Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Generative Molecular Design (GMD) for AI-Based Drug/Material Discovery

Nathan Urban and Byung-Jun Yoon

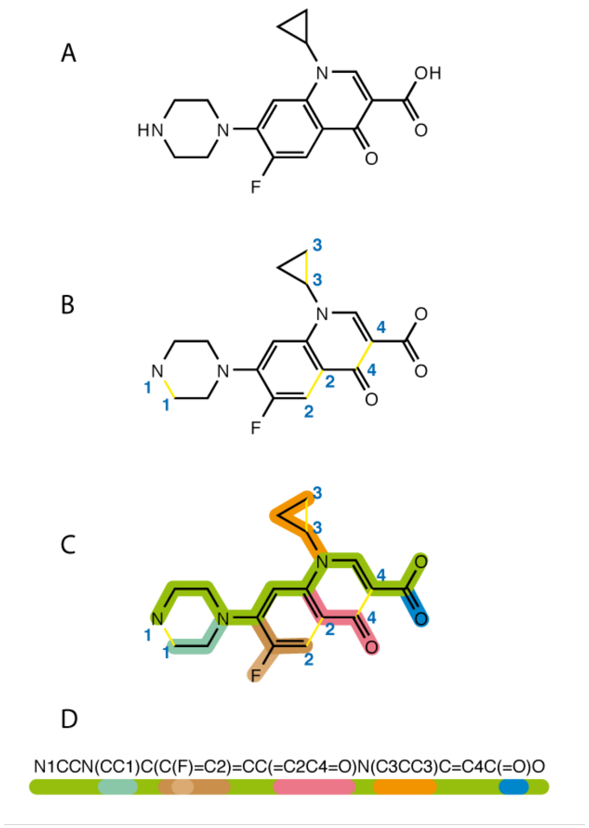
May 7, 2024

@BrookhavenLab

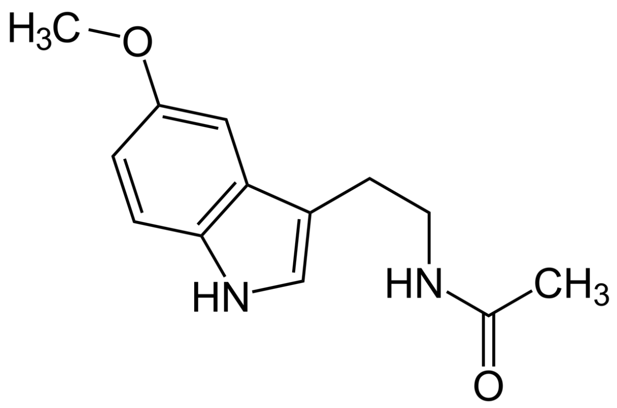
Small Molecule Design

SMILES

(Simplified molecular-input line-entry system)



Melatonin (C13H16N2O2)



```
CC(=O)NCCC1=CNc2c1cc(OC)cc2
```

Question:

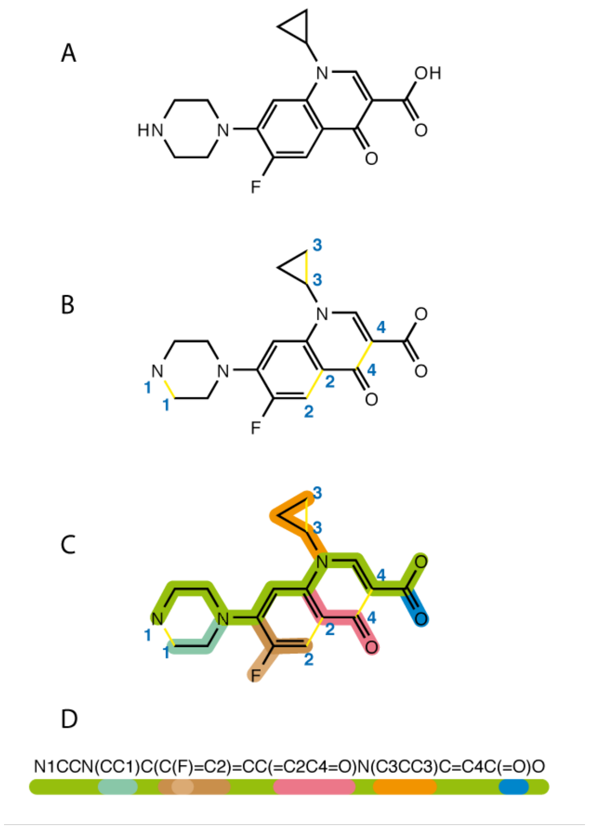
How to generate molecules with desired properties?

Image source: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

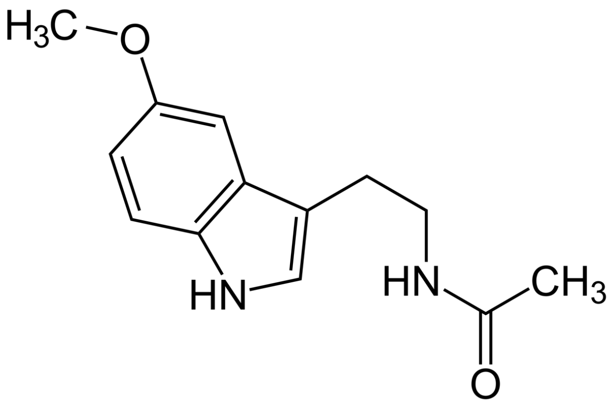
Small Molecule Design

SMILES

(Simplified molecular-input line-entry system)



Melatonin ($C13H16N2O2$)



```
CC(=O)NCCC1=CNc2c1cc(OC)cc2
```

Challenges:

Huge design space that is **discrete** and **high-dimensional**

Image source: https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

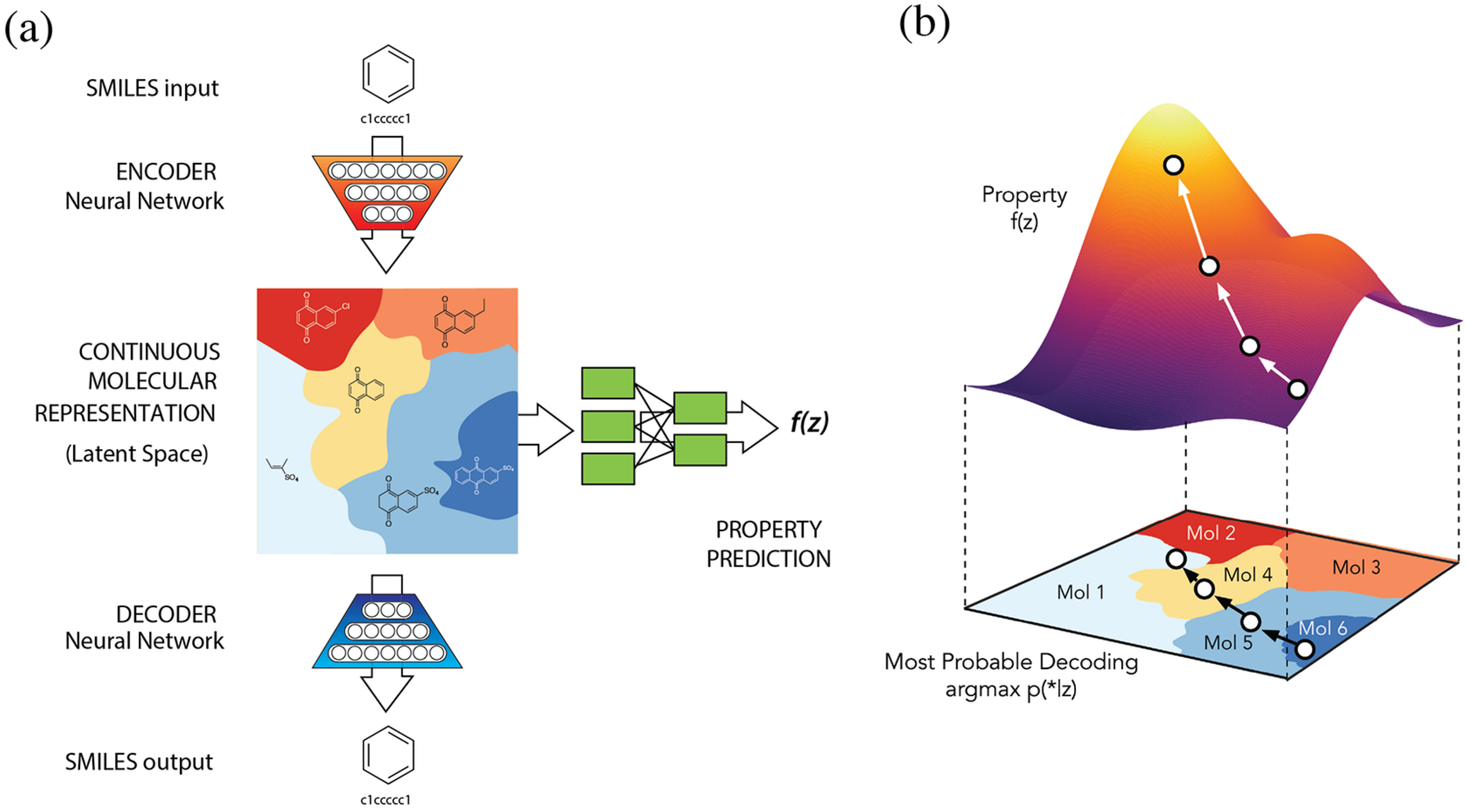
Generative Design

- How to generate new (molecules, materials, …) from an extremely high-dimensional (and discrete) space?
- **Generative design loop**
  - *Collect* a library of existing designs
  - *Train* a ML model to generate more designs “like them”
    - New molecules, material structures, etc.
    - (may or may not be better than library designs, but different)
  - *Screen* each generated design to see if it is better
  - *Adjust* the generator to preferentially suggest higher-quality designs

Training a Generator

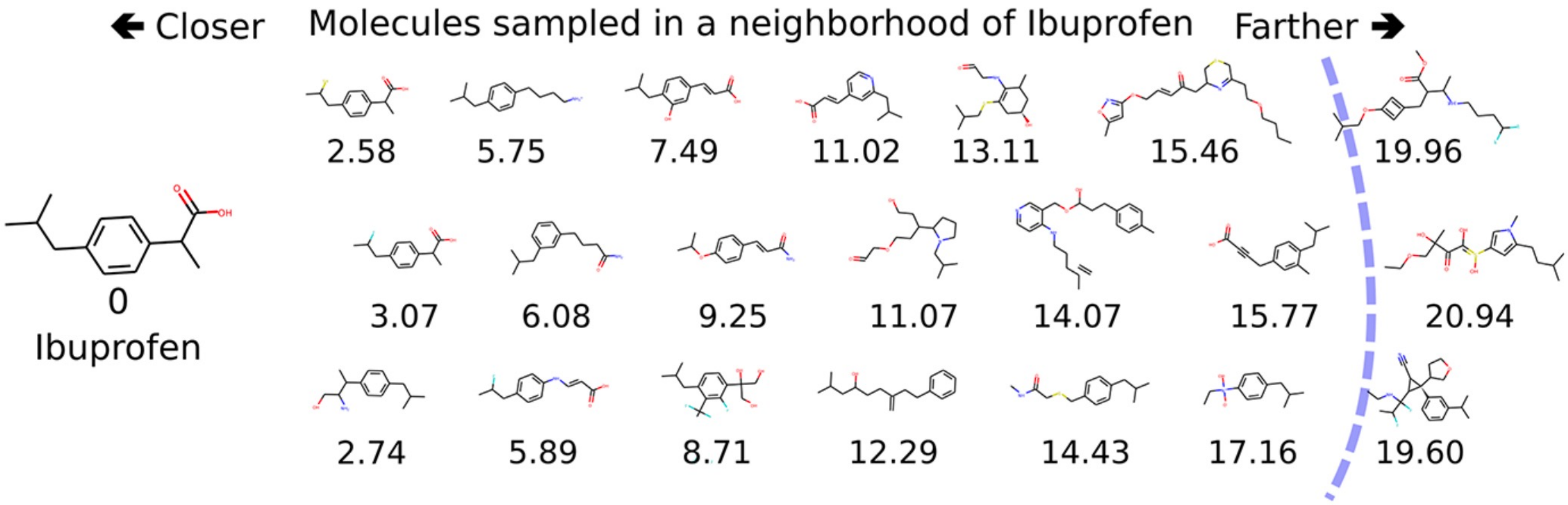
- Mathematically each design has a *representation* (e.g., encoding of a molecular structure)
- The design library becomes a set of *points* in this representation space
- A generative AI model learns to sample this distribution

Generative Molecular Design (GMD) Using A Variational Autoencoder (VAE)



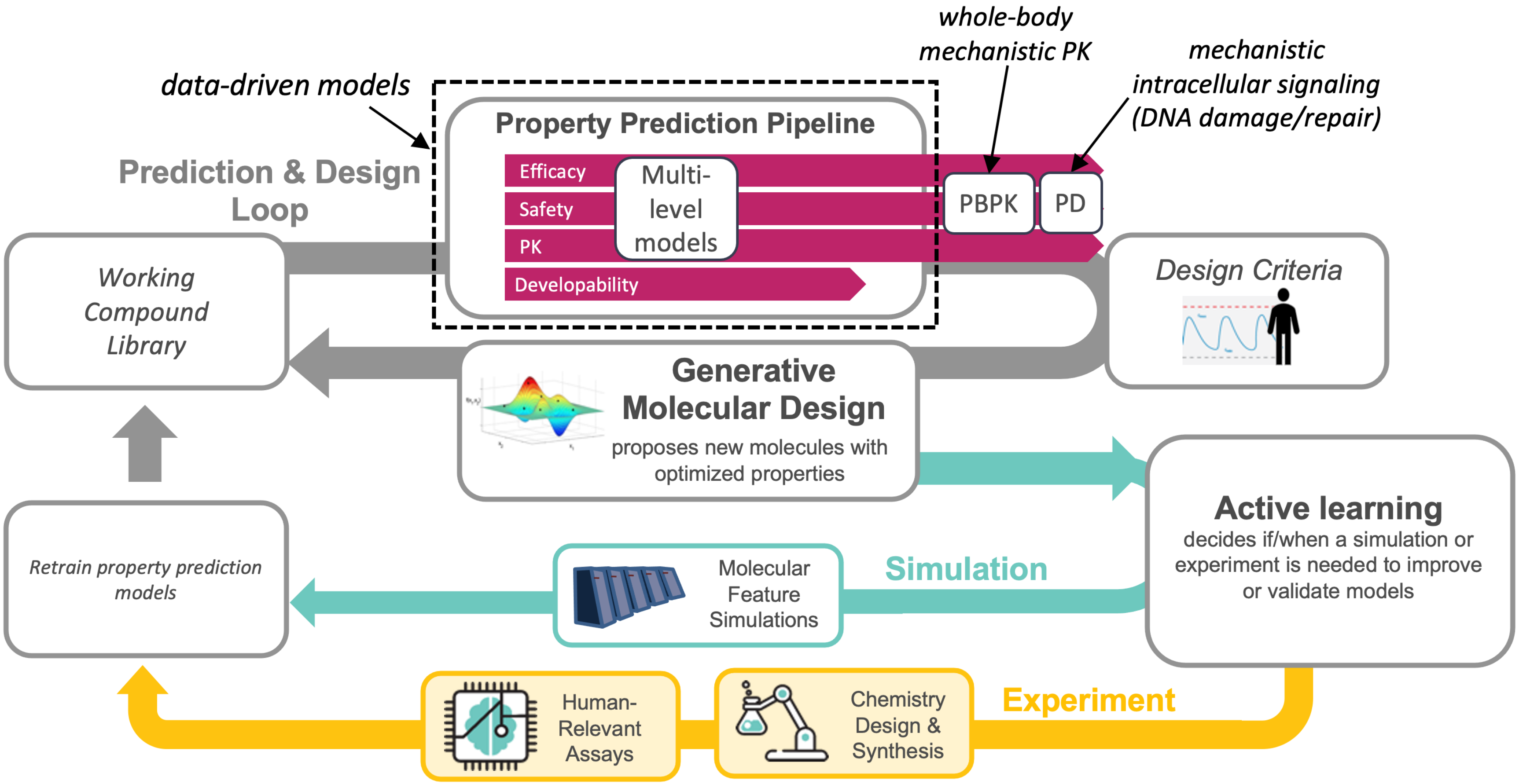
Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS central science* 4.2 (2018): 268-276.

Generative AI has been having huge impact on small molecule design



Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.

ATOM GMD | Accelerating Therapeutics for Opportunities in Medicine



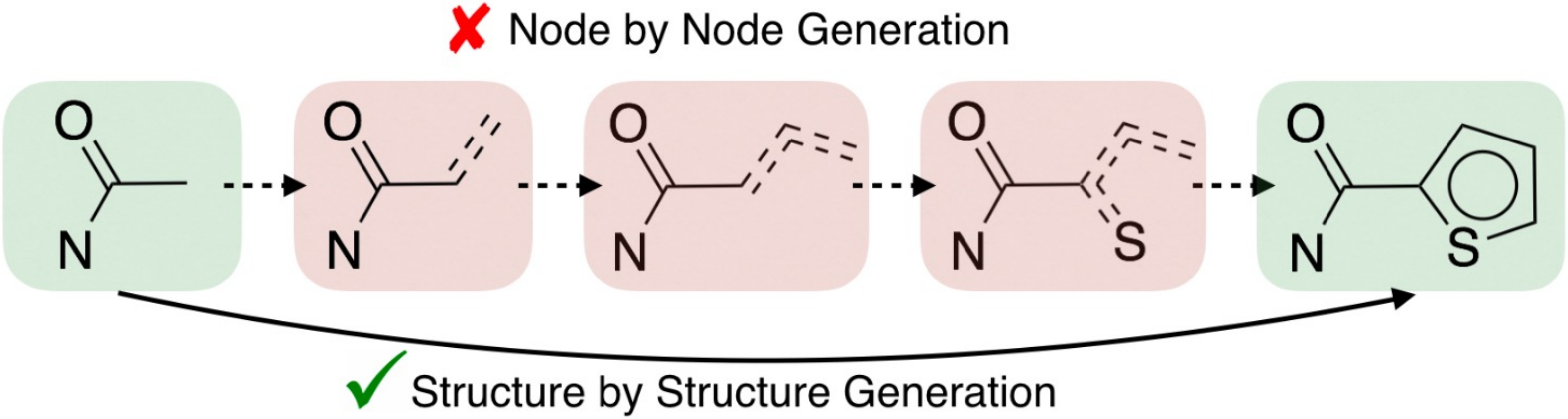
data-driven models
Property Prediction Pipeline
Efficacy
Safety
PK
Developability
Multi-level models
whole-body mechanistic PK
mechanistic intracellular signaling (DNA damage/repair)
PBPK
PD
Prediction & Design Loop
Working Compound Library
Design Criteria
Generative Molecular Design
proposes new molecules with optimized properties
Active learning
decides if/when a simulation or experiment is needed to improve or validate models
Retrain property prediction models
Molecular Feature Simulations
Simulation
Human-Relevant Assays
Chemistry Design & Synthesis
Experiment

Brookhaven National Laboratory

A note on the ATOM GMD loop

- The VAE works by mapping the high-dimensional representation of a molecule into a low-dimensional latent space
- However, ATOM pipeline doesn’t actually use VAE generatively to sample new molecules
  - It only uses the latent space
- To generate new molecules, it uses a genetic algorithm to propose and improve candidate designs within the latent space
  - VAE is used as dimension reduction of design space for GA
- However, versions of GMD where the generative AI samples new molecules also exist

Structure preservation: Junction Tree Variational Autoencoder


✗ Node by Node Generation
O
N
S
✓ Structure by Structure Generation

Naïve generative models can propose candidates that decode to invalid molecules

Structure-by-structure graph generation approach is preferred as it avoids invalid intermediate states (marked in red) encountered in node-by-node approach (exploits graph structure of molecules)

Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." International conference on machine learning. PMLR, 2018.

Some Practical Questions

1. **How can we extend the capability of a generative model** for suggesting novel molecules with enhanced properties that go **beyond the initial training data**?

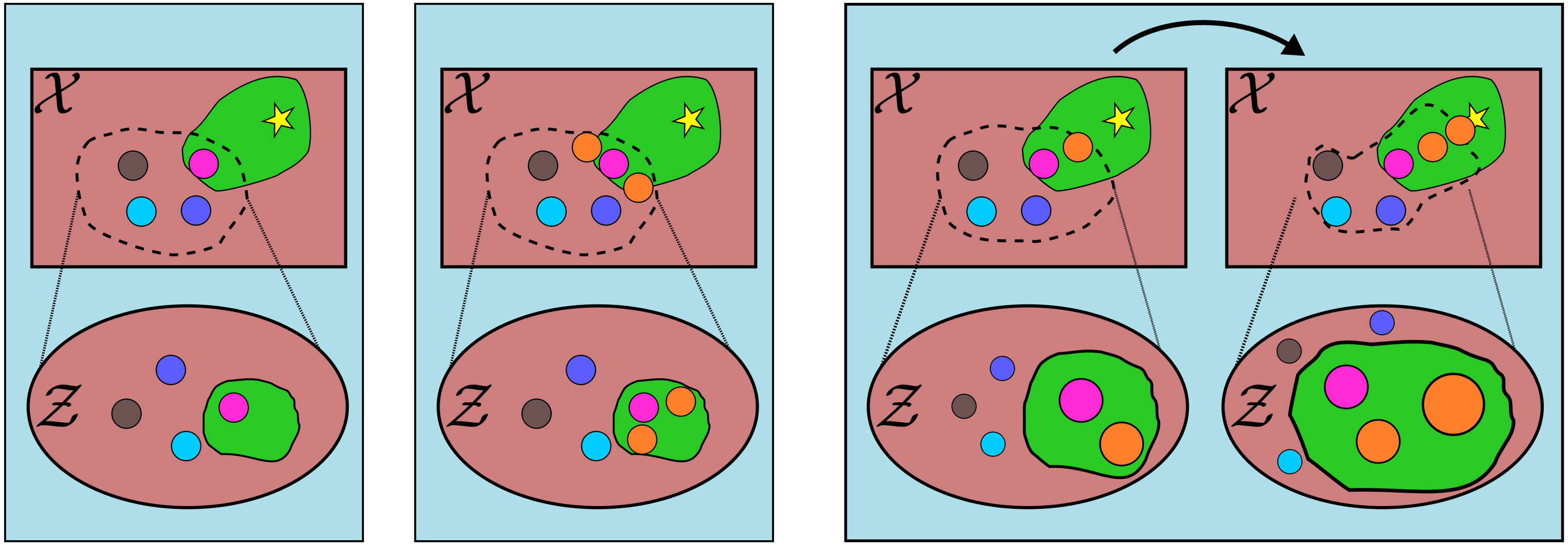
2. Considering that the initial training dataset is typically huge, **how can we augment the dataset** such that it can effectively **steer the model towards** molecules with more **desirable properties**?

3. How can we **improve "data-driven" generative models** by taking advantage of other **mechanistic models** (e.g., pathway models)

4. How to incorporate **uncertainty quantification** for large ML models with huge parameter spaces, and use that to guide exploration?

Brookhaven
National Laboratory

Extending GMD Capability | Latent Space Optimization (LSO)



Starting point

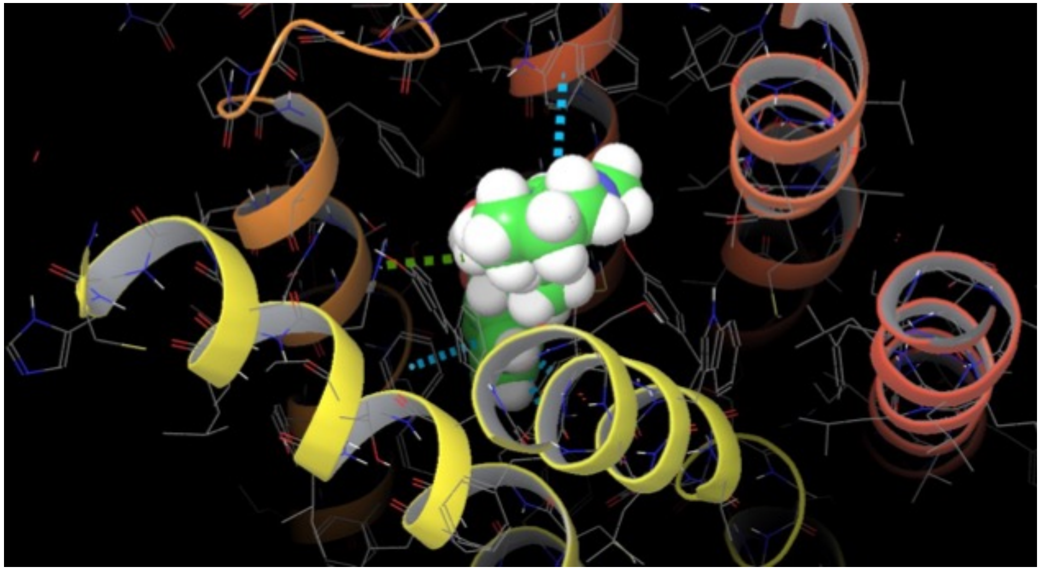
Standard LSO

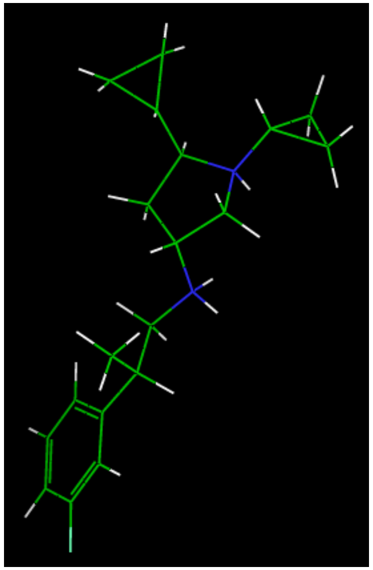
LSO with weighted retraining

As we learn more about which candidates are good, iteratively retrain the generator to preferentially suggest good candidates.

Austin Tripp, Erik Daxberger, and José Miguel Hernández-Lobato. Sample-efficient optimization in the latent space of deep generative models via weighted retraining, 2020.

Examples | DRD2 inhibitors





- Molecules designed based on the optimized GMD was able to compete molecules generated by QSAR (Quantitative structure-activity relationship) for binding and various properties

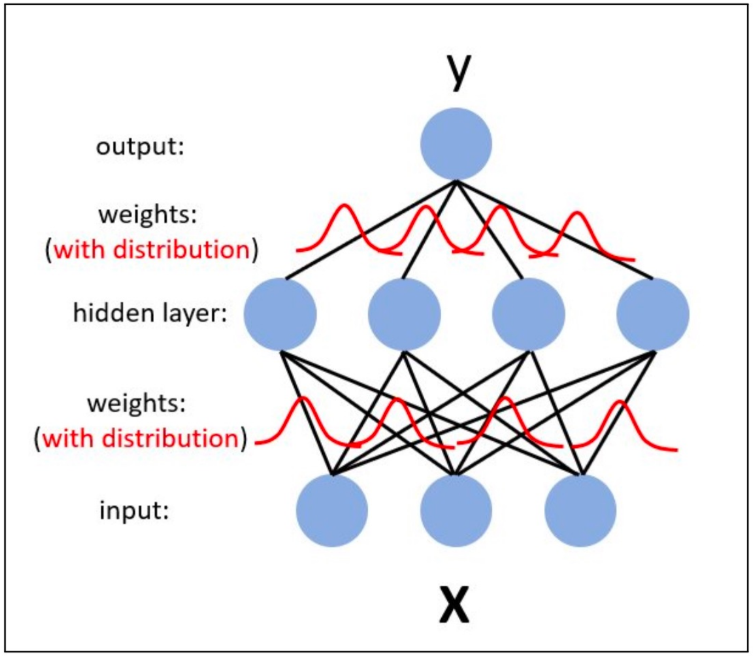
- In this case, although GMD-generated molecules are not structurally aware of the target, they score better than co-crystalized ligand and long duration MD shows very stable binding

Brookhaven National Laboratory

Enabling Effective UQ & OED

How can we enable effective UQ and OED / active learning for deep neural networks?

- **Plain Feedforward Neural Networks (frequentist approach):**
  (1) tend to overfit
  (2) incapable of quantifying training data uncertainty
  (3) make overly confident decisions
- **Bayesian Neural Networks (BNN):**
  (1) improved predictions
  (2) reliable uncertainty estimates
  (3) principled model comparison
  (4) support decision-making under uncertainty



Brookhaven National Laboratory

Summary

1. **GMD enables efficient search for novel molecules** with desired properties

2. **Various ML models have been proposed**, where VAE-type of models have been especially popular

3. **Diverse techniques have been introduced to fine-tune / optimize** generative models for specific downstream tasks

4. **Further research is needed for effective (Bayesian) UQ and OED techniques** for such generative models