



2024 KRICT ChemDX Hackathon

Korea Research Institute of Chemical Technology
Chemical Data–Driven Research Center

Dr. Su-Hyun Yoo



Self-introduction Research path



2013 2015 2020 2022 2023

BSc, Master @Yonsei > PhD@MPIE > Postdoc@MPIE > Research fellow @ICL > Senior researcher @KRICT



Materials Theory Group Prof. Soon

Nanomorphology



Computational Materials Design Prof. Neugebauer

Environment-dependent surface stability



Semiconductor surface/defect



Materials Design Group Prof. Walsh

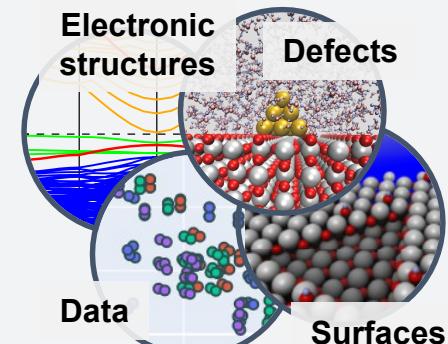
High-throughput @ data-driven research



Chemical Data-Driven Research Center

AI-driven autonomous experimentation

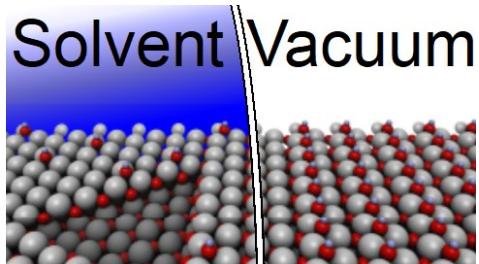
Computational materials science



Self-introduction

Research path

Environment-dependent surface stability

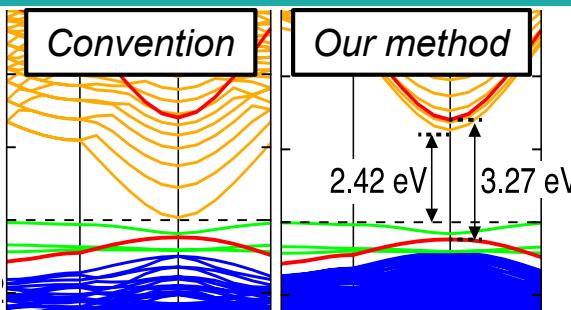


MPIE
Prof. Neugebauer

Phys. Rev. Lett. 120, 066101 (2018)

Editor's suggestion

Computational method development

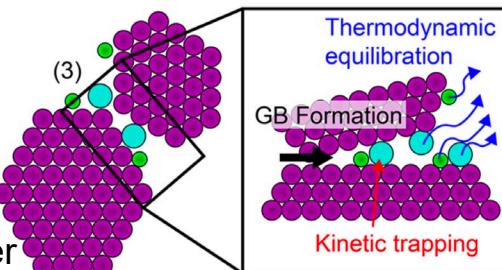


npj Comput. Mater. 7, 58 (2021)

Phys. Rev. Mater. 5, 044605 (2021)

Finalists in the Gerhard Ertl Award 2022

Impurity incorporation in metal nanostructures



Adv. Mater. 2203030 (2022)

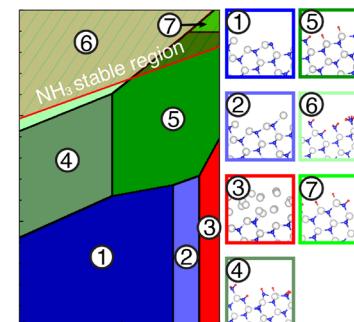
JACS 144, 2, 987 (2022)

New J. Phys. 24, 013008 (2022)

High-throughput research



UCSB
Prof. Van de Walle



On-going project (2024)

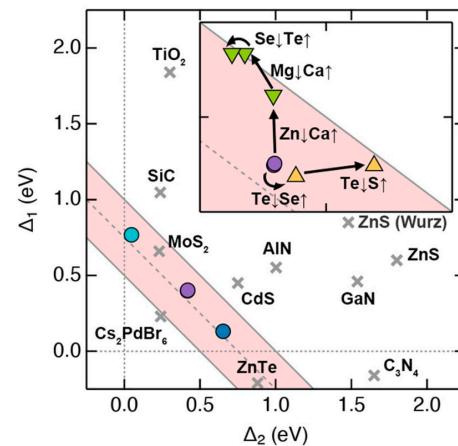


Korea Univ.
Prof. Seho Kim



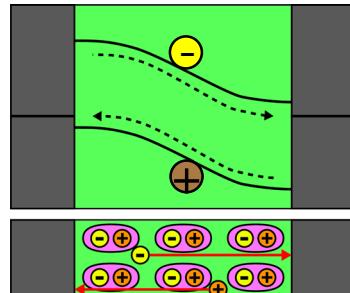
MPIE
Dr. Gault

Data-driven research



npj Comput. Mater. 10, 166 (2024)
J. Phys. Chem. Lett. 15, 6451-6457 (2024)

Next-generation photovoltaic materials



Prof. Walsh

Energy for Future (MSCA co-funded) fellow

Autonomous lab KRICT K2030 project

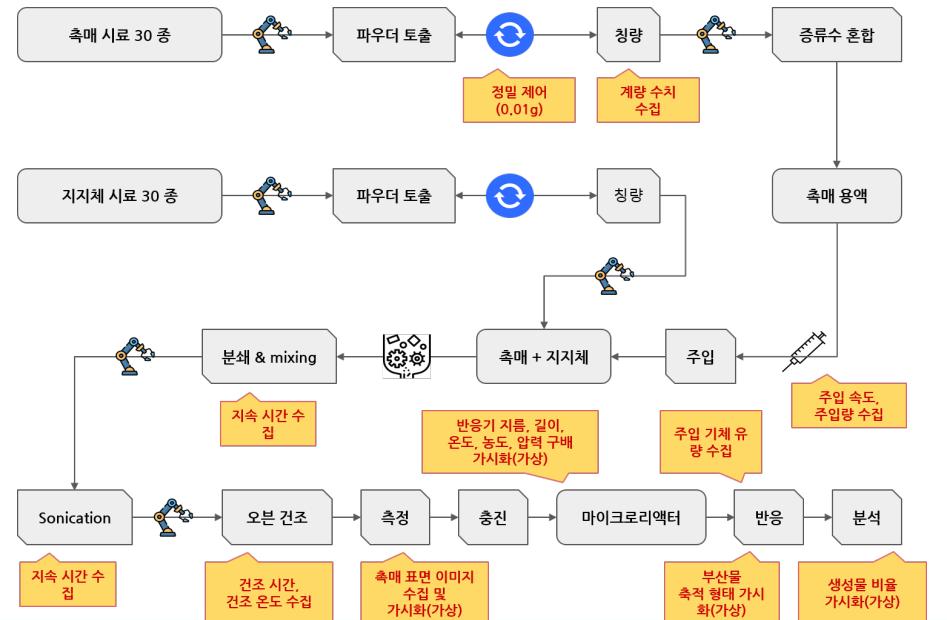


- Design catalysts using AI and intelligent robots

-5 years project

-Accumulate experimental data and design optimized experiments with AI

–Tech. for dealing with powders and solutions, constructing database, and machine-learning research





1

Key Research Area

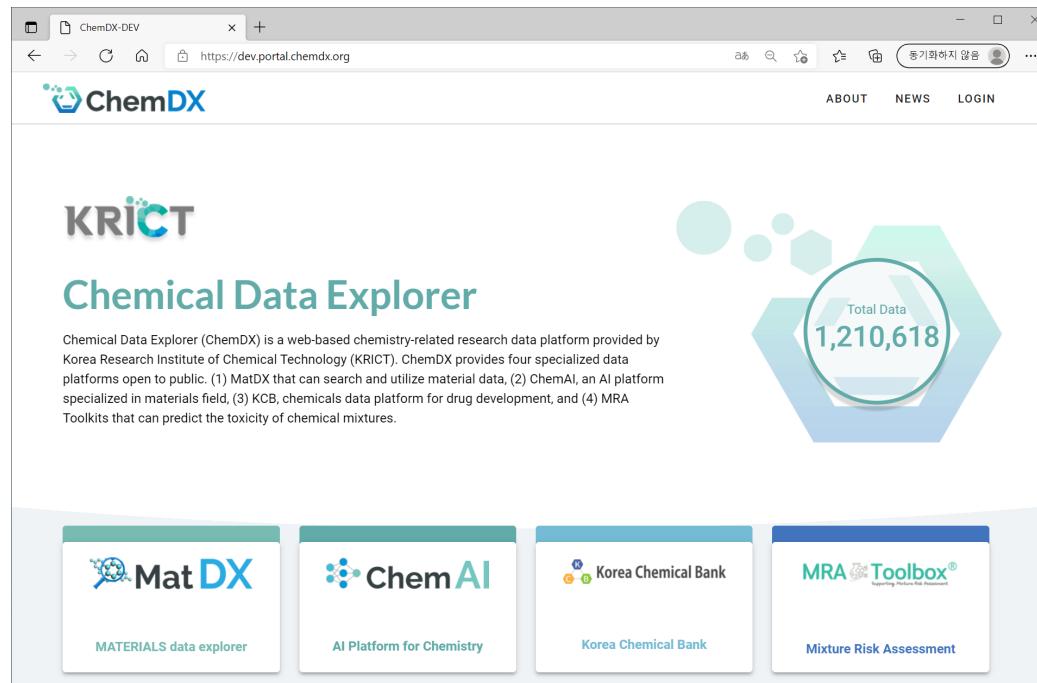


- **Building a Big Data Platform for Chemical Materials**
 - Development of a platform for utilizing chemical materials data
 - Collection and utilization platform for material data
 - Development of a web-based artificial intelligence platform
- **Research on Data-Driven Material Development**
 - Development of material property prediction technology using artificial intelligence
- **Research on Simulation-Based Material Development**
 - Prediction and analysis of material properties using computer simulations

2

ChemDX (Chemical Data Explorer) <https://chemdx.org>

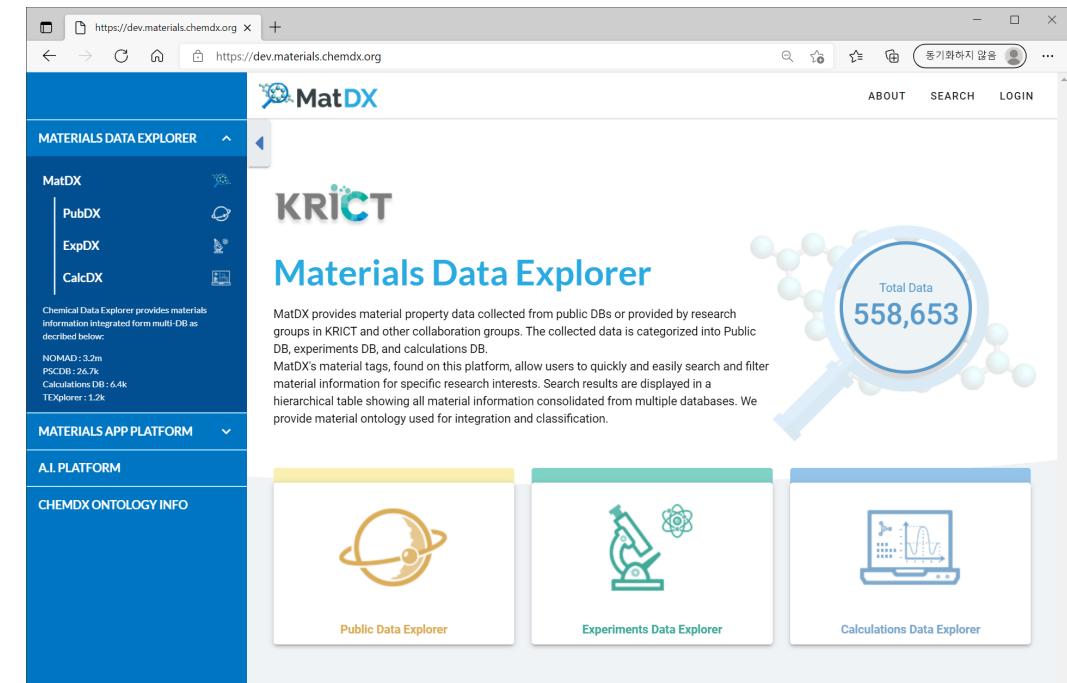
- Integrated Platform for Chemical Material Data
- Chemical material data search
- Compound library, Web-based machine learning, Compound bank, Compound toxicity data



3

MatDX (Materials Data Explorer)

- Public DB: NOMAD, Materials Project, continuously expanding
- ExpDB (to be public): Experimental data on material properties, standard material analysis data, compound toxicity data
- CalcDB: DFT calculation data



4

ChemAI: AI Platform for Chemistry

The screenshot shows the ChemAI platform interface. At the top, there's a navigation bar with links for Home, Toolkits, Machine Learning, Applications, Documentations, and Settings. A search bar is also present. The main content area features a dark blue background with a molecular structure graphic. On the left, there's a section titled "Toolkits: Pre-Trained Machine Learning Models" which lists three pre-trained models: Band Gap Prediction, Formation Energy Prediction, and Thermoelectricity Prediction, each with an R² score. Below this, there's a table showing more details about these toolkits.

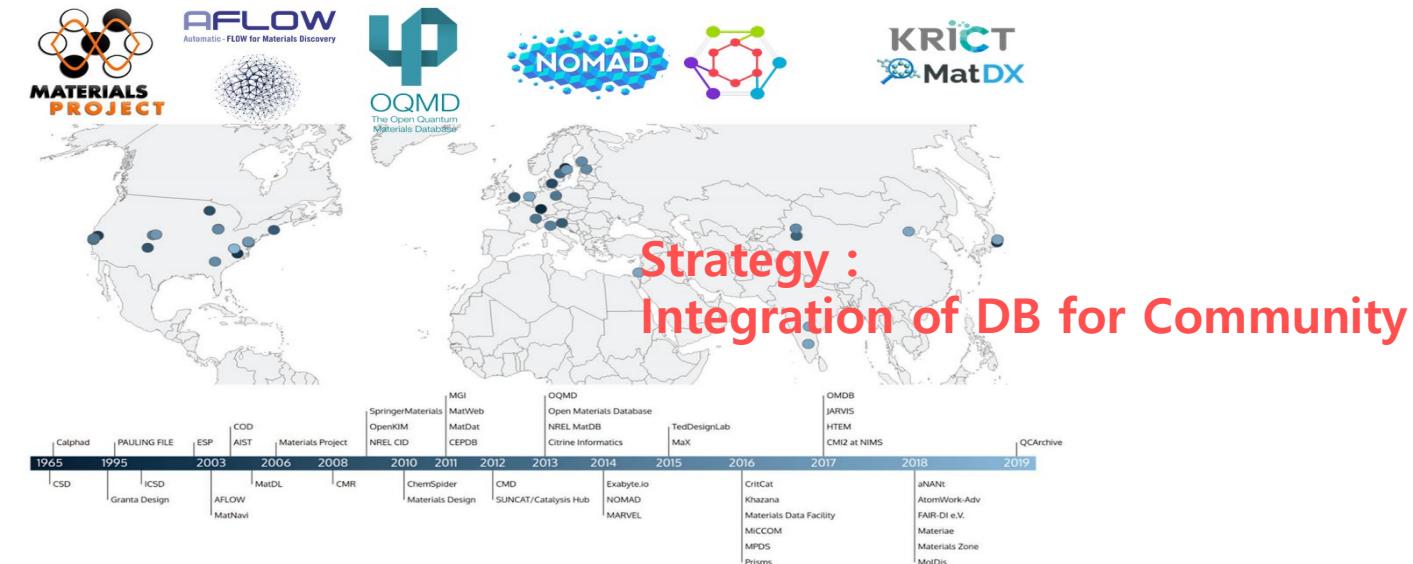
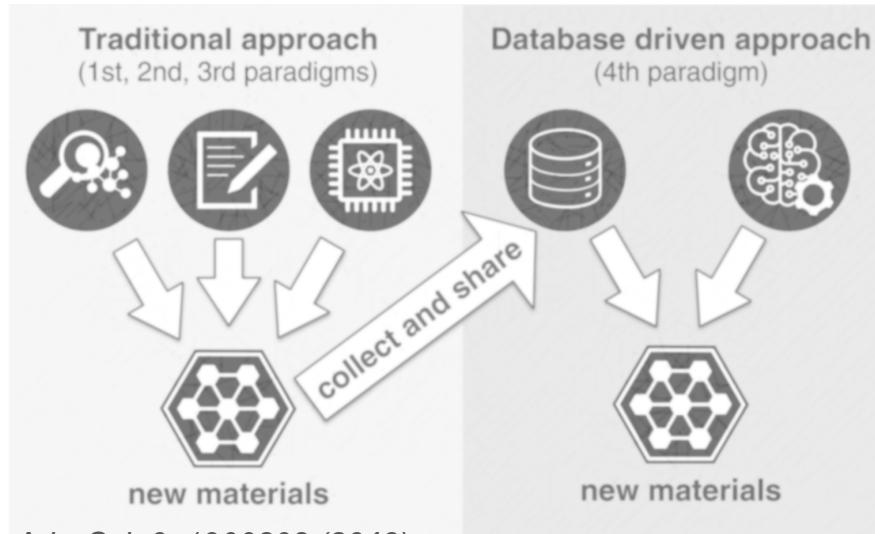
Toolkit	Required Input(s)	Target Property (Unit)	R ² Score
Band Gap Prediction [run] Prediction of experimental band gap	Composition	Experimental band gap (eV)	0.909
Formation Energy Prediction [run] Prediction of experimental formation energy	Composition	Experimental formation energy (eV/atom)	0.907
Thermoelectricity Prediction [run] Prediction of ZT (Figure of merit)	Composition	ZT (Figure of merit)	0.867

5

Data-driven research

- Machine Learning
 - Provides a platform for generating machine learning prediction models using user data
 - Supports various types of user data (chemical composition, molecular structure, crystal structure)
 - Provides example data
- Toolkits
 - Machine learning prediction tools developed by KRICT are available publicly and continuously expanding

ChemDX & MatDX Purpose and vision



The question:

"Why do we need to integrate data repositories?"

which also means,

"What kind of benefits there are if we do integration of data repositories"

"We would like to construct a **dataset** including all the materials information including properties, applications, input data in the materials science field"

ChemDX & MatDX



 <https://chemdx.org> ABOUT NEWS LOGIN

KRICT
Chemical Data Explorer *BETA*

Chemical Data Explorer (ChemDX) is a web-based chemistry-related research data platform provided by Korea Research Institute of Chemical Technology (KRICT). ChemDX provides four specialized data platforms open to public. (1) MatDX that can search and utilize material data, (2) ChemAI, an AI platform specialized in materials field, (3) KCB, chemicals data platform for drug development, and (4) MRA Toolkits that can predict the toxicity of chemical mixtures.



MatDX
Materials Data Explorer

ChemAI
AI Platform for Chemistry

Korea Chemical Bank
Korea Chemical Bank

MRA Toolbox®
Supporting Mixture Risk Assessment

TEXplorer

Catalysis

F-Polymer

SENSOR MATERIALS

Solar Cell

2D Materials

OPD
Organic Photodiodes

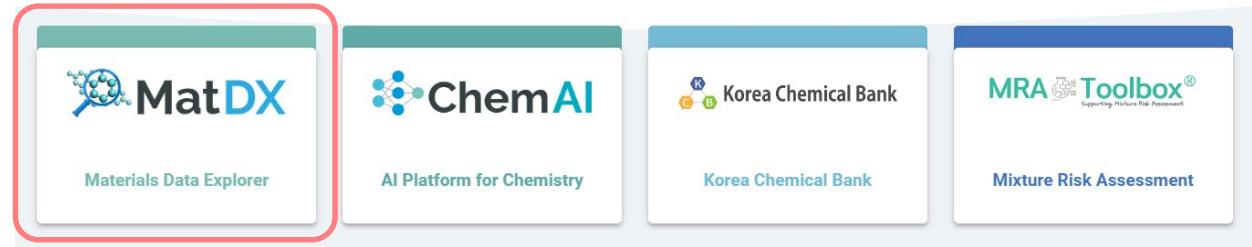
OPV
Organic Photovoltaics

ChemDX & MatDX Structure

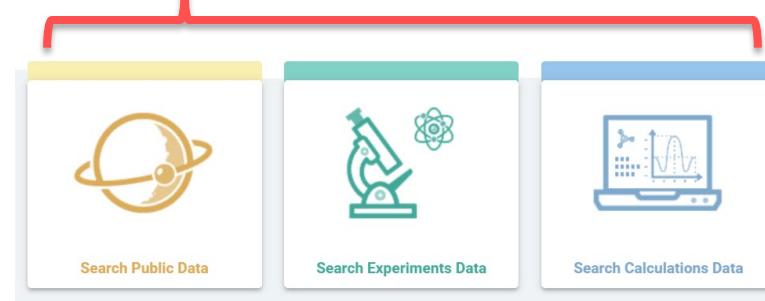


KRICT

Chemical Data Explorer *BETA*



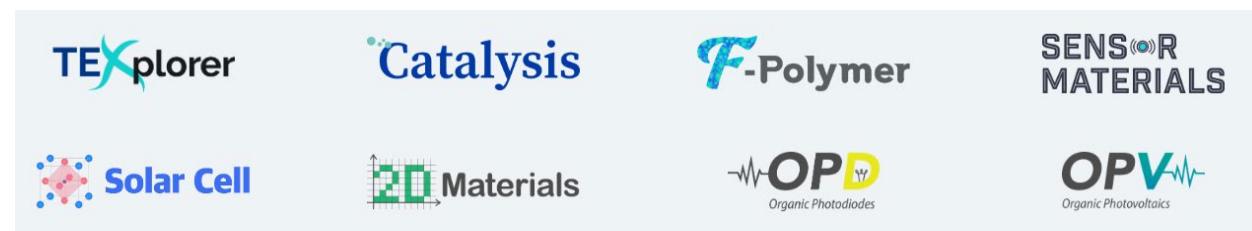
Materials Data Explorer *BETA*



Experimental Data Explorer
ExpDX *BETA*



Materials Applications



ChemDX Hackathon Purpose and vision



- ChemDX in beta version: aiming to gather professional feedbacks to move closer to a public release.
- The main goal of this hackathon is to receive expert/general opinions:
 - Development of general ML models using ChemDX-MatDX data
 - Creative features for data visualization (API) to enhance data interaction
 - UI/UX improvement suggestions to refine user experience
 - Feedback on database usage (particularly for MatDX, SolarCell, and 2D Materials data)
- Freely share your ideas while using DBs and developing your ideas
- Task examples will also be provided to guide discussions.
- We plan to use the outcomes from this hackathon to make meaningful improvements to KRICT's database

Day 1

- Morning: Fresh talk & Team formation
- Lunch
- Afternoon: Development
- Dinner: (한우천국: 0507-1349-1140) walk within 10min

Day 2

- Morning: Development
- Lunch
- Afternoon: Development
- Dinner: (더함뜰: 0507-1402-9294) walk within 15min

Day 3

- Morning: Presentation & Evaluation
- Lunch
- Afternoon (~3-4PM): Award ceremony
- Dinner: (더함뜰: 0507-1402-9294) walk within 15min

Hackathon Guidelines



- Team Composition
 - Teams of 2-3 members, formed through 1-2 minute introductions
- Project Topics
 - Choose freely after exploring ChemDX or select from the example sets
 - Apply your skill for machine learning, data visualization, or UI/UX enhancements to ChemDX datasets
- ChatGPT account
 - Free to use only for the hackathon (<https://chatgpt.com/>)
 - ID: syoo.gpt@gmail.com
 - PW: gpt.syoo123
 - Available only during the hackathon
- Communication
 - Use GitLab issue tracker for questions or support

Project examples



Modeling Tasks

MatDX: ML model to predict formation energy of materials

LitDX - TE: ML model to predict properties of thermoelectric materials (e.g., seebeck coefficient)

Solar cell: ML model to predict fill factor, efficiency, V_{oc} , and J_{sc}

LitDX - IPO: Build ML model to predict properties of inorganic phosphoroptical properties (e.g., Emission max.)

Visualization Tasks

Develop functions for graphing and representing data, such as MatDX, SolarCell and LitDX

Develop Raman spectrum fitting functions using data from Sensor Materials

UI/UX Improvement Tasks

Enhance visualization, improve webpage layout, add statistical analysis and Raman spectrum fitting for Sensor Materials

General Tasks

Identify and suggest specific improvements for user experience

Project Evaluation and Awards

Evaluation

10-min presentation
on the final day

Submit materials
syoo@krikt.re.kr

PPT: explain
your idea, work flow,
data utilization,
and how it works, etc

Source code: e.g., jupyter
notebook, python file, etc

- Evaluation committee
 - KRICT: Dr. Jungho Shin
 - Yonsei University: Prof. Woosun Jang
 - Fordham University: Prof. Joshua Scrier

Bayesian Optimization for Composition Space Exploration

Guiding experimental catalyst design under the scheme of one-at-a-time of materials design

- Find next composition based on the Bayesian optimization
- Gaussian process (GP) regression implemented in scikit-learn with Matern($\nu=1.5$) kernel
- Kernel parameter is fixed, and therefore will not be optimized during training of GP model
- In current implementation, only one variable is used as input of GP

```
import numpy as np
import pickle
import pandas as pd

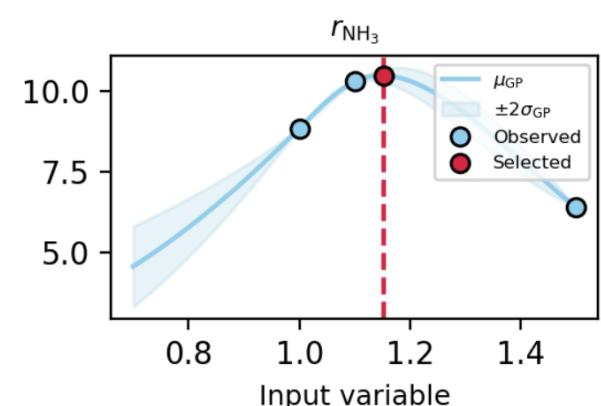
from scipy.optimize import minimize
from scipy.stats import norm

from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import *

import matplotlib.pyplot as plt
```

```
plt.subplot(2,1,2)
plt.plot(x_grid, EI_grid, 'r', c='skyblue', label='EI')
plt.plot(x_EI_train, 'o', c='skyblue', mec='k', label='Observed')
plt.axvline(x_grid[np.argmax(EI_grid)], c='crimson', ls='--')
plt.plot(x_grid[np.argmax(EI_grid)], np.max(EI_grid), 'o', c='crimson', mec='k', label='Selected')
plt.title('Expected Improvement', fontsize=9)
plt.xlabel('Input variable', fontsize=9)
plt.legend(loc='upper right', fontsize=6)
plt.tight_layout()

# Save figure
plt.savefig(filename_figure, dpi=200)
```



Project Evaluation and Awards



Evaluation

10-min presentation
on the final day

Submit materials (syoo@kRICT.re.kr)

PPT: explain your idea, work flow, data utilization, and how it works, etc

Source code: e.g., jupyter notebook, python file, etc

• Evaluation committee

- KRICT: Dr. Jungho Shin
- Yonsei University: Prof. Woosun Jang
- Fordham University: Prof. Joshua Scrier

• Evaluation criteria

1. Data Utilization and completeness (30%):
 - How effectively did the team utilize the provided chemical and material property data?
 - How complete is the solution or prototype in addressing the problem within the given time?
2. Creativity and Innovation (40%):
 - How original and innovative is the proposed solution?
 - Did they present new perspectives and creative solutions for problem-solving?
3. Technical Feasibility (15%):
 - How well did the team overcome technical challenges and limitations?
4. Teamwork and Presentation (15%):
 - How effectively did the team collaborate with each other?

Q&A

- Use the GITLAB issue board
- <https://gitlab.chemdx.org/global-network/2024-krikt-chemdx-hackathon/-/issues>

The image shows two screenshots of the GitLab interface. The left screenshot displays the main issue board for the 'global-network > 2024-krikt-chemdx-hackathon' project. It features a search bar at the top, a sidebar with navigation icons, and a central area with a rocket and magnifying glass illustration. A red box highlights the 'New issue' button in the top right corner. An arrow points from this button to the right screenshot, which shows a detailed view of a single issue. The issue is titled 'test' and was created by 'Suhyun Yoo' 6 days ago. It has a status of 'Closed'. The right screenshot also shows a comment from 'Suhyun Yoo' just now, replying to the test issue.

Project examples



Modeling Tasks

MatDX: ML model to predict formation energy of materials

LitDX - TE: ML model to predict properties of thermoelectric materials (e.g., seebeck coefficient)

Solar cell: ML model to predict fill factor, efficiency, V_{oc} , and J_{sc}

LitDX - IPO: Build ML model to predict properties of inorganic phosphoroptical properties (e.g., Emission max.)

Visualization Tasks

The screenshot shows the PubDX search interface. At the top, there are three tabs: CHEMDX, MATDX, and SEARCH. Below the tabs is a sidebar with categories: MATERIALS DATA, ExpDX DATA COLLECTIONS, and MATERIALS APPLICATIONS. The main area is titled "Search" and features a "Selected Tags:" section with a "FIND TAGS" button. A "POPULAR" tab is selected, showing various tags like Binary, O, NOMAD, Semiconductor, Solar Cell, C, AI, Total energy, Ternary, Bulk, Ti, Ag, DOS, Li, Zn, H, Si, Fe, Cs, V, Co, Oxide, Mn, Metal, Ni, Pb, Thermoelectric, Zr, Cubic, and Eigenvalues. Below this is a link "Search with periodic table". The search results section displays "563,812 compositions found" with a link to "View details". The results are presented in a table with columns: Formula/name, Composition & Structure, Property & Application, and Reference. Two entries are shown: Rh₄SnTe and LiNbSe. For Rh₄SnTe, the composition is listed as Ternary, Bulk, Cubic, and the properties include Total energy, LUMO, HOMO, Eigenvalues, all from the NOMAD reference. For LiNbSe, the composition is listed as Ternary, Bulk, Cubic, Tetragonal, Trigonal, and Orthorhombic, and the properties include LUMO, DOS, HOMO, Fermi energy, Eigenvalues, and Total energy, also from the NOMAD reference.

UI/UX Improvement Tasks

Project examples



Modeling Tasks

MatDX: ML model to predict formation energy of materials

LitDX - TE: ML model to predict properties of thermoelectric materials (e.g., seebeck coefficient)

Solar cell: ML model to predict fill factor, efficiency, V_{oc} , and J_{sc}

LitDX - IPO: Build ML model to predict properties of inorganic phosphoroptical properties (e.g., Emission max.)

Visualization Tasks



Intro

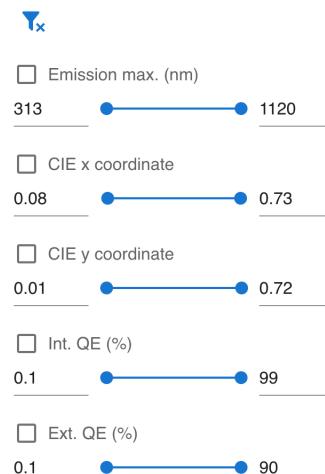
Thermoelectric Materials

Inorganic Phosphors

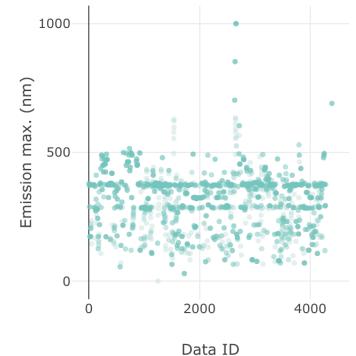
Inorganic Phosphors

Intro

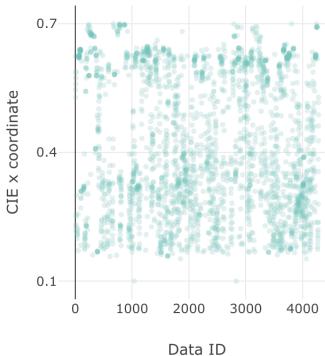
We thoroughly searched published literature to gather data on experimentally synthesized inorganic phosphors and their corresponding physical properties. We have named this comprehensive collection of data the Inorganic Phosphor Optical Property (IPOP) dataset. This dataset contains information on various physical properties of inorganic phosphors, such as photoluminescence (PL) maximum wavelengths, CIE coordinates, lifetime, and quantum efficiency. The IPOP dataset comprises 16,023 observations from 2,238 host materials reported in 553 published papers. More details about the collected data can be found in the following paper. (<https://www.nature.com/articles/s41598-024-58351-w>)



Emission max. (nm)



CIE x coordinate



ChemDX MatDX

age plan

nts for

Project examples



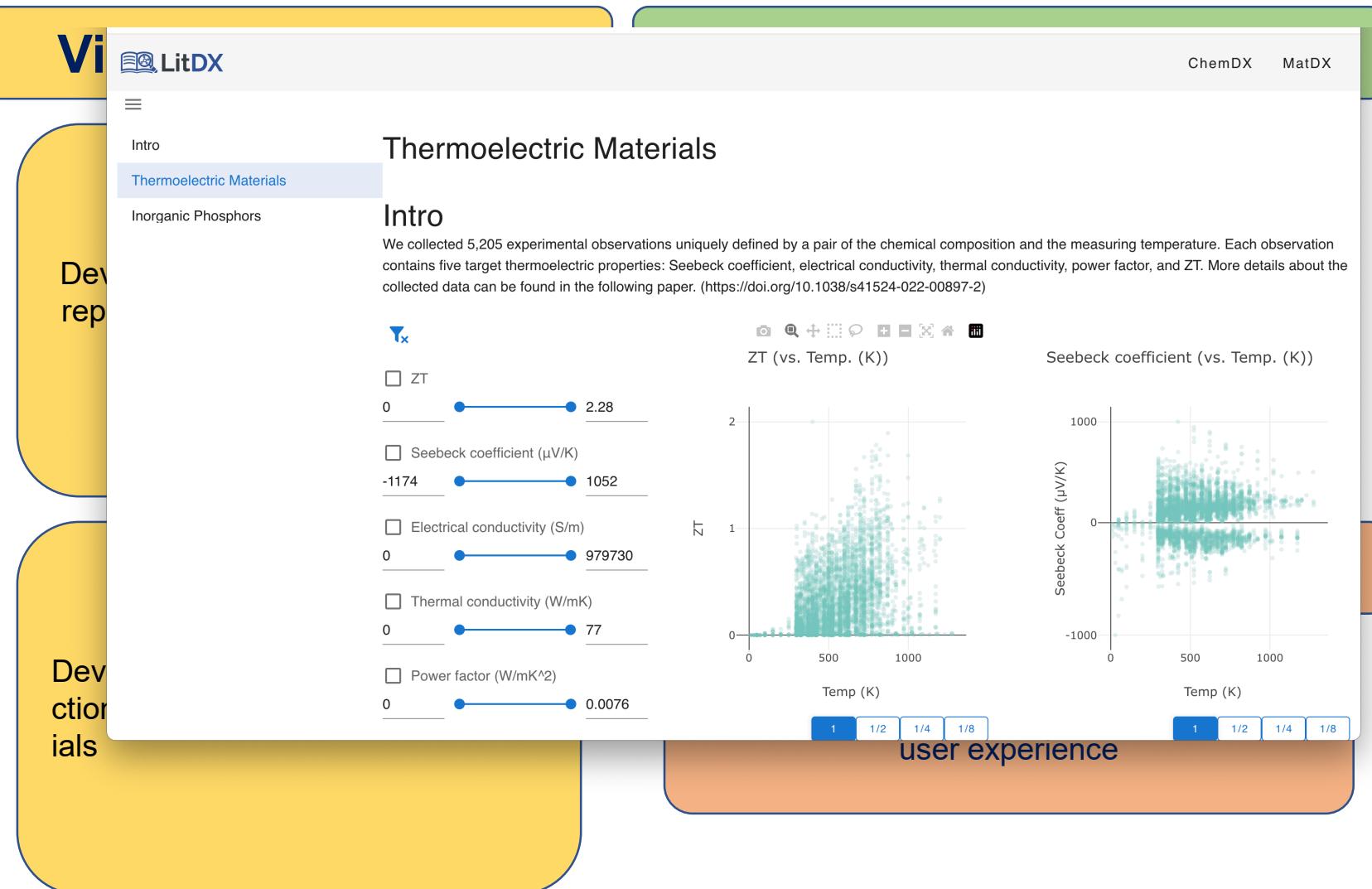
Modeling Tasks

MatDX: ML model to predict formation energy of materials

LitDX - TE: ML model to predict properties of thermoelectric materials (e.g., seebeck coefficient)

Solar cell: ML model to predict fill factor, efficiency, V_{oc} , and J_{sc}

LitDX - IPO: Build ML model to predict properties of inorganic phosphoroptical properties (e.g., Emission max.)



Project examples



Modeling Tasks

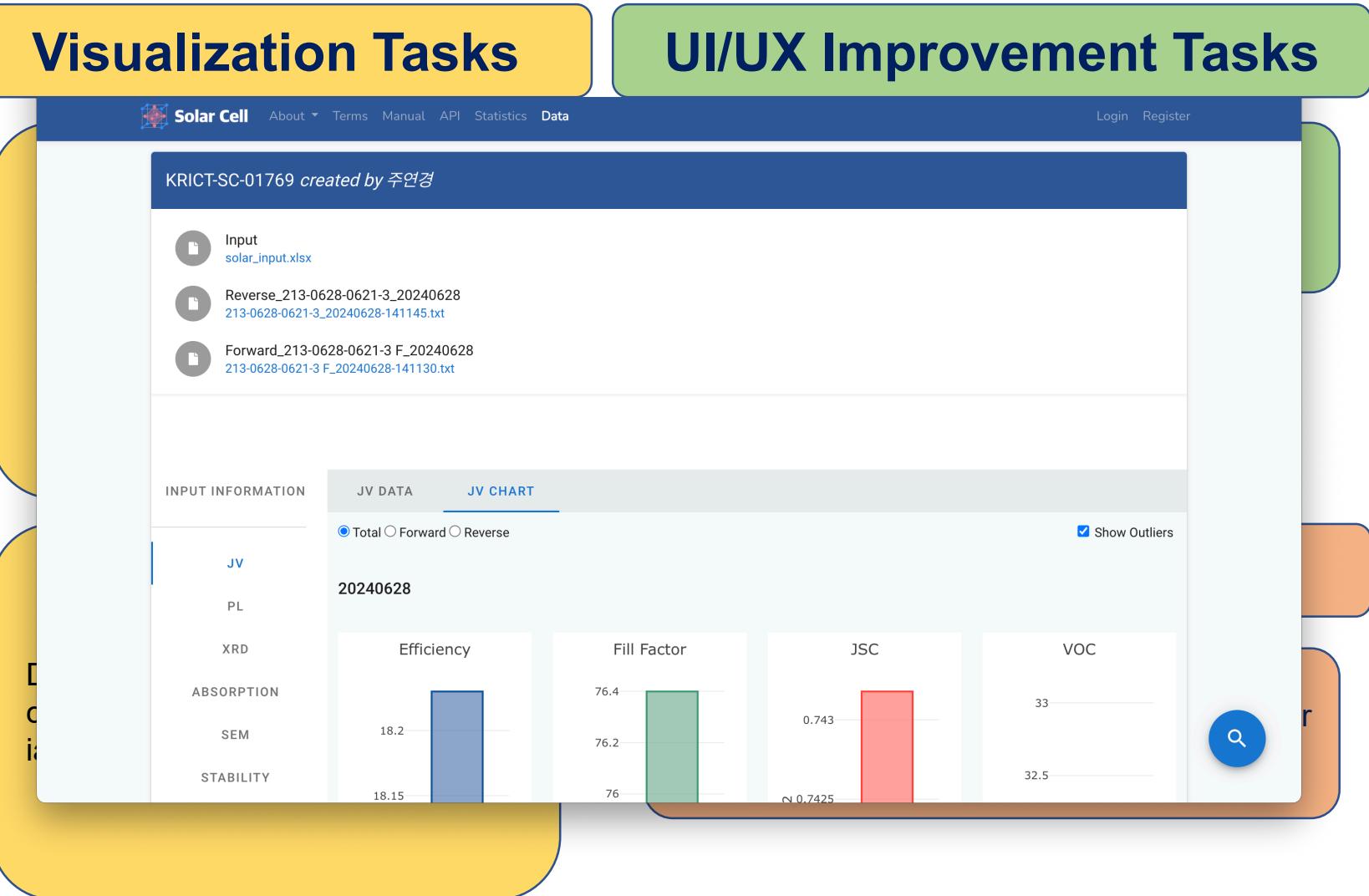
MatDX: ML model to predict formation energy of materials

LitDX - TE: ML model to predict properties of thermoelectric materials (e.g., seebeck coefficient)

Solar cell: ML model to predict fill factor, efficiency, V_{oc} , and J_{sc}

LitDX - IPO: Build ML model to predict properties of inorganic phosphoroptical properties (e.g., Emission max.)

Visualization Tasks



Thank you
for your attention



포럼



- 장소: 디딤돌플라자 강당
- 아침 이동: 온천교 704번
- 스트리밍 준비: 더에쓰시브이
- 다과: 케이터링 50인분
- 오전 세션 (장우선 교수 좌장)
- 점심: 성심당 도시락 40개 (교육장 48석)
- 오후 세션1 (손알로이시우스 교수 좌장)
- 오후 세션2 (김병현 교수 좌장)
- 저녁: 식당 추천?

- 5일 포럼 오후 행사 진행 서폿
- 포럼 점심 – 도시락 (40명 이상 참가시
구내식당 식권)
- 포럼 저녁 – 외부식당
- 포럼/해커톤 개회사 (원장님?)

- 클라우드 크레딧
- CPU서버

- 3명 한팀

해커톤



- 장소: S8-110 국립한밭대학교
- 아침 이동: 셔틀버스 8:28분 유성온천역 6번출구
- 행사장 준비: 더비엠씨 (박찬식)
- 다과: 유수현
- 개회사: 원장? 과기부?
- 오전
 - Fresh talk (인당 2분)
 - Project Intro (유수현 – 20분)
 - Teaming (30–60분)
- 점심: 도시락 (더비엠씨)
- 오후
 - Development
- 저녁: 한밭대 주변 한식당 (한우천국: 0507-1349-1140)
- 저녁 이동: 셔틀버스 18:10분 한밭대 주차장

- 장소: S8-110 국립한밭대학교
- 아침 이동: 셔틀버스 8:28분 유성온천역 6번출구
- 행사장 준비: 더비엠씨 (박찬식)
- 다과: 유수현
- 오전
 - Development
- 점심: 도시락 (더비엠씨)
- 오후
 - Development
- 저녁: 한밭대 주변 한식당 (더함뜰: 0507-1402-9294)
- 저녁 이동: 셔틀버스 18:10분 한밭대 주차장

- 장소: 디딤돌플라자 강당
- 아침 이동: 온천교 704번
- 행사장 준비: 더비엠씨 (박찬식)
- 다과: 케이터링
- 오전
 - Presentation & Evaluation
- Eval. Committee: Dr. Shin, Prof. Jang, Prof. Schrier
- 점심: 도시락 (더비엠씨), 3층 교육장
- 오후
 - 시상 (최우수: 아이패드 3개, 상장 / 우수 2팀: 에어팟프로 3개씩, 상장)
- 저녁 이동: 704번 한국화학연구원 정류장

```
[33]: import numpy as np
from scipy.optimize import curve_fit
import matplotlib.pyplot as plt
import pandas as pd

# Function to calculate the Gaussian with constants a, b, and c
def gaussian(x, a, b, c,d):
    return a*np.exp(-np.power(x - b, 2)/(2*np.power(c, 2)))+d

# ref: https://mathworld.wolfram.com/LorentzianFunction.html
def lorentzian(x,a,b,c,d):
    return 0.5*c*a/np.pi/(np.power(x - b,2)+np.power(0.5*c,2)) + d

rawdata=pd.read_csv('./화학연구원 과제_데이터 형식/cvd ws2 pl mapping data_pl-pos.txt', sep='\s+')

range_cutting_start=400
range_cutting_end=1200

raw_numpy=rawdata.to_numpy()
x=list(raw_numpy[:,0])

height=list()
center=list()
fullwidth=list()

pl_center_input1=630.0 # read from xls file
pl_center_input2=640.0
pl_cutoff_count_input=1000.0 # read from xls file

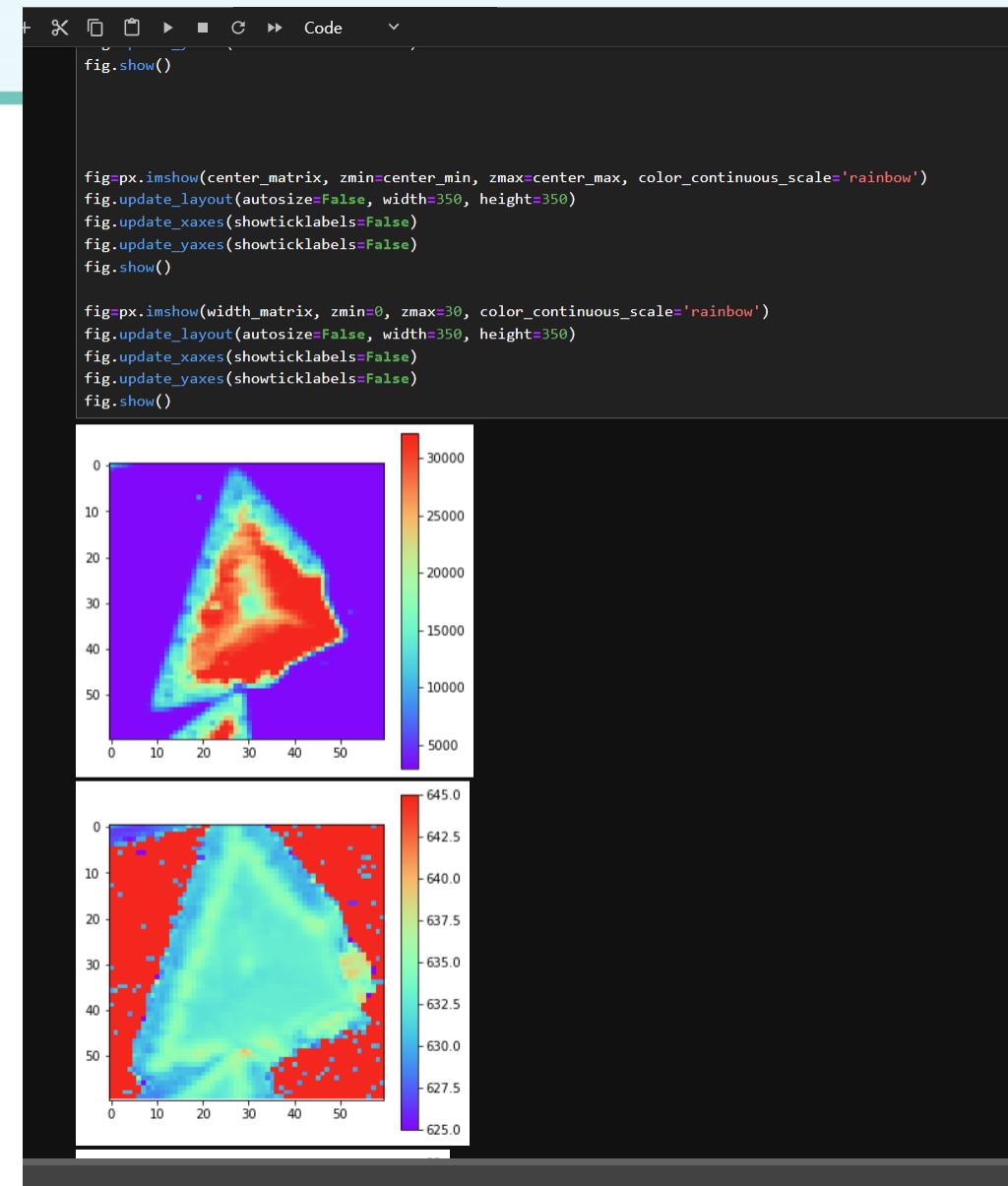
for i in range(1,rawdata.shape[1]):
    y=list(raw_numpy[:,i])

    pars, cov = curve_fit(f=lorentzian, xdata=x, \
                           ydata=y, \
                           p0=[10000, pl_center_input1, 10, pl_cutoff_count_input], \
                           bounds=(-np.inf, np.inf))

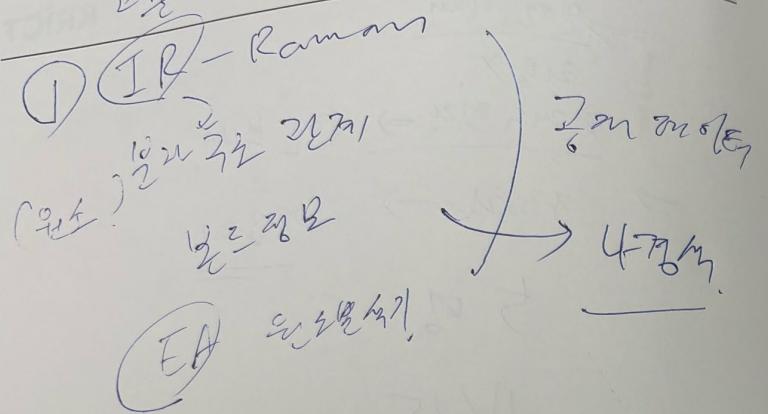
    height.append(2.0/np.pi*pars[2]*pars[0]+pars[3])
    center.append(pars[1])
    fullwidth.append(pars[2])

    #print(pars)
    #print(cov)
    #print(_lorentzian(x[range_cutting_start:range_cutting_end],1000.0, pl_center_input1, 30.0, pl_cutoff_count_input))

    plt.plot(x[range_cutting_start:range_cutting_end],y[range_cutting_start:range_cutting_end],'bo')
    plt.plot(x[range_cutting_start:range_cutting_end],lorentzian(y[range_cutting_start:range_cutting_end],x[range_cutting_start:range_cutting_end],pars[1],pars[2],pars[3]),'r')
    plt.show()
```



2월 21 DB (분석설명)



EPS.
⇒ 분석설명 / 분석.

②

Digital

web → Jupyter
HTML → Jupyter
Lit DX. → 예술로
하기 위해서 → Visualiz
modeling
분석 DOJ.

EXP DX
Dataset
→ Smiles 분석 방법
Spectra (예술 작품)
분석 방법

* SBD
기술 / 분석
기술 분석
기술 분석
기술 분석
Spectrum
고급 분석/EI
구조 분석

erial.
12월 3일

DF1
100M.

ing face