A Unified Active Learning Framework for Designing Energy-Relevant Molecules

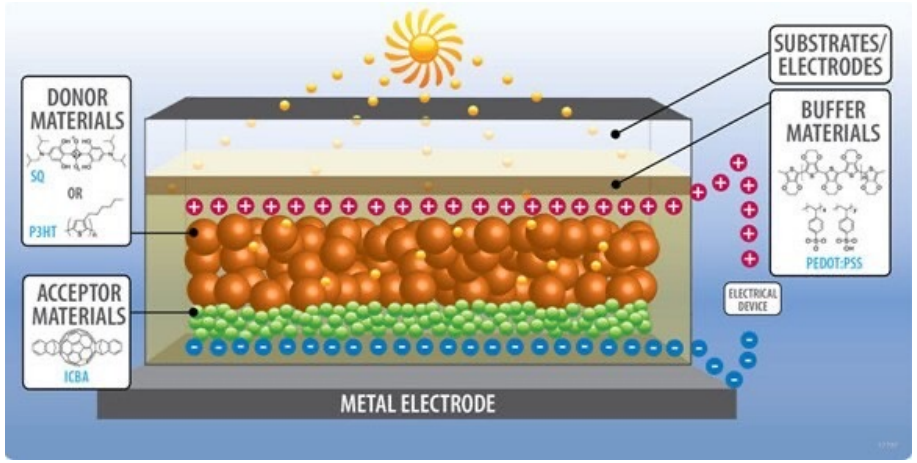
Shomik Verma[1], Jiali Li[2], Kevin Greenman[1],
Rafael Gomez-Bombarelli[1], Xiaonan Wang[2,3], Aron Walsh[4]

[1]Massachusetts Institute of Technology, [2]National University of Singapore,
[3]Tsinghua University, [4]Imperial College London

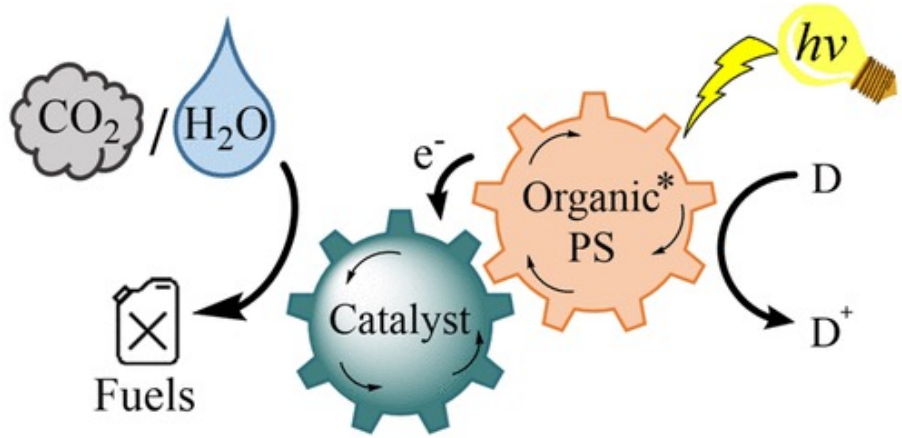
Massachusetts Institute of Technology

Molecules relevant to energy applications

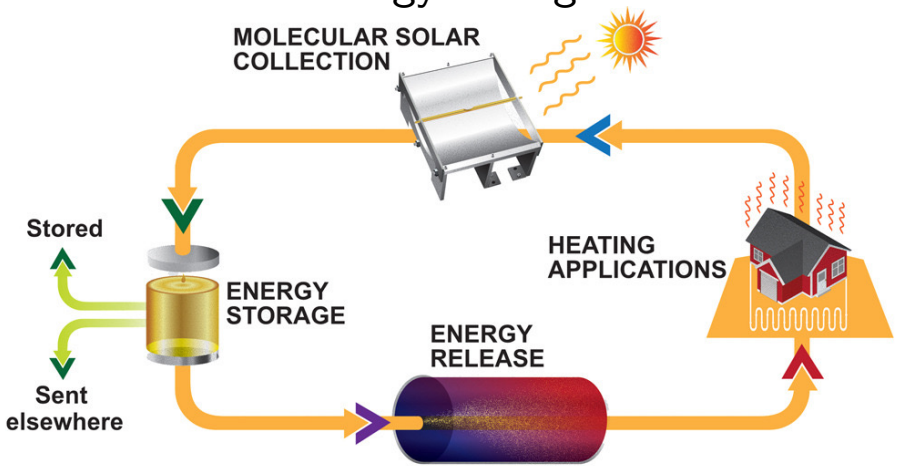
Organic photovoltaics



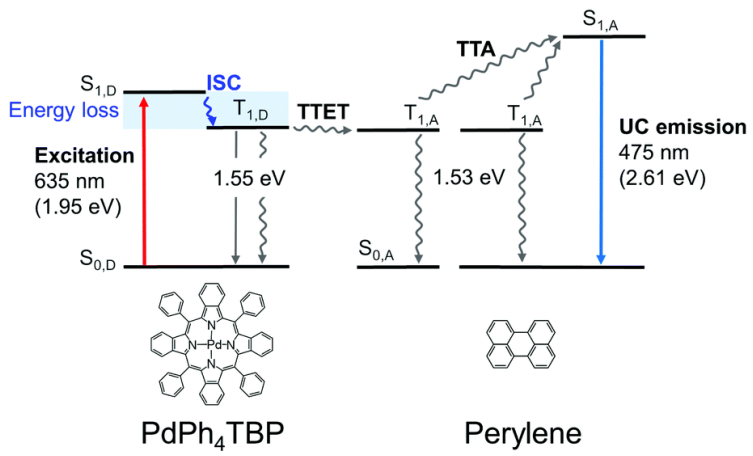
Solar fuels



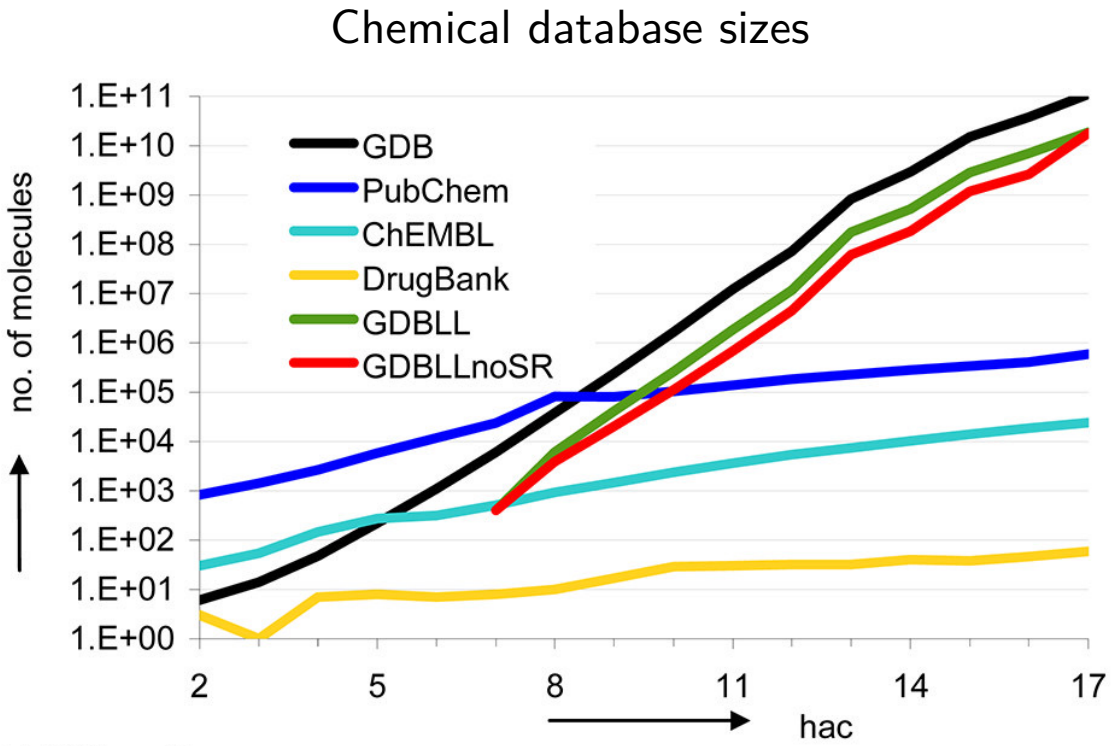
Energy storage

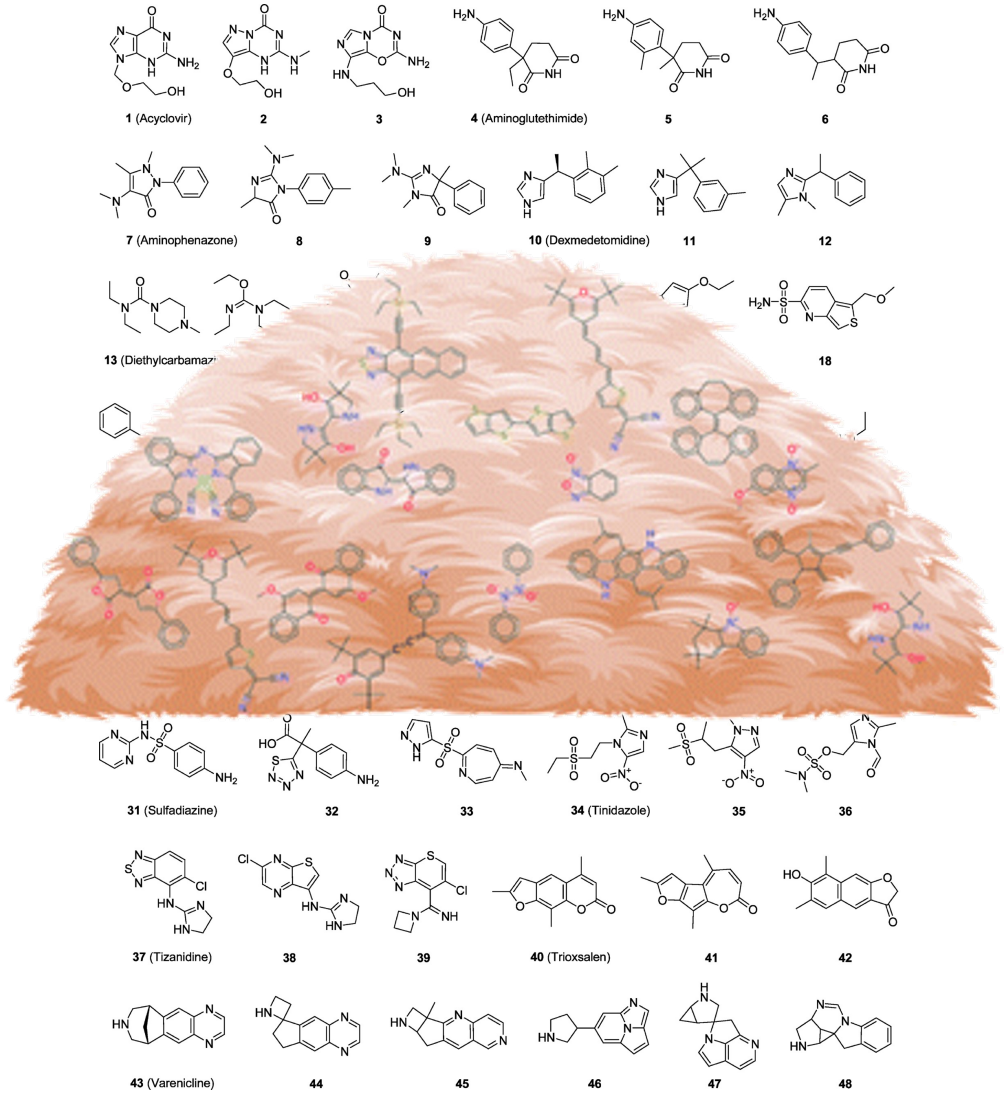


Photon conversion



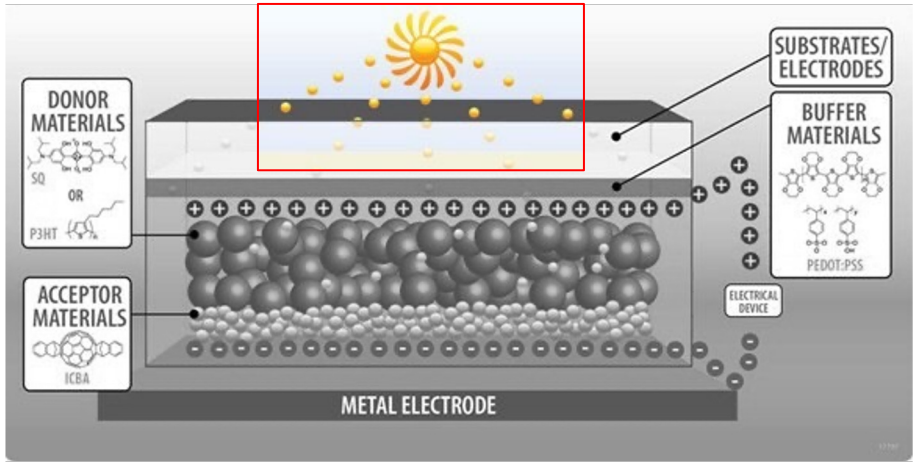
Molecular space is massive



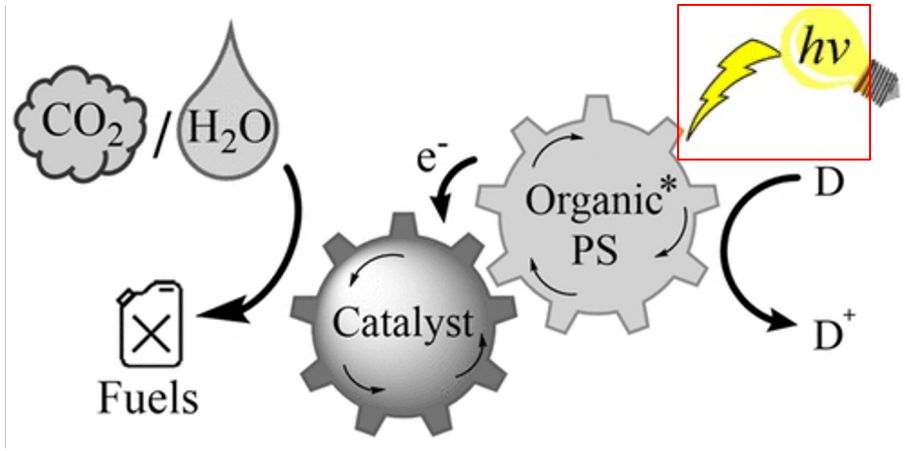


Light interactions = excited state energies required

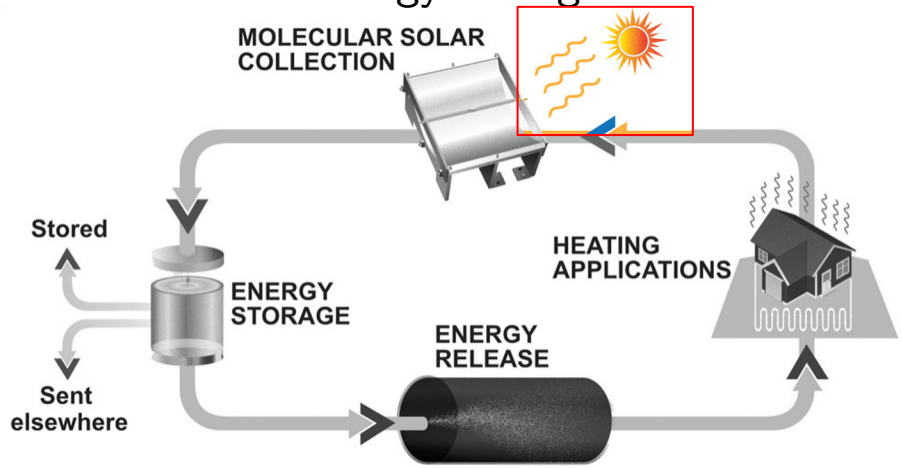
Organic photovoltaics



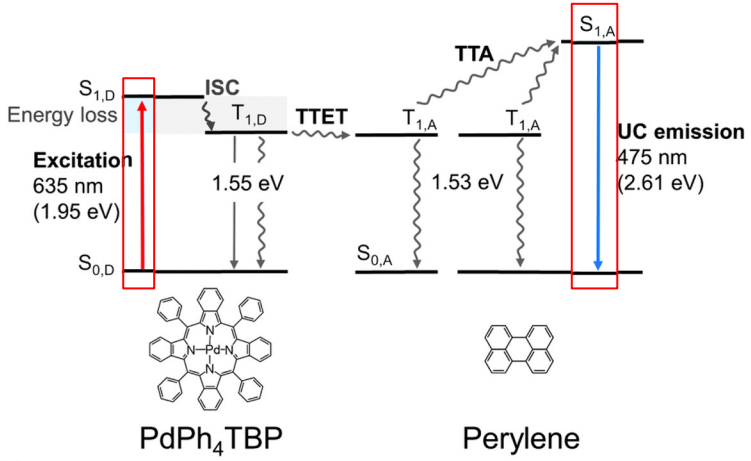
Solar fuels



Energy storage

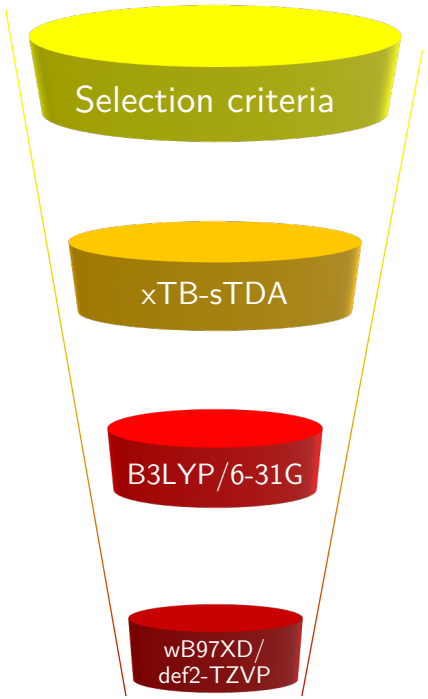


Photon conversion



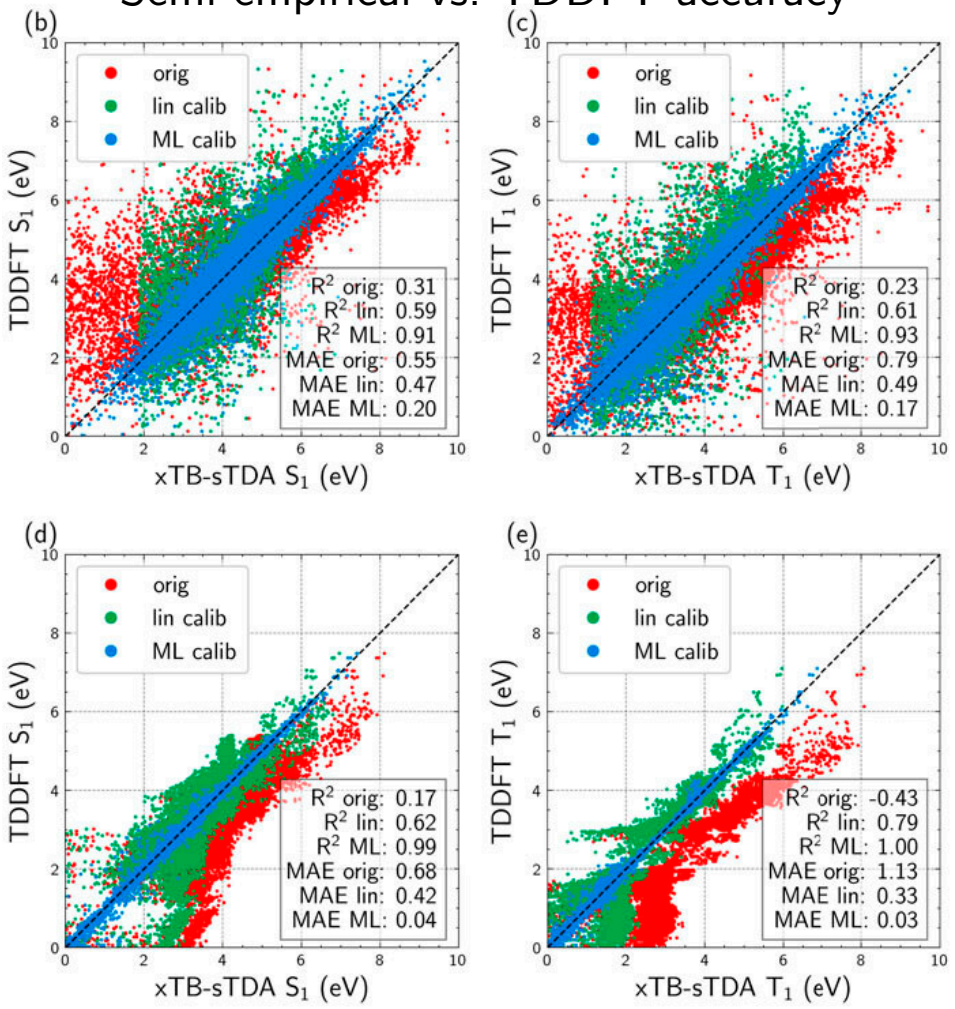
Fast and accurate chemical screening

Molecular dataset



Filtered molecules

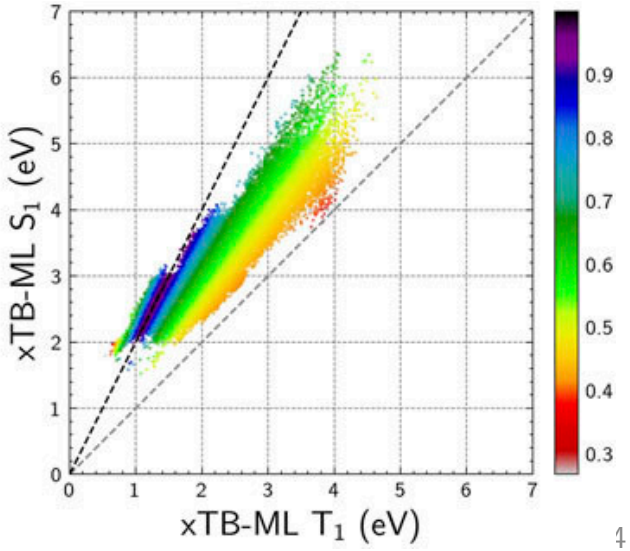
Semi-empirical vs. TDDFT accuracy



Calibration results

	xTB-sTDA T_1	xTB-ML-20k Class 1 T_1	xTB-ML-300k Class 1 T_1	xTB-ML-20k Class 2 T_1	xTB-ML-300k Class 2 T_1
MOPSSAM (143)	0.59	0.17	0.13	0.10	0.08
MOPSSAM (1k)	0.39	0.19	0.15	0.14	0.12
INDT (10k)	0.40	0.13	0.15	0.60	0.38
Verde (1.5k)	0.48	0.19	0.22	0.26	0.29

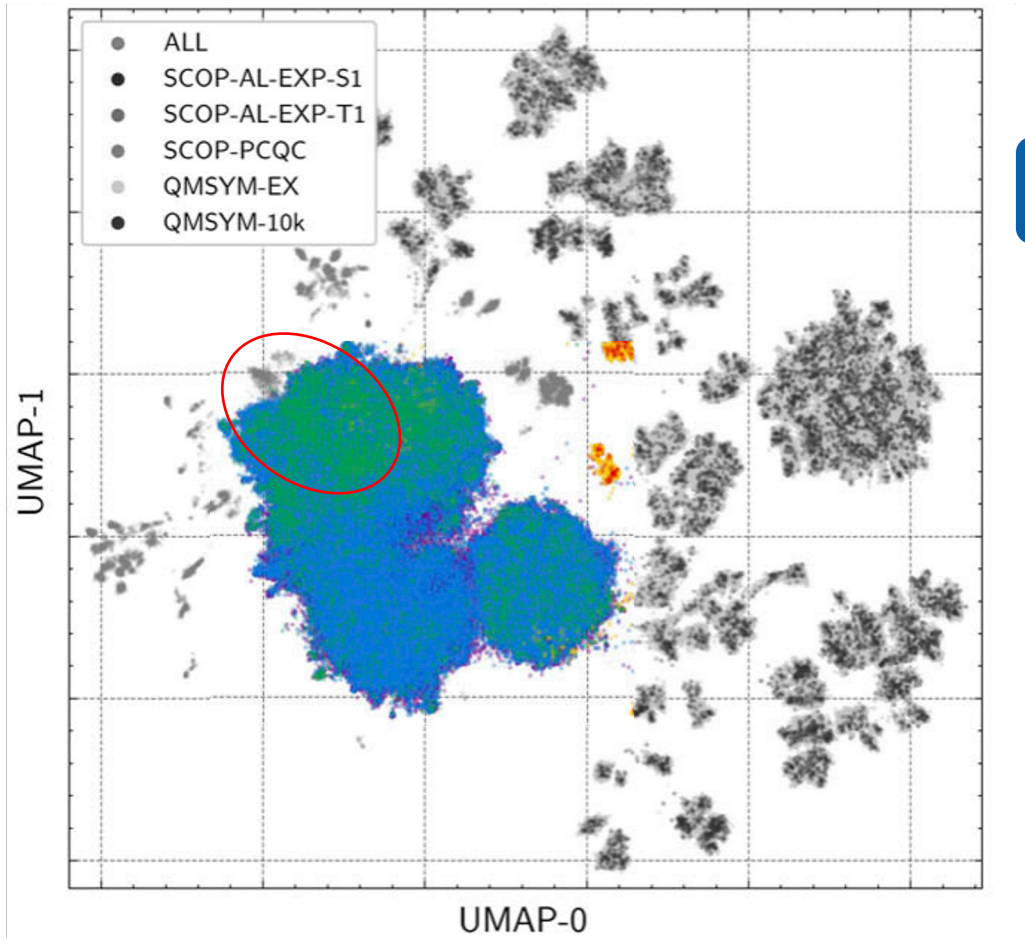
HTVS for photon conversion molecules



xTB-ML

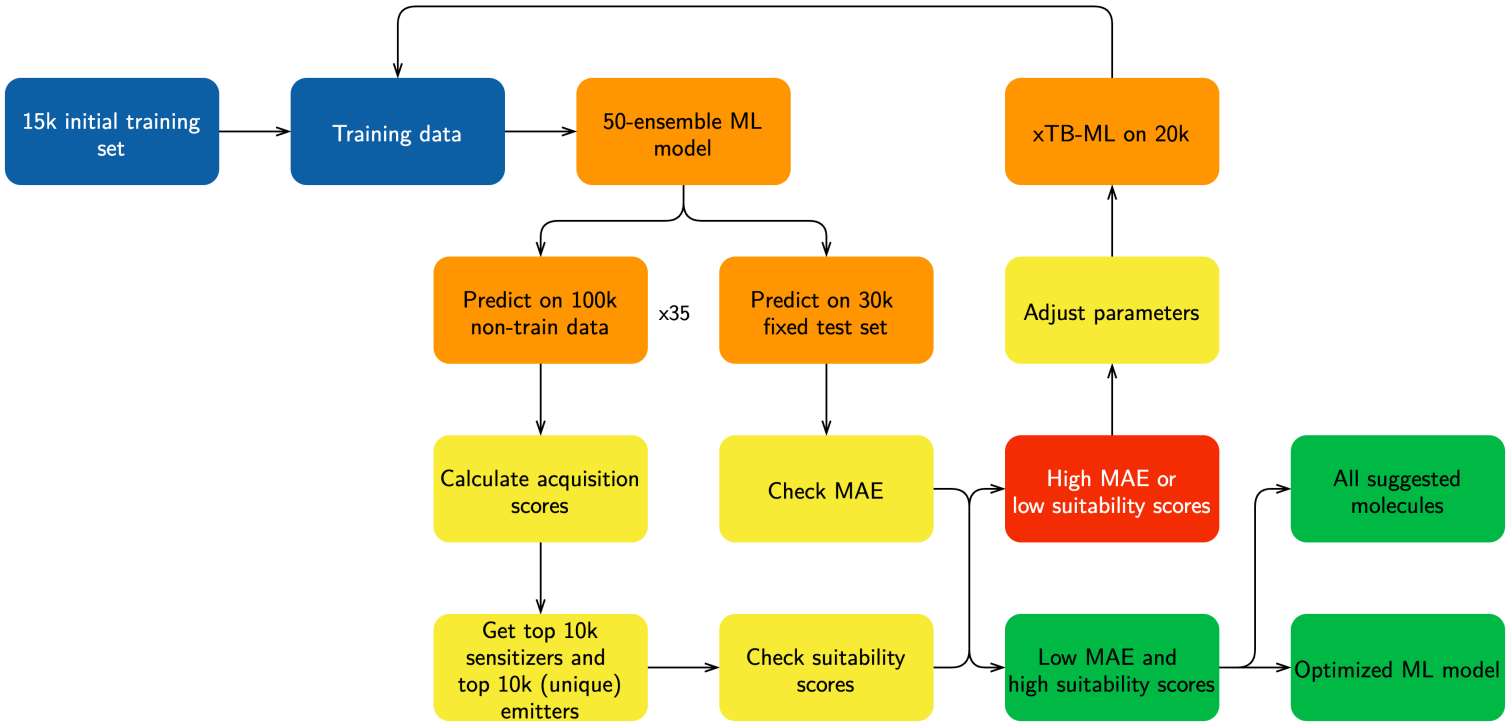
Efficient exploration with active learning

Global chemical space map



250k HTVS space

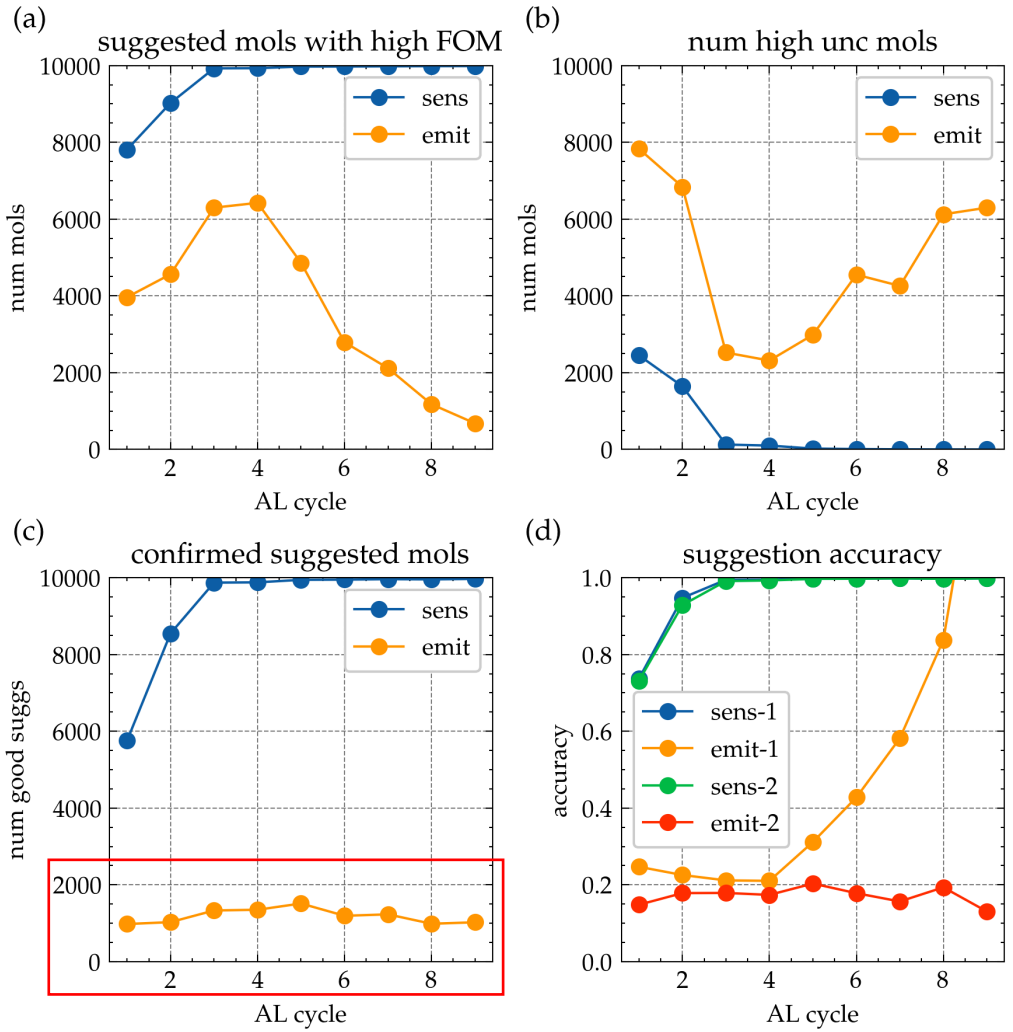
AL workflow

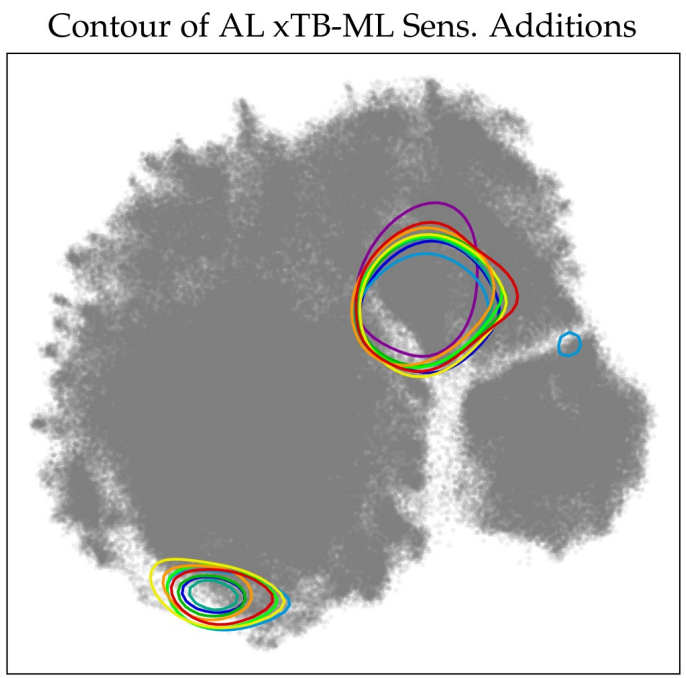


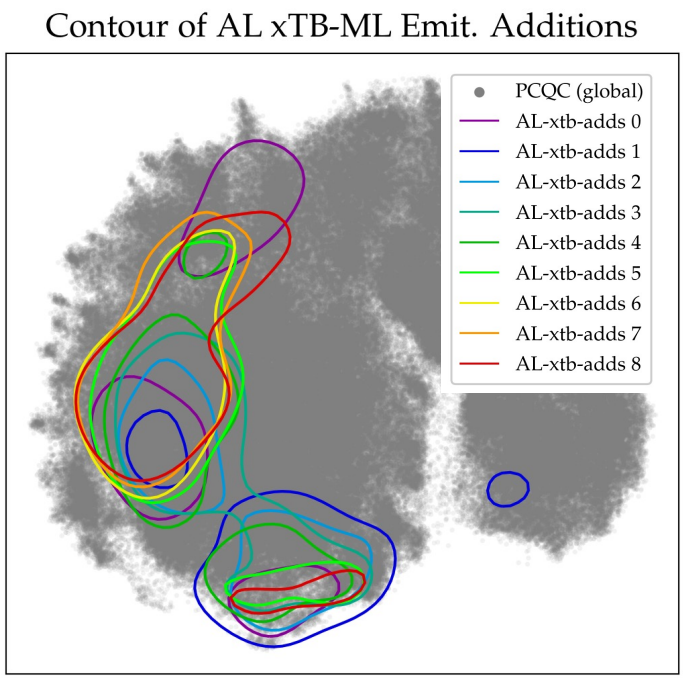
$$\alpha_{sens} = B \cdot \varepsilon_{sens} + \sigma_{T1} + \sigma_{S1}$$

$$\alpha_{emit} = B \cdot \varepsilon_{emit} + \sigma_{T1} + \sigma_{S1}$$

AL results on small molecule database







Low AL accuracy for emitters

How to improve results?

1

Expand database

2

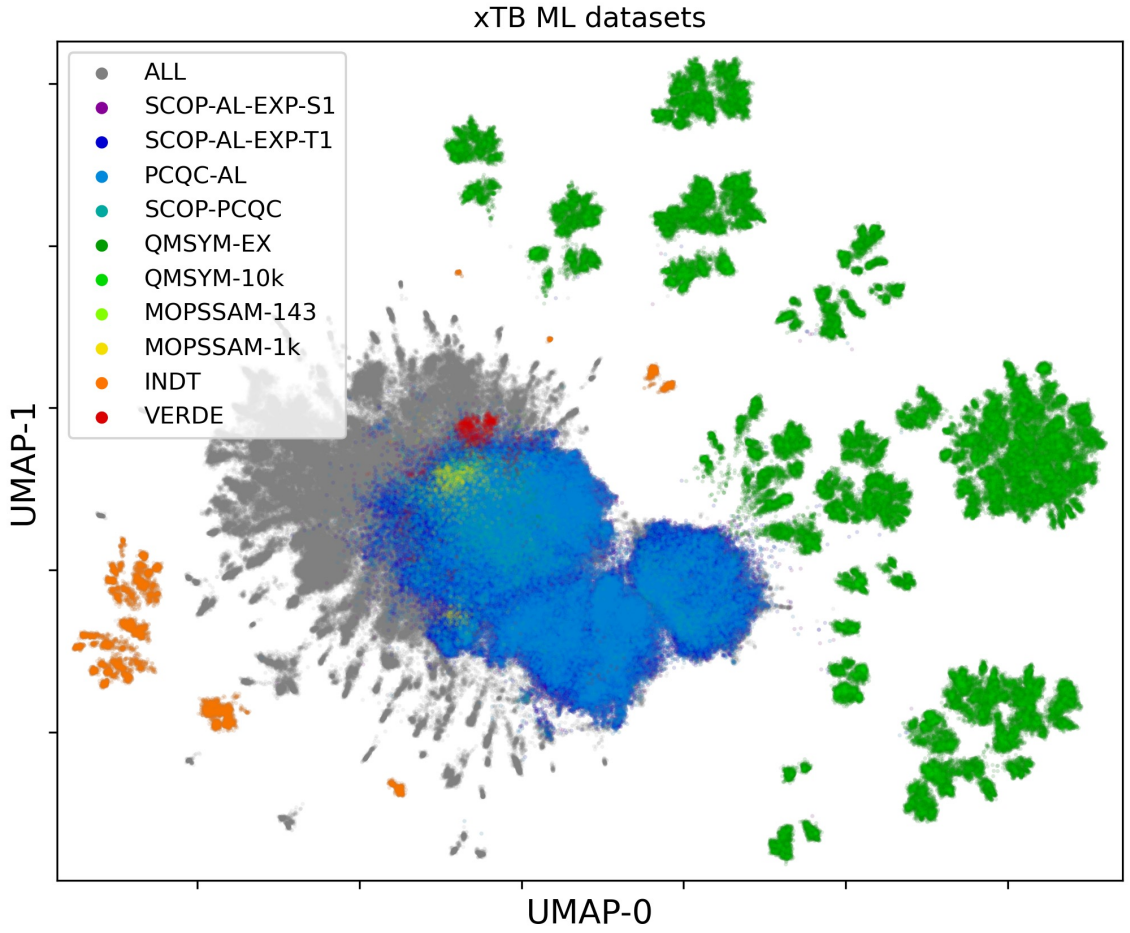
Improve AL framework

3

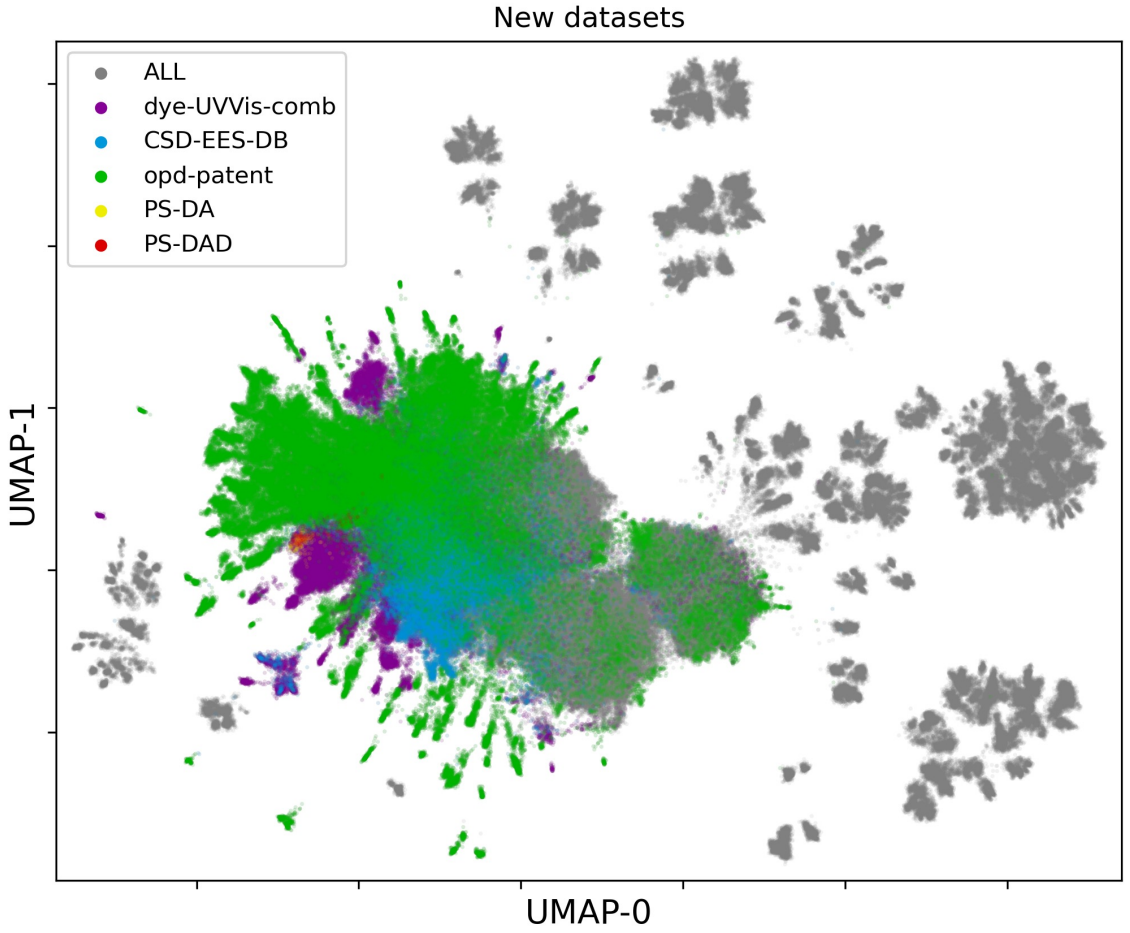
Implement generative ML

Expanding to large molecules

(a)



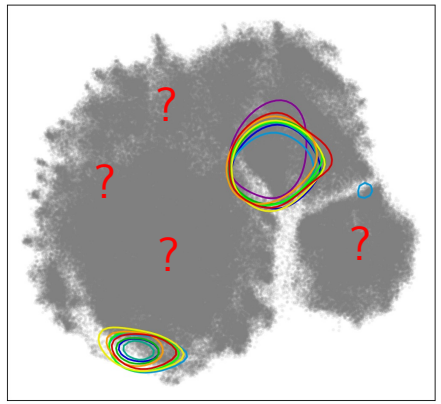
(b)



Improving the AL framework

Uncertainty

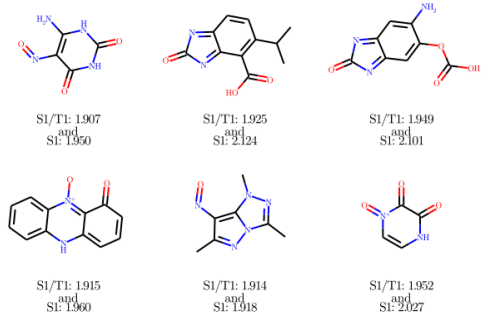
Training data locations



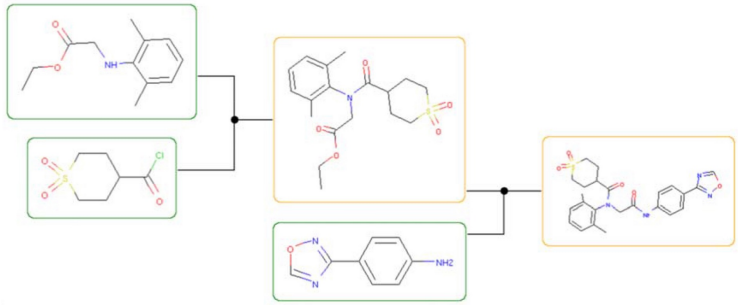
$$\sigma_{T_1} = \frac{\sum_{i=1}^{n}(T_{1_i} - T_{1_{mean}})^2}{n}$$

$$\sigma_{S_1} = \frac{\sum_{i=1}^{n}(S_{1_i} - S_{1_{mean}})^2}{n}$$

Domain Knowledge / Synthesizability

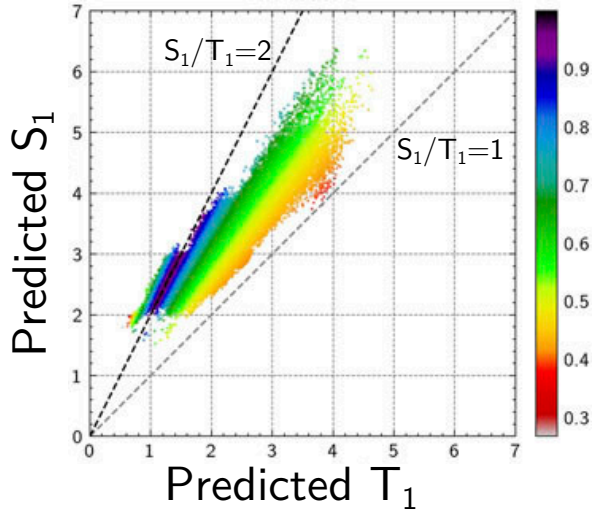


Known sens/emit



AiZynth (retrosynthetic planning)

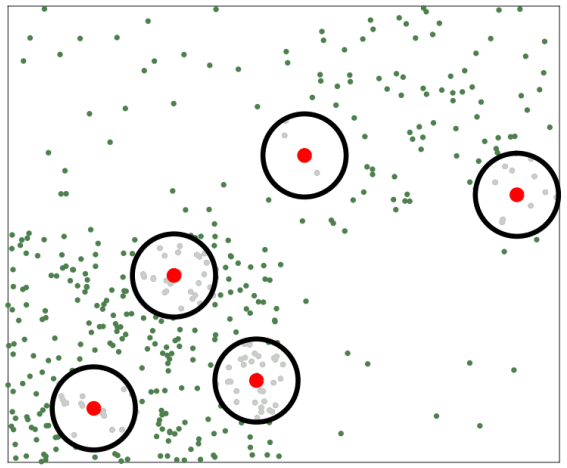
Target Property



$$\varepsilon_{sens} = e^{-A\left(1-\frac{S_1}{T_1}\right)}$$

$$\varepsilon_{emit} = e^{-A\left(2-\frac{S_1}{T_1}\right)}$$

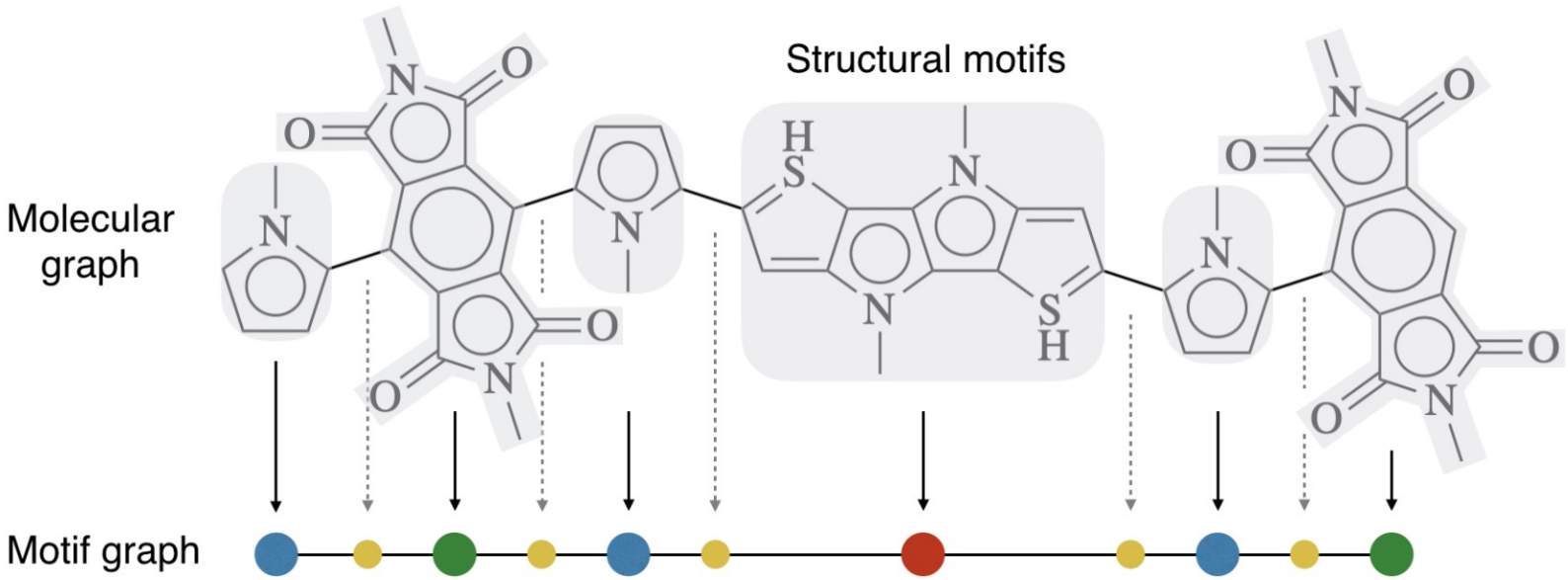
Batch diversification



$$r_{new} - r_i > \theta, \forall i$$

Generative models

Large molecules → Motif Based Generative Models



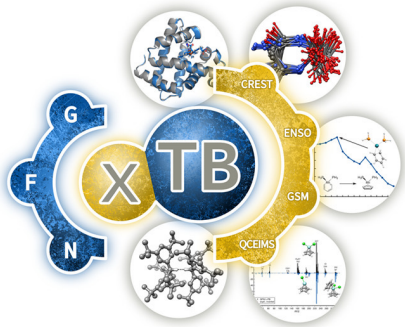
Create a design space generative model based on representative molecules pool

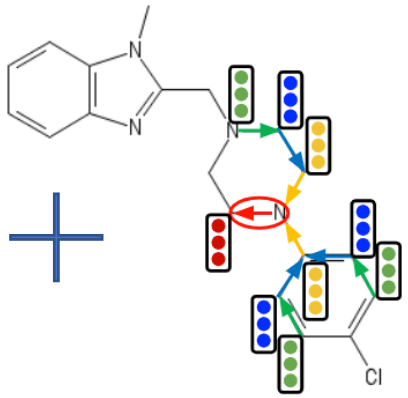
↓

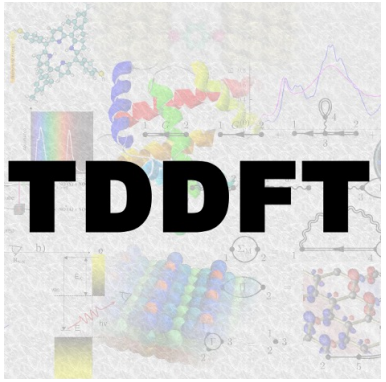
Sample new molecules according to different generative strategies

Generative Model

Summary







Method 1: xTB+ML

Design Space Selection

Expert Based Design Space Pool

Direct Generation with Surrogate Model

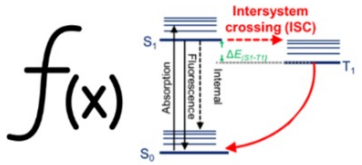
Acquisition Function Formulation

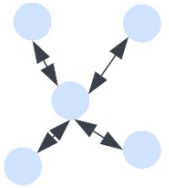
Property Part

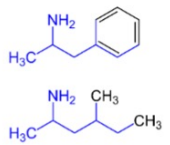
Uncertainty Part

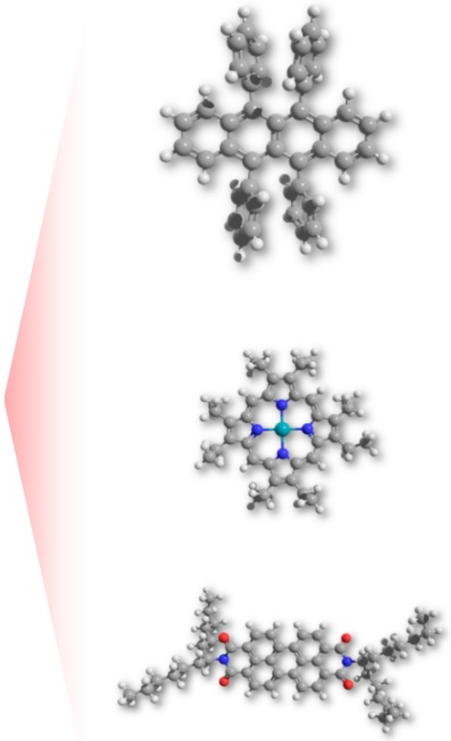
Diversity Part

Domain Kownledge Part









Method 2: Unified AL

Thanks!

Questions?

Massachusetts Institute of Technology

Imperial College London

NUS National University of Singapore