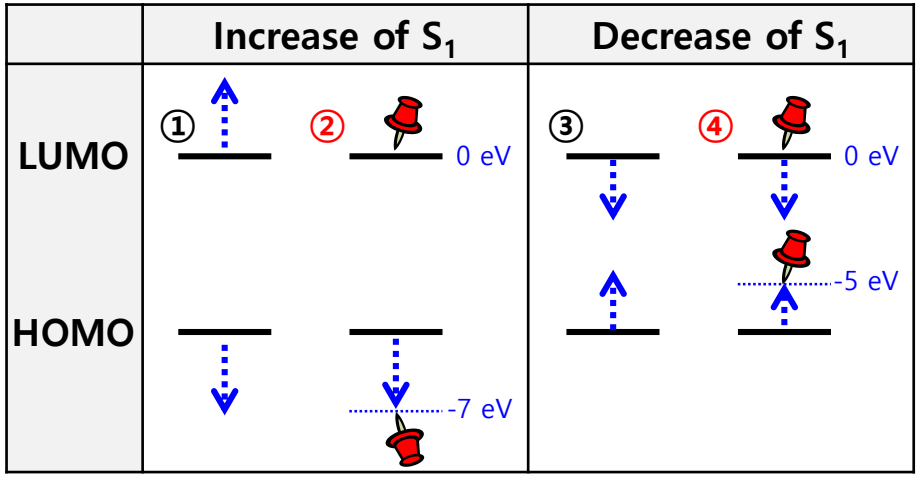
Deep Learning-Based Evolutionary Design

- **Evolution under the constraint of HOMO and LUMO (eV)**
  - **Seed: randomly selected 50 molecules ($3.8<S_1<4.2$)**
  - **Number of training data = 50k**
  - **Constraint: -7.0<HOMO<-5.0, LUMO<0.0**



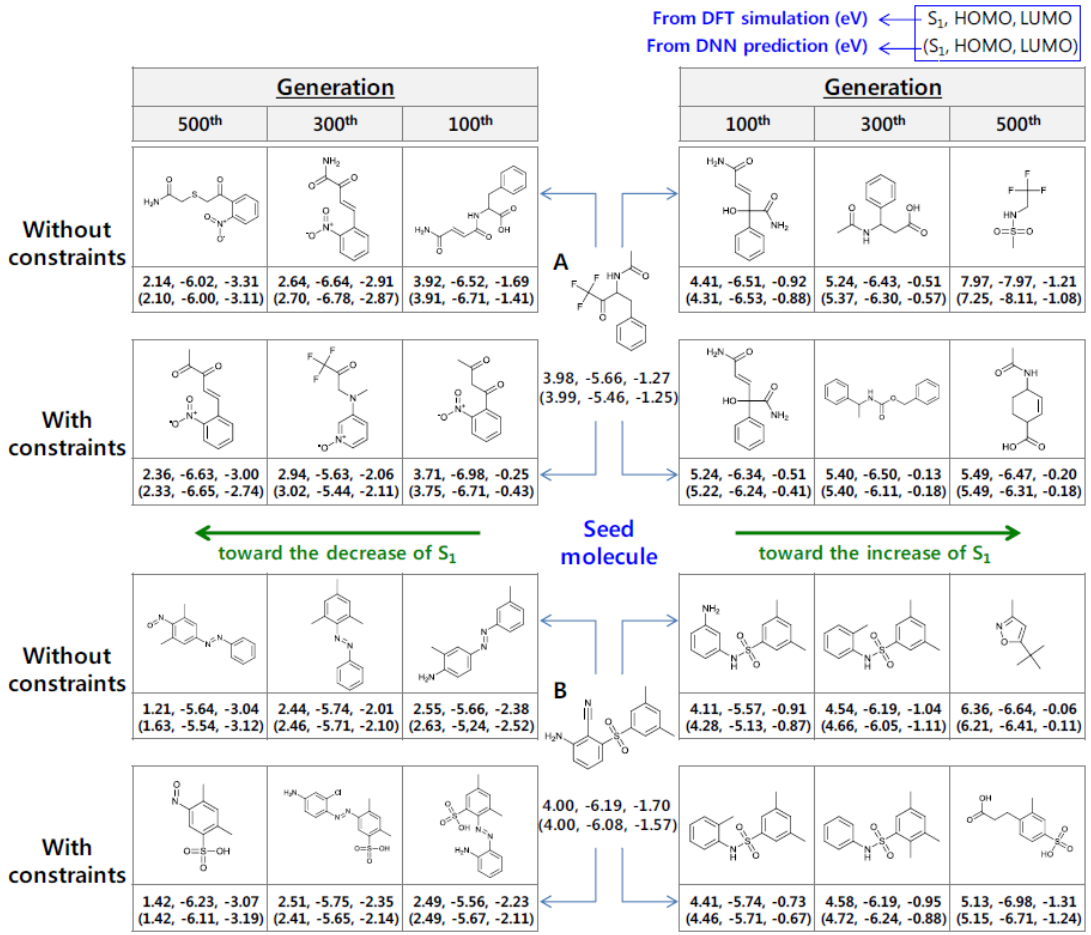


HOMO & LUMO distributions in the training data (50k)



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

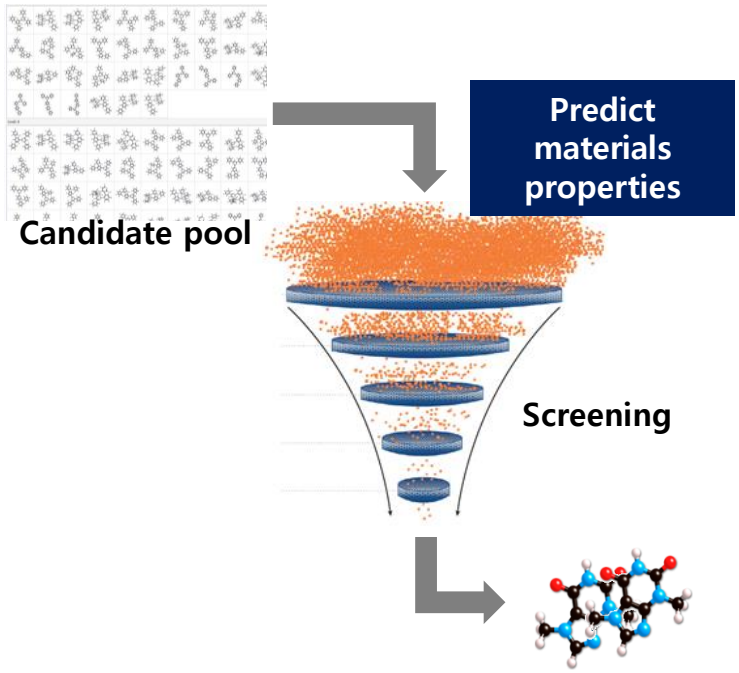
npj Comput. Mater., **4**, 67, 2018

Deep-Learning-Based Inverse Design

- **Paradigm shift of ML in computer-aided materials discovery**

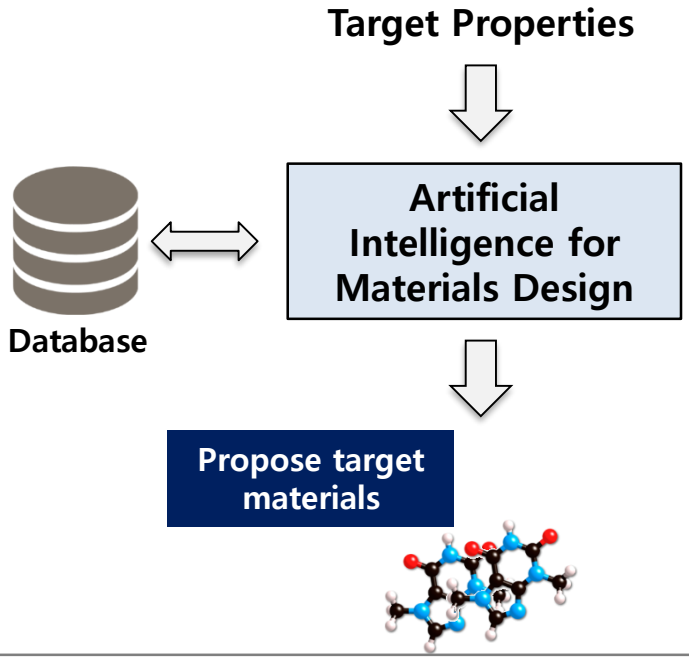
Passive Role

- Efficient screening based on property prediction
- Highly depends on explicit knowledge of chemists



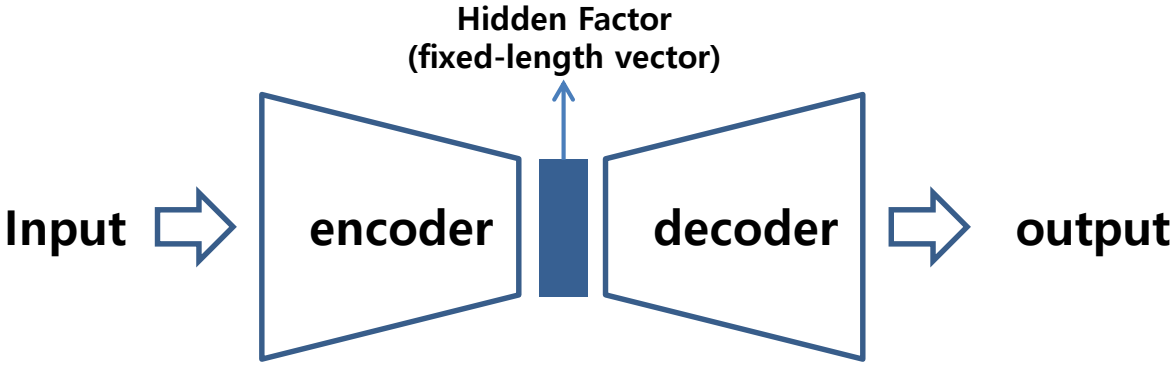
Active Role

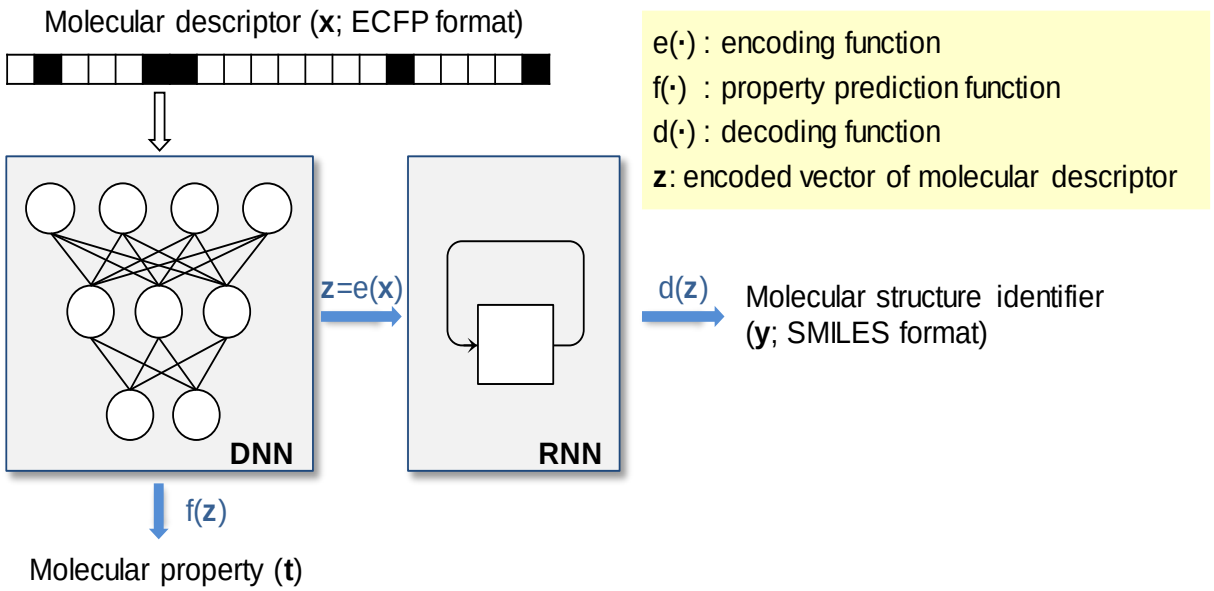
- Propose candidates via automated design
- Provides implicit knowledge from data



Deep-Learning-Based Inverse Design

- **Implementation of inverse-design model**





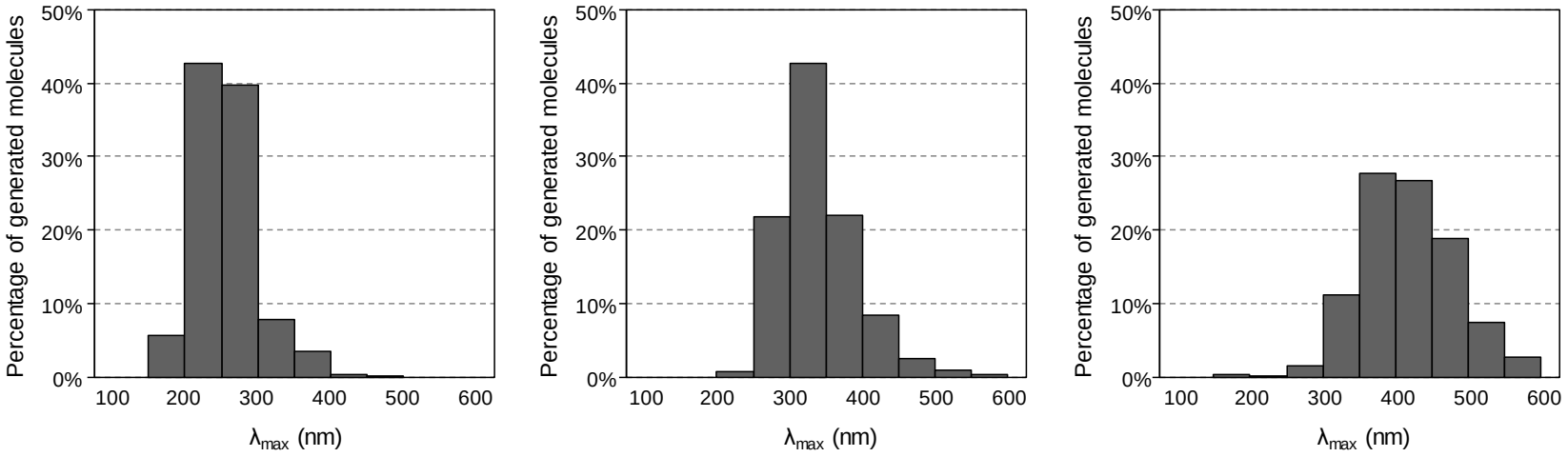
e(·) : encoding function
f(·) : property prediction function
d(·) : decoding function
z: encoded vector of molecular descriptor

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 λ_{max} of the inverse-designed molecules

*Simulation values for the 500 molecules in each target



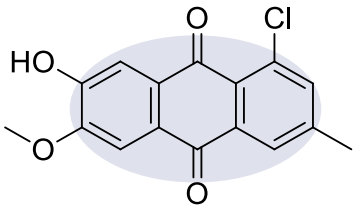
Target	λ_{max}=200–300 nm	λ_{max}=300–400 nm	λ_{max}=400–500 nm
Hit rate	82.6%	64.8%	45.6%

※ About 10% of the designed molecules were found in PubChem even though those were not included in the randomly selected training library.

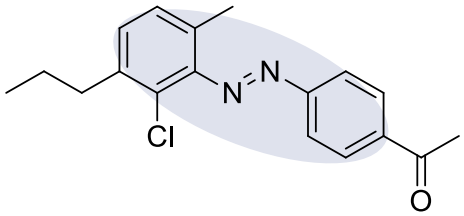
Deep-Learning-Based Inverse Design

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

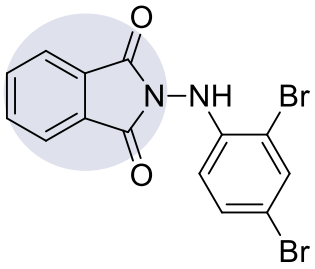
a. Antraquinone derivative
(λ_{max}=433.4 nm)



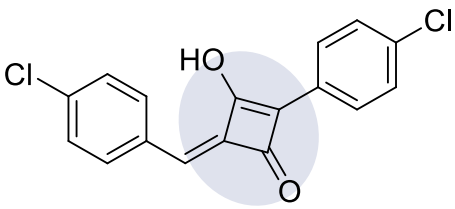
b. Azobenzene derivative
(λ_{max}=527.5 nm)



c. Isoidoline derivative
(λ_{max}=434.4 nm)



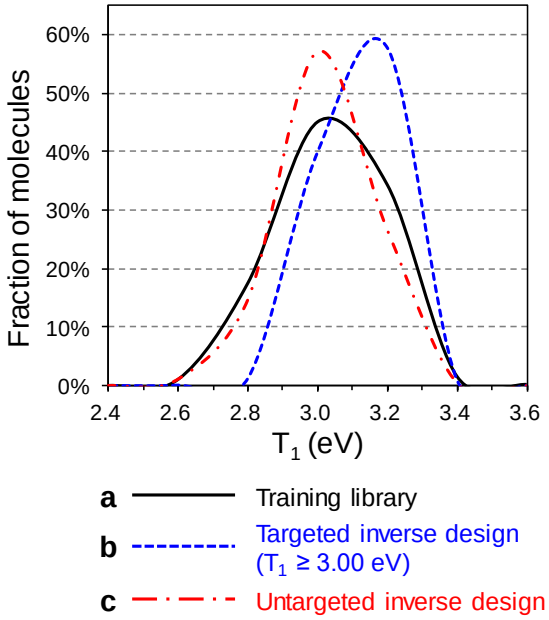
d. Squaraine derivative
(λ_{max}=503.5 nm)



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

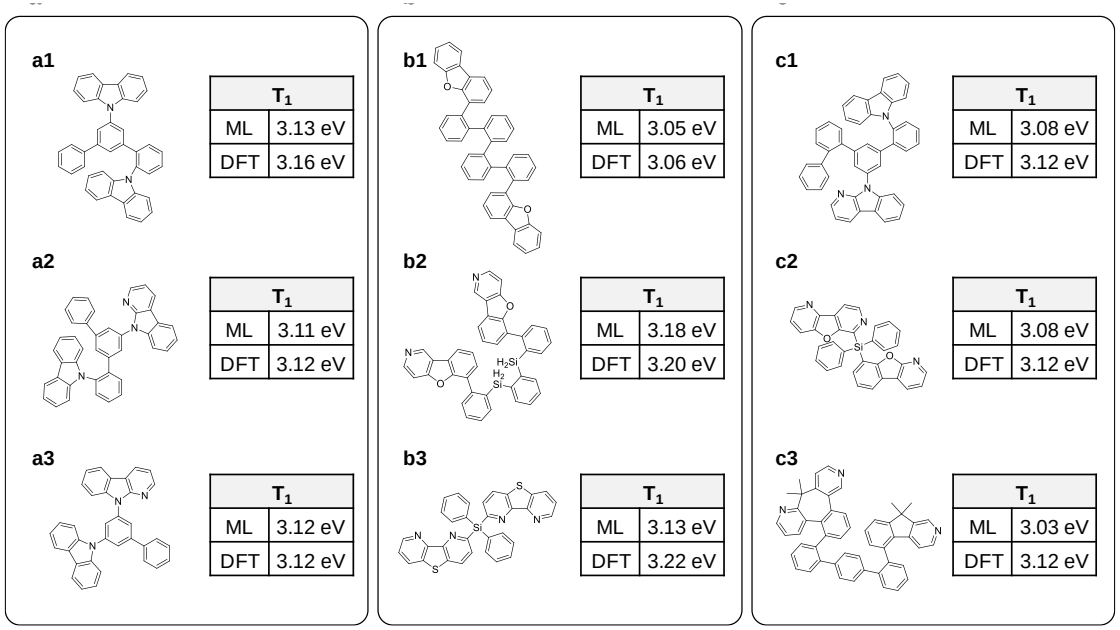
The fractions of the hosts that satisfied the target ($T_1 \geq 3.00$ eV)

- 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



Asymmetric molecules with the given fragments in the training library

Symmetric molecules where the new fragments were introduced

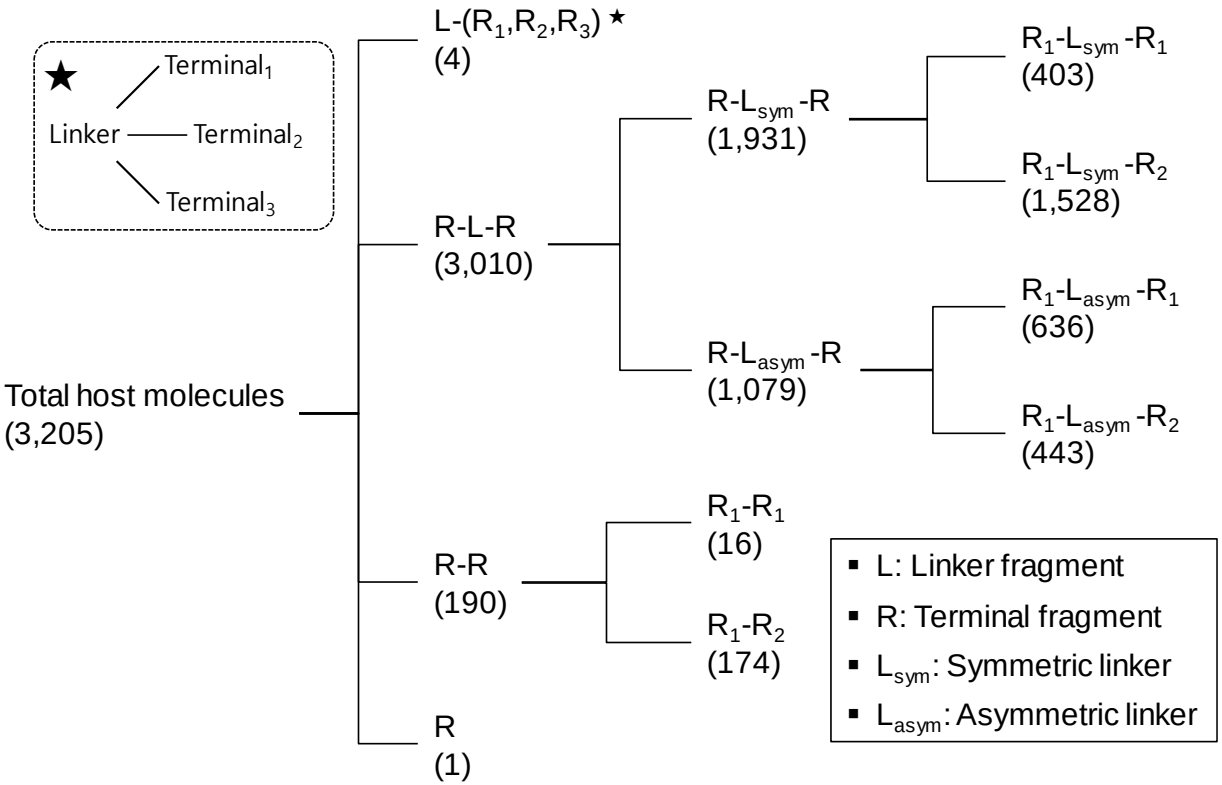
Asymmetric molecules where the new fragments were introduced

	Experiment (eV)				
	HOMO (eV)	LUMO (eV)	S_1 (eV)	T_1 (eV)	ΔE_{ST} (eV)
a1	-5.98	-2.43	3.56	3.06	0.55
b1	-5.96	-2.14	3.64	2.93	1.01
c1	-6.07	-2.65	3.38	2.97	0.46

Deep-Learning-Based Inverse Design

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

The connection rules of the inverse-designed molecules



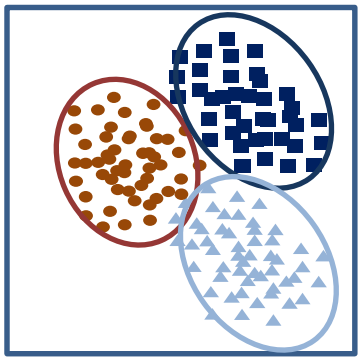
Conclusions

- **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



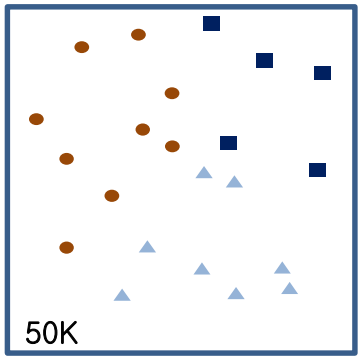
1st trial: 1M Candidates ●
QC simulations take 1.5 years
Fail to find the target structure

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

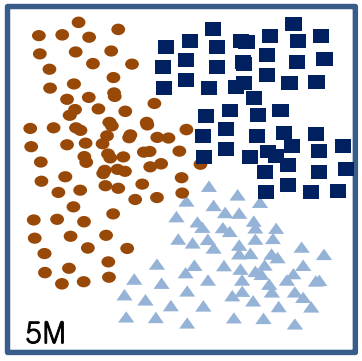


[Step1] Building the training dataset
Needs only QC sim. for 50k molecules (27 days)

Inverse Design

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

Full search



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

Total TAT takes 1 month

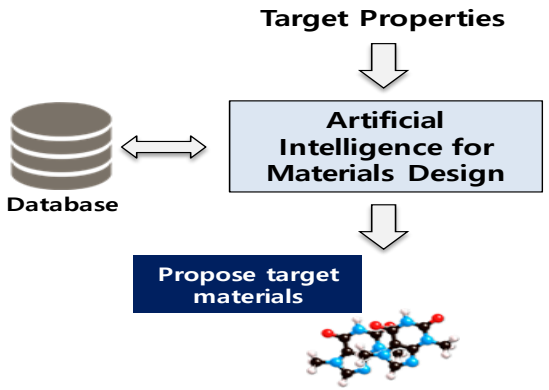
more than 50X speed up (4.5 years *vs.* 1 month)

"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"

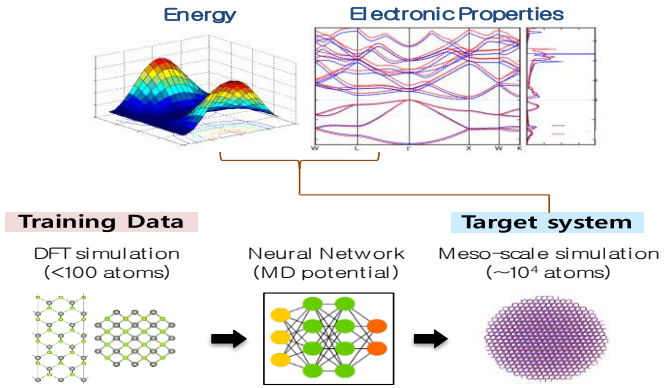
* QC simulation tool : turbomole
Total computational resources=10,000 CPU
In case of 10 CPU computing per molecule, the simulation requires about 13 hrs.

Prospects for AI-based Materials Development

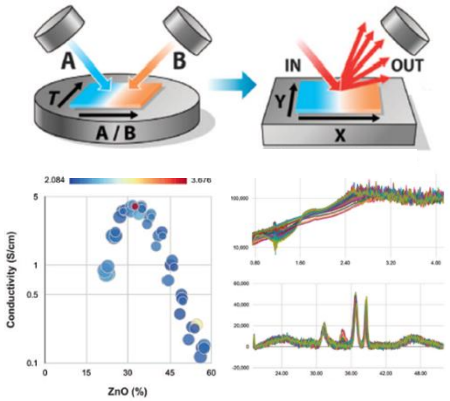
Design



Simulation



Analytical Chemistry



Synthesis

Thank you
(Q&A)

ysuk.choi@samsung.com