Towards inverse design in chemistry: from prediction to deep generative models

A few months before:

Prof. Alán Aspuru-Guzik

PhD Advisor

Matter lab

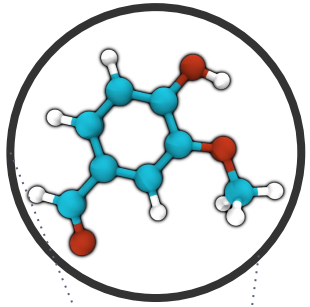
Benjamin Sanchez-Lengeling

@ AISIS 2019 CDMX

Research Scientist @

Google AI

One example with Vanillin



Material need

TIME

WORLD • MADAGASCAR

Vanilla Is Nearly as Expensive as Silver. That Spells Trouble for Madagascar

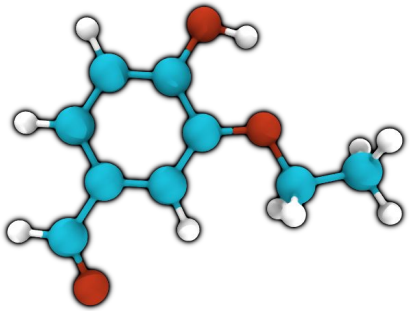
BY ARYN BAKER/SAHABEVAVA UPDATED: JUNE 13, 2018 3:15 PM ET

Can we find a substitute Vanillin that might be more sustainable?

Predicting functionality

What does Vanillin smell and taste like?

“vanilla”, "sweet", "creamy" and "chocolatey"

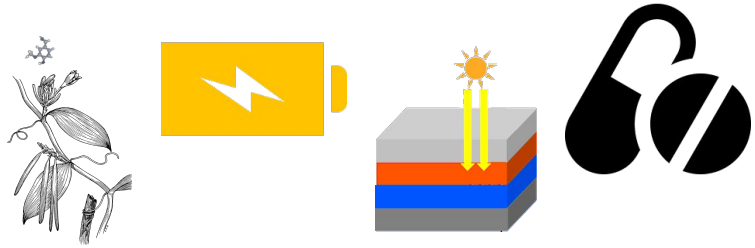


Generation based on functionality

How can we find more molecules like Vanillin?

In a broader context

Material needs



Predicting functionality

odor precepts, redox potential, solar cell efficiency, binding affinities

Generation based on functionality

This is an inverse design problem

Inverse design in chemistry

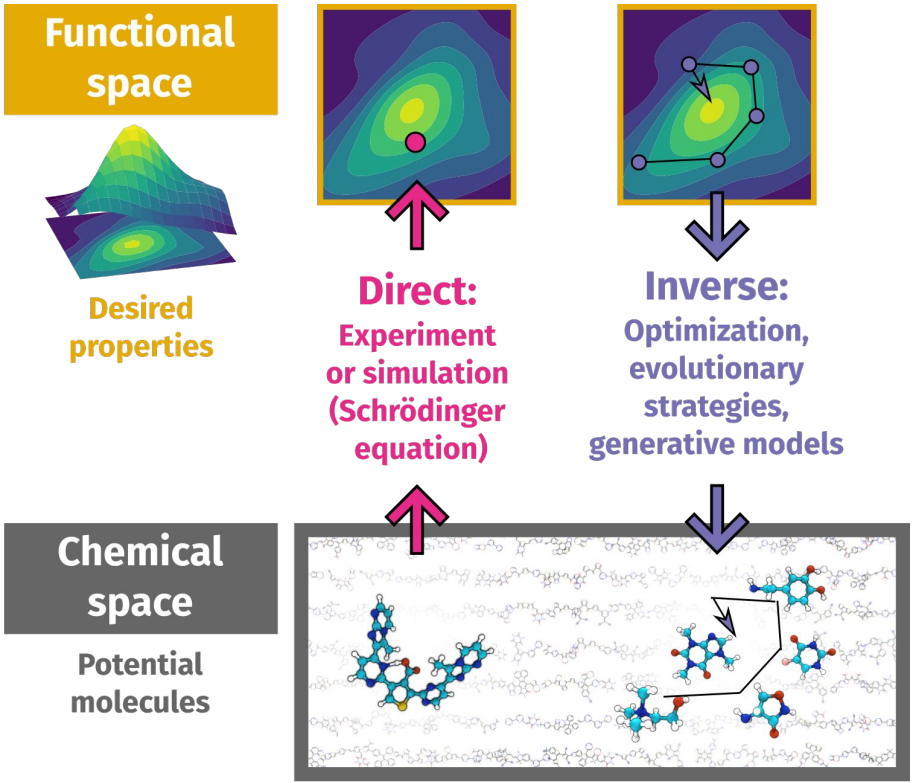
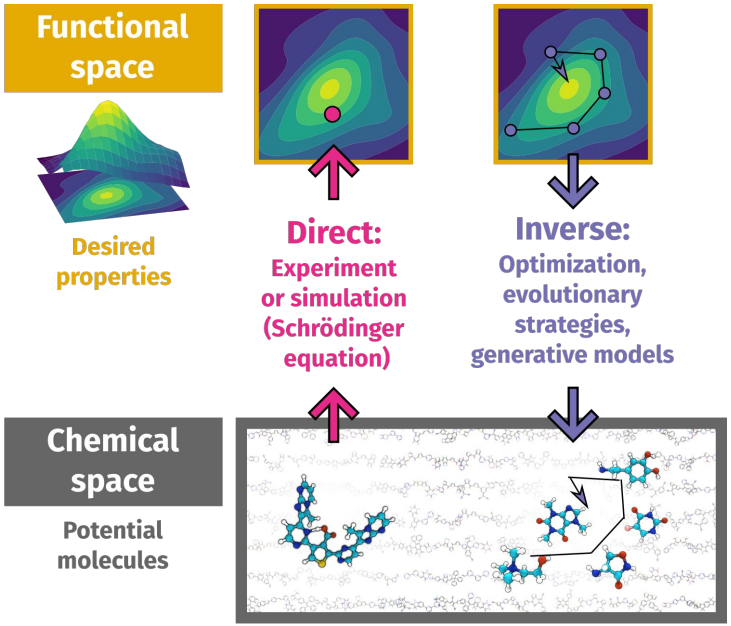


Figure modified from "Inverse molecular design using machine learning: Generative models for matter engineering" Science 2018, 10.1126/science.aat2663, **Benjamín Sanchez-Lengeling** and Alan Aspuru-Guzik

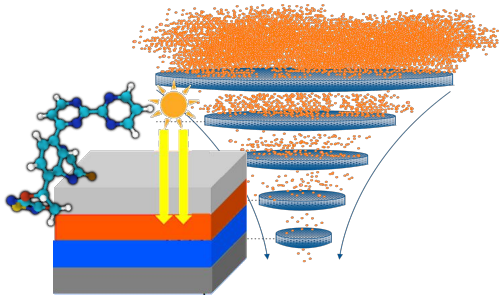
Context:

Using machine learning to build computational models for prediction and generation of organic molecules.



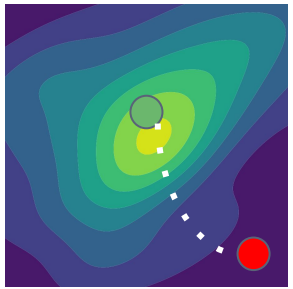
How can we find molecules according to functionality?

High throughput virtual screening (HTVS)



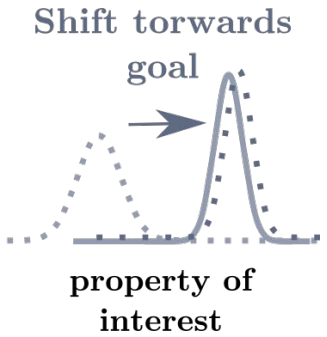
Library generation, Quantum chemistry, Gaussian Process prediction.

Explicit optimization



Variational Autoencoders, exploring and optimizing in latent space

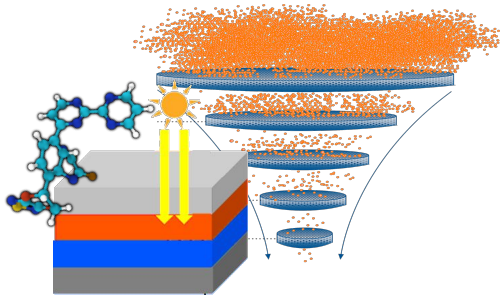
Implicit optimization



Reinforcement learning and generative adversarial networks

How can we find molecules according to functionality?

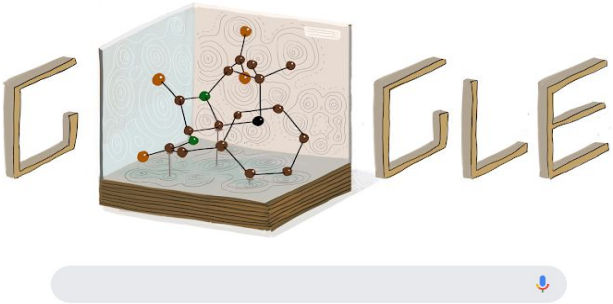
High throughput virtual screening (HTVS)



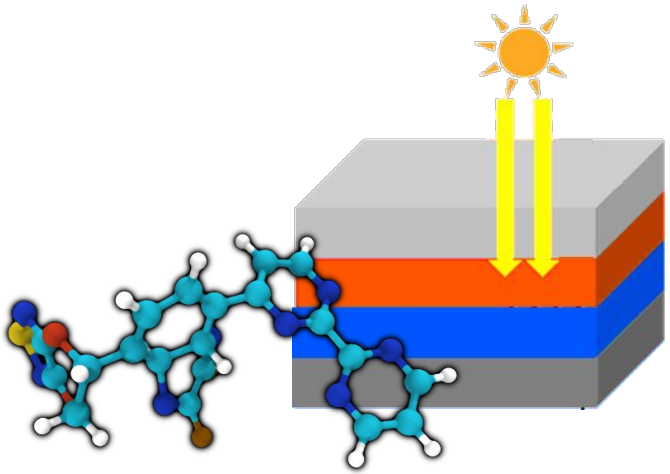
Quantum chemistry, Gaussian Process prediction and molecular structure interpretation.

Searching for molecules with a specific criteria

Inverse design for organic photovoltaics

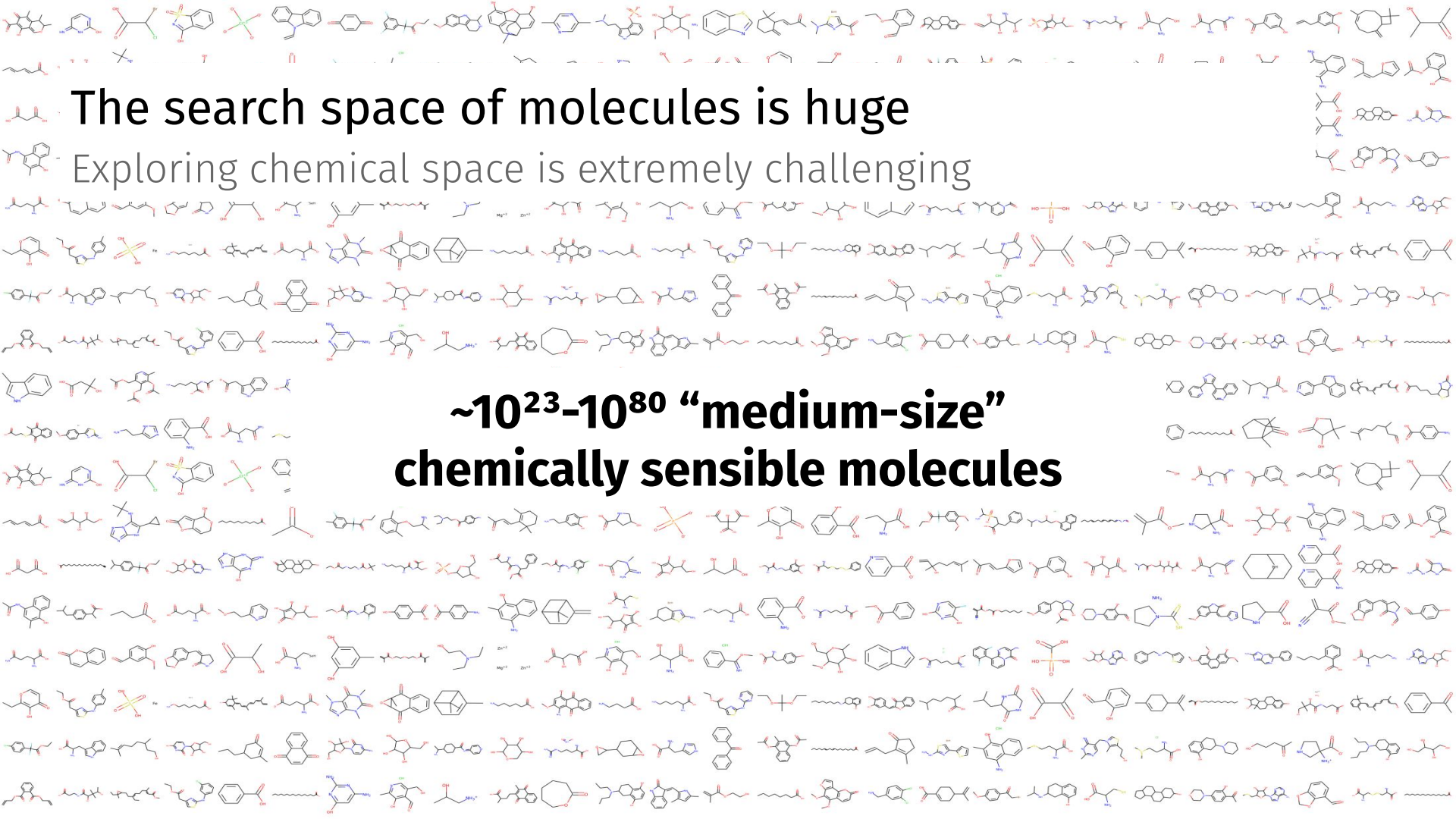


“I want organic, non-toxic and cheap molecules that will have a high efficiency in a solar cell”



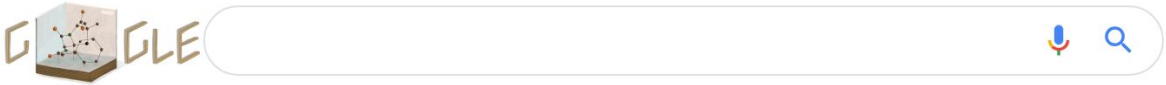
Design Principles and Top Non-Fullerene Acceptor Candidates for Organic Photovoltaics, Joule, Cell Press, 2017

Steven A. Lopez*, **Benjamin Sanchez-Lengeling***, Julio de Goes Soares, Alán Aspuru-Guzik


The search space of molecules is huge
Exploring chemical space is extremely challenging
~10^23-10^80 “medium-size” chemically sensible molecules

Building a molecular search engine

Going from query to top candidates



High throughput virtual screening (HTVS):

1) Build a search space
2) Predict accurately
3) Ranking results

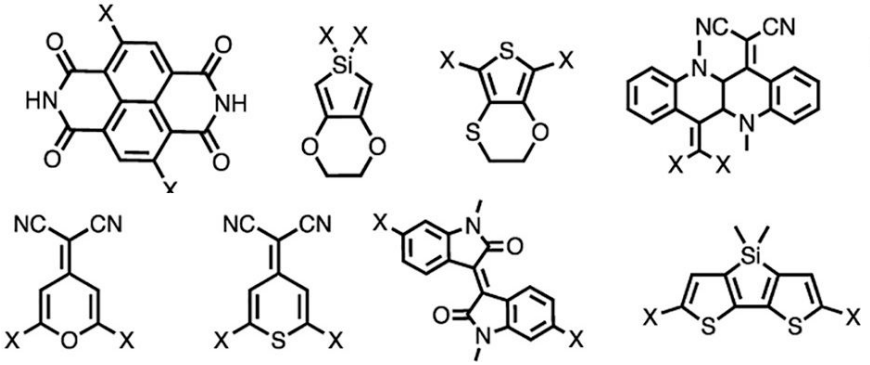
Extra goal:

Make results intuitive

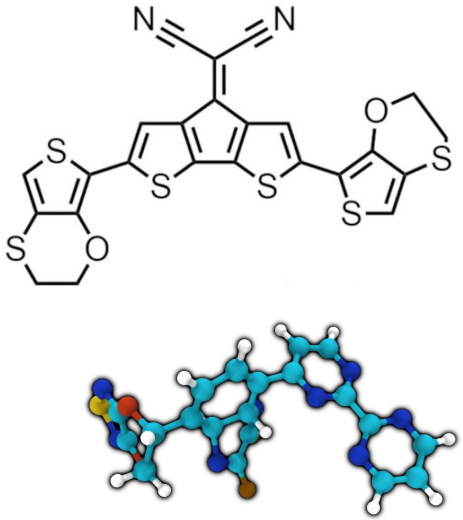
Step 1: Constraining our search space

We consider only combinations of select fragments

Library of fragments, known to be in good molecules



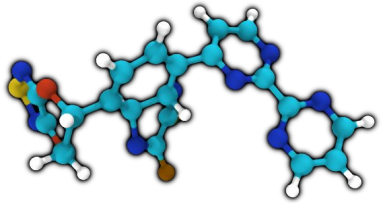
~51k molecules



Step 2: Predict efficiency

We can relate molecular properties to solar cell efficiency

Simulation via
Quantum Chemistry
to calculate
HOMO–LUMO gap



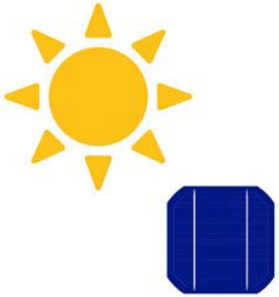
Expensive,
takes hours to days

Uses bandgap
and solar
information

Scharber model

Approximate,
takes seconds

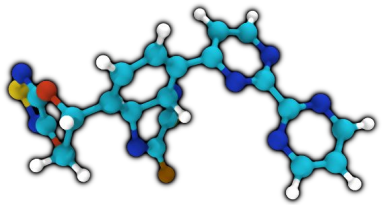
Solar cell
efficiency (%)

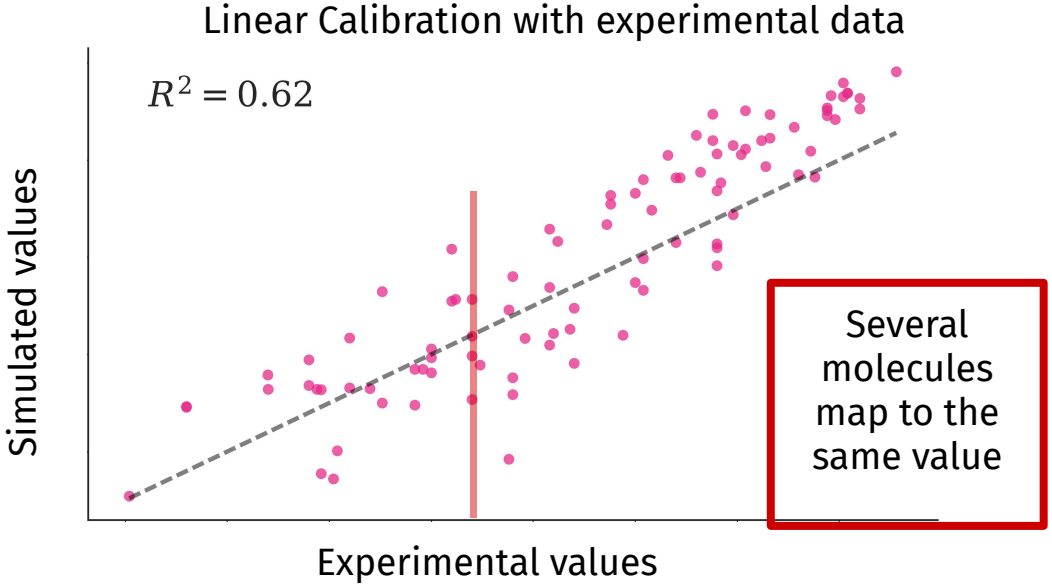


Step 2: Improving predictions

We can improve simulation by leveraging experiments

Simulation via
Quantum Chemistry
to calculate
HOMO–LUMO gap



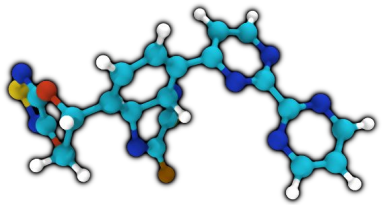


$$\begin{aligned} \text{bandgap} &= f\,(HOMO - LUMO\ gap) \\ &= aHOMO - LUMO\ gap + b \end{aligned}$$

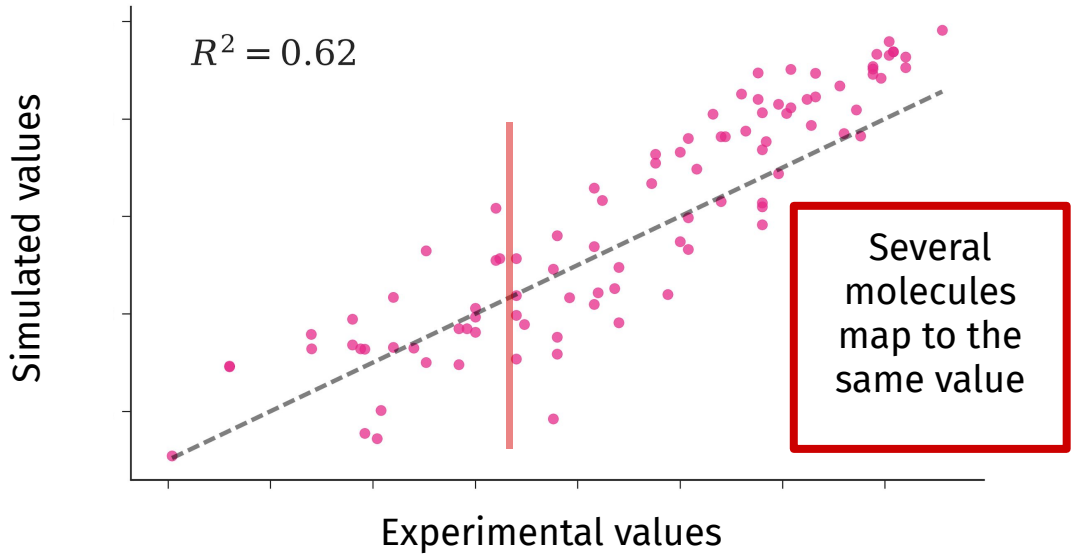
Step 2: Improving predictions

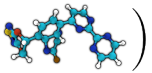
We can improve simulation by leveraging experiments

Simulation via
Quantum Chemistry
to calculate
HOMO–LUMO gap



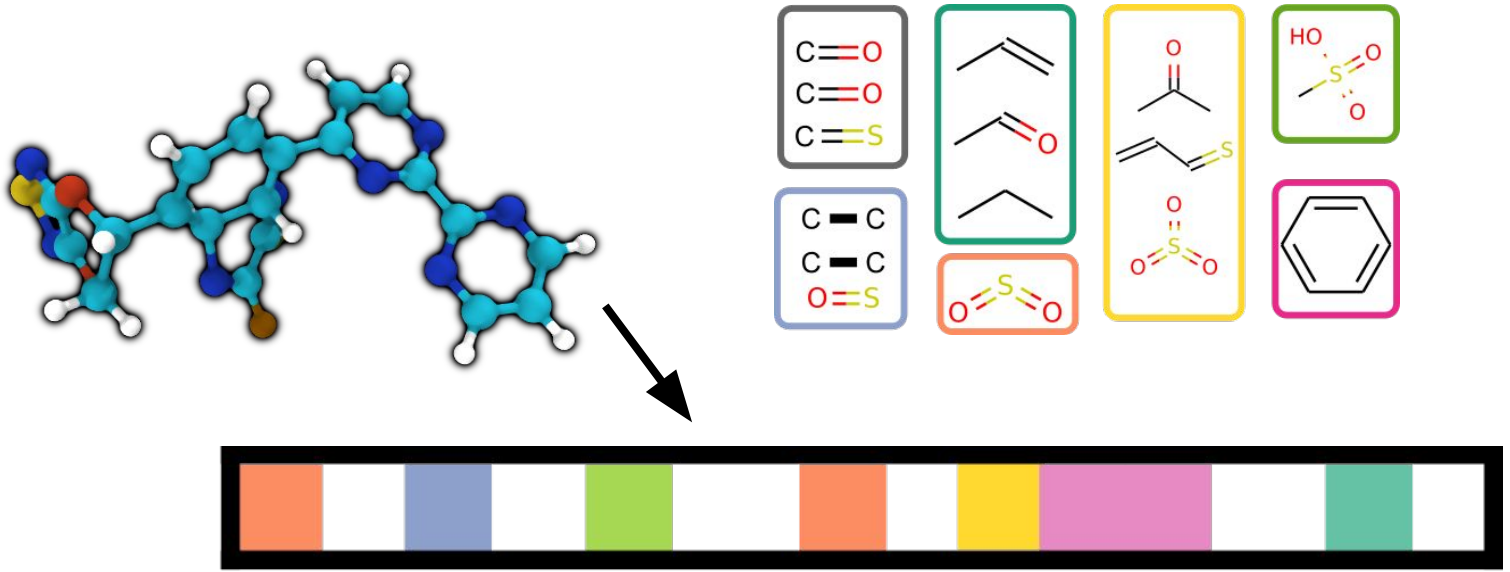
Linear Calibration with experimental data



$$\text{bandgap} = f(HOMO - LUMO\ gap,$$  $$)$$

How can we represent a molecular structure?

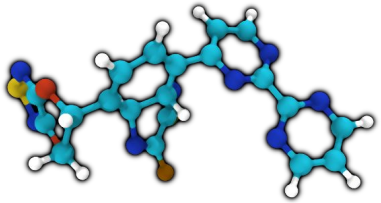
Fingerprints (FP): An effective bag-of-fragments representation

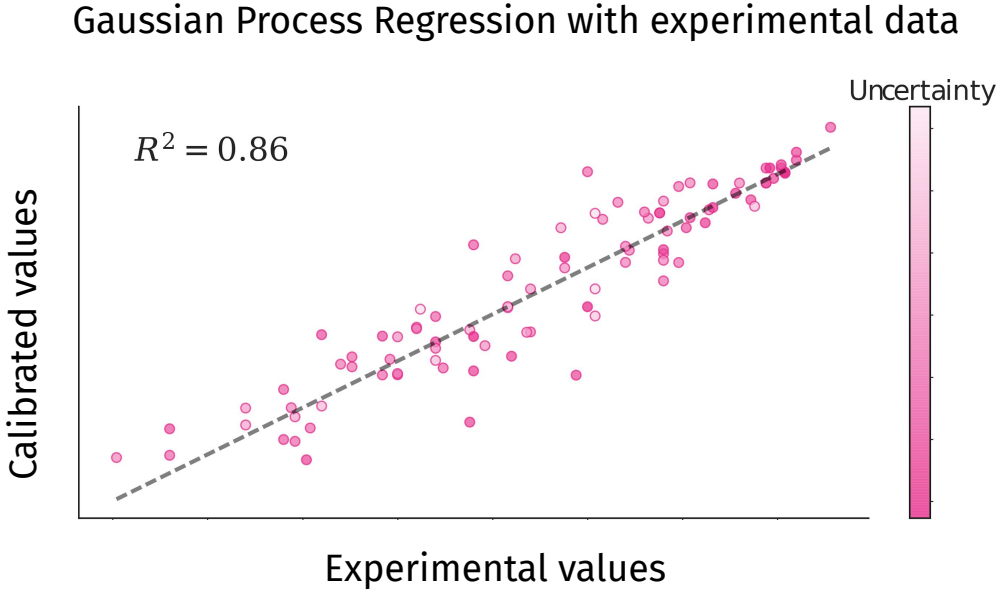


Step 2: Improving predictions

FP + GP improve prediction result on simulation baselines

Simulation via
Quantum Chemistry
to calculate
HOMO–LUMO gap





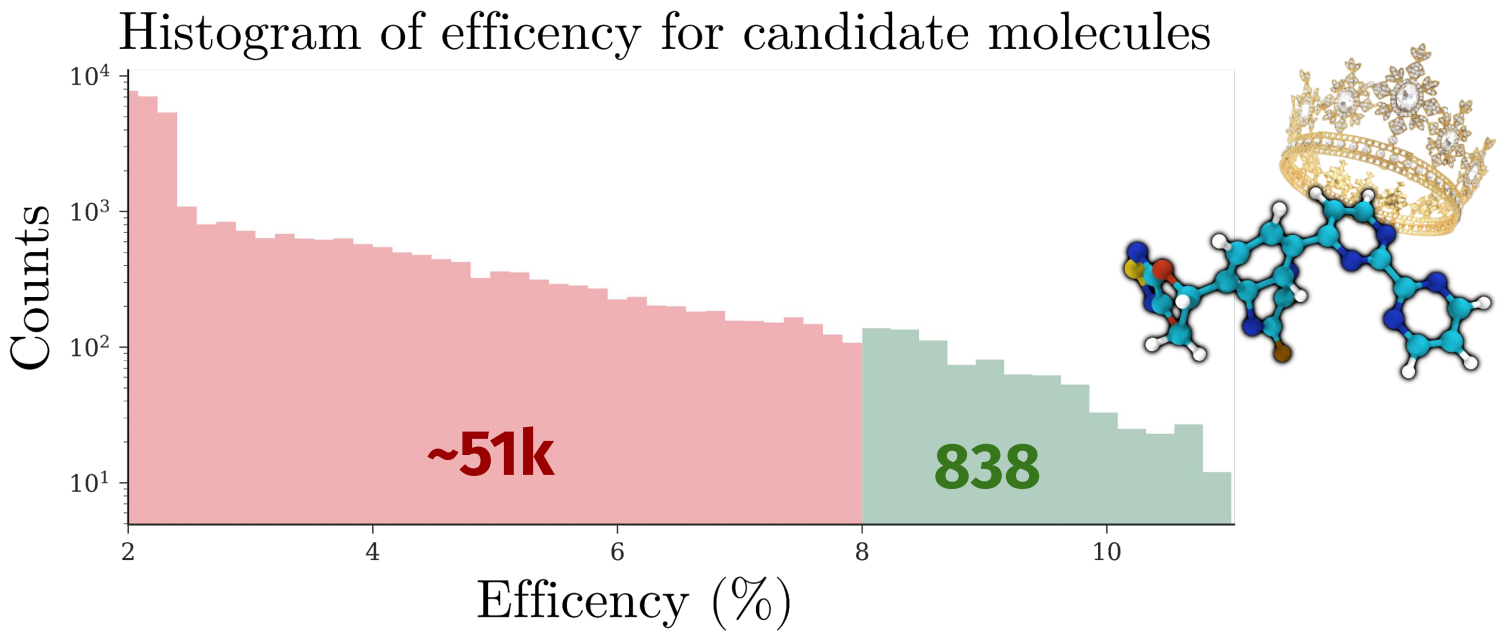
$$\text{bandgap} = GP(HOMO - LUMO\ gap + b, fp)$$

gpmol

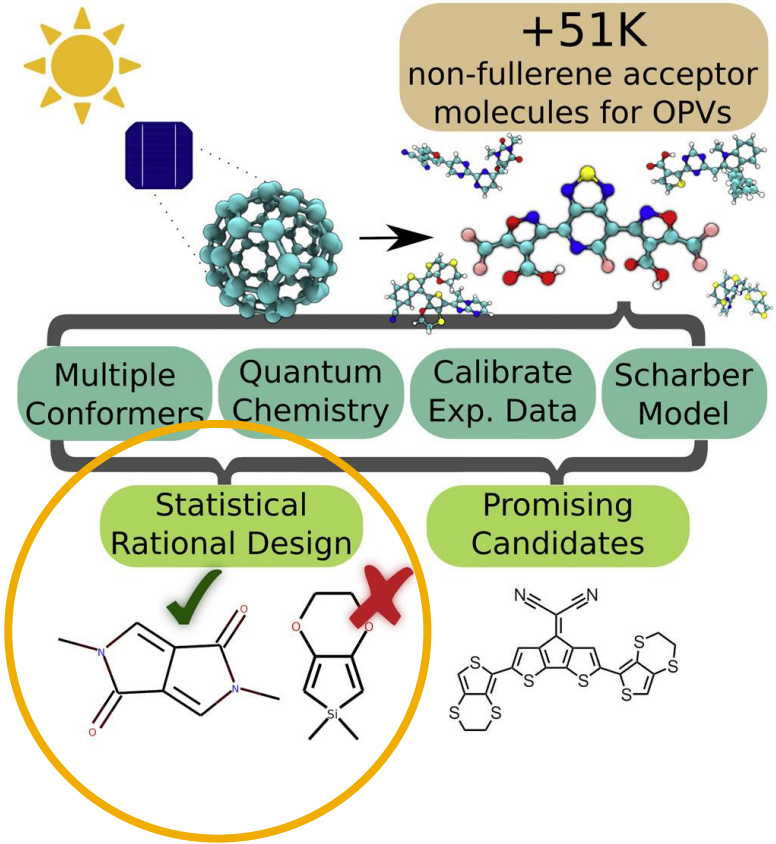
(coming in Nov'19 to a
github repo near you)

Step 3: Rank molecules

We can can pick our top candidates based on performance



At the end of the screening procedure

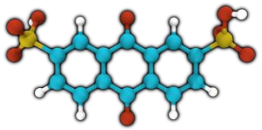


Collaboration with Christoph Brabec group, has led to two other works

Machine learning (ML) and deep learning (DL)

Learning representations of our data, optimized to a task

Input:
Molecular
representation

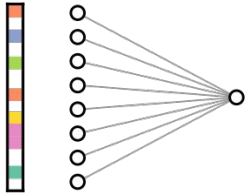


$x \longmapsto$

Traditional
machine learning
approach

Hand-crafted
features

x'



Predictive model:
e.g. Generalized
linear model

$$y = f(w \cdot x' + b)$$

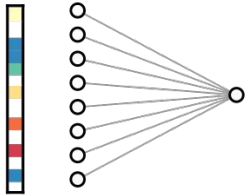
Requires
more data
to generalize

$x \longmapsto$

Parametrized,
optimizable
transformations

Deep learning
approach

z

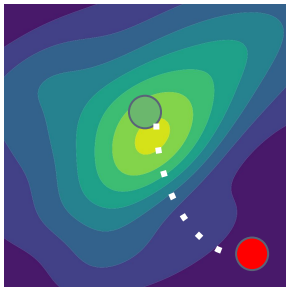


Learned
representation

$$y = f(w \cdot z + b)$$

How can we find molecules according to functionality?

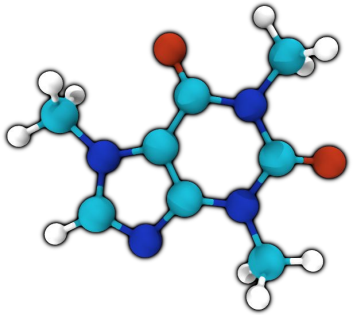
Explicit optimization



Variational Autoencoders, exploring and optimizing in latent space

Molecular data

Most data is unlabeled, and few public experimental sets exists



Caffeine

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

SMILES is a discrete sequence encoding the molecular graph

- Unlabeled datasets: upto 166B (GBD-17)
  - **ZINC ~ 980M molecules**
- Experimental datasets: upto 437k (PBCBA)
- Simulated data sets:
  - Semi-empirical upto 91M, most are less than 1M
  - High quality quantum chemistry: upto 134k (qm9)

Deep generative models

Neural networks optimized for a data generation task. The field of natural language processing has introduced many of these methods for discrete sequences.

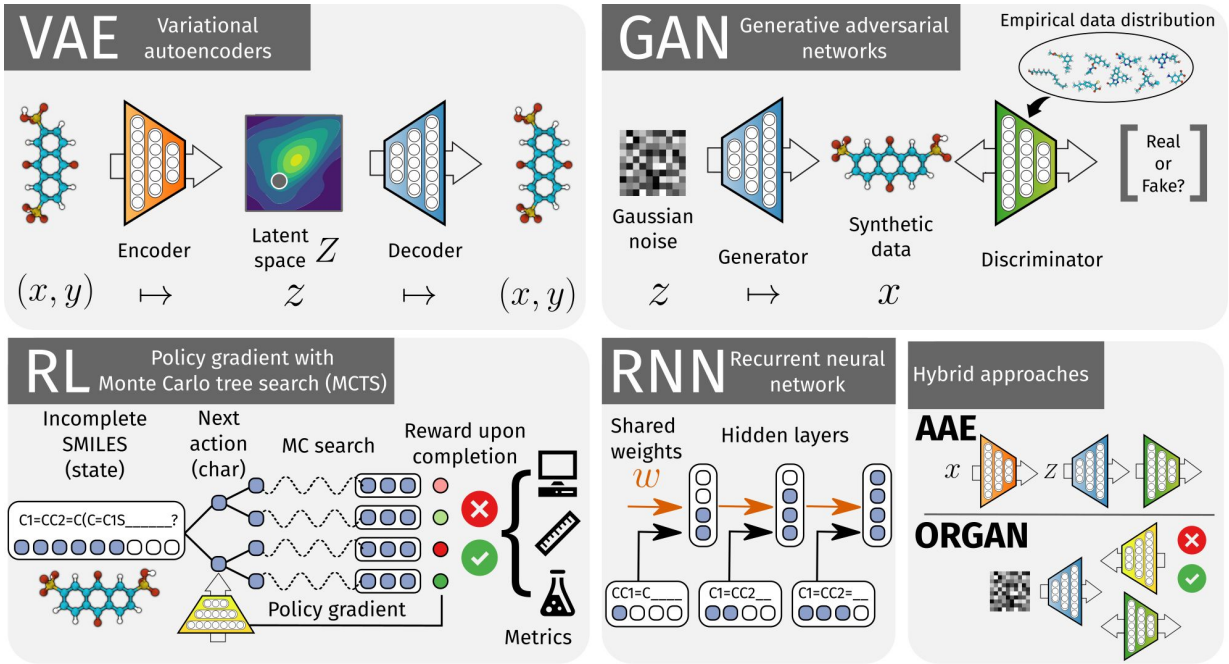
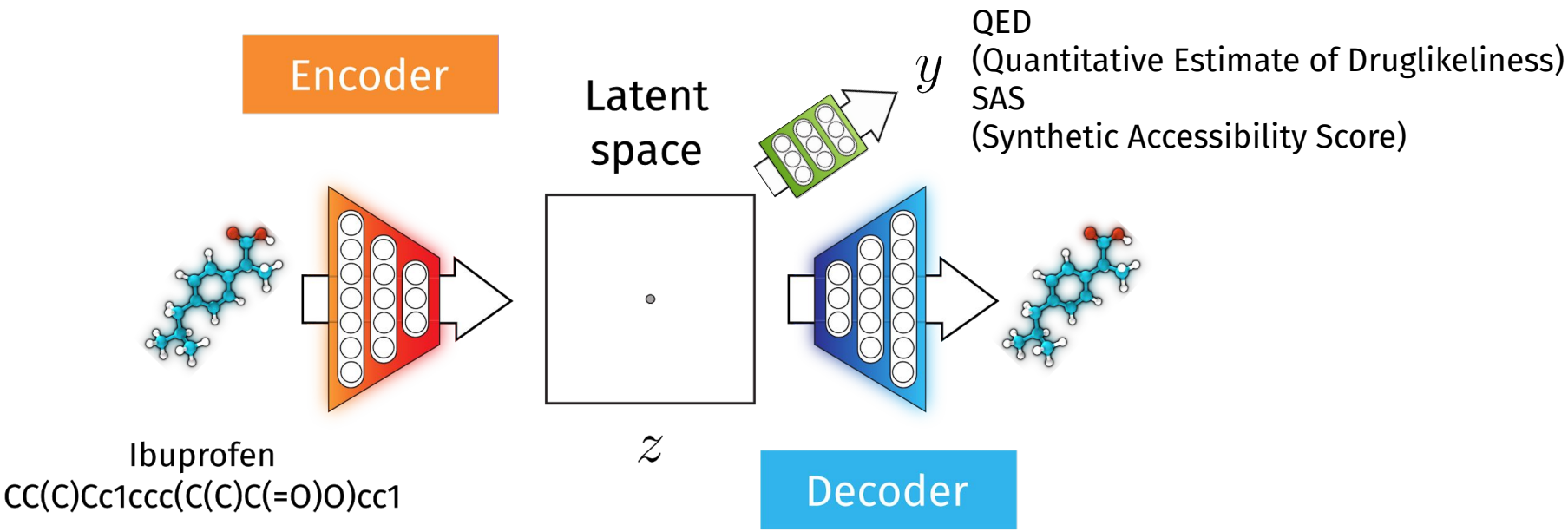


Figure modified from "Inverse molecular design using machine learning: Generative models for matter engineering" Science 2018, 10.1126/science.aat2663, **Benjamín Sanchez-Lengeling** and Alan Aspuru-Guzik

Variational Autoencoders (VAE)

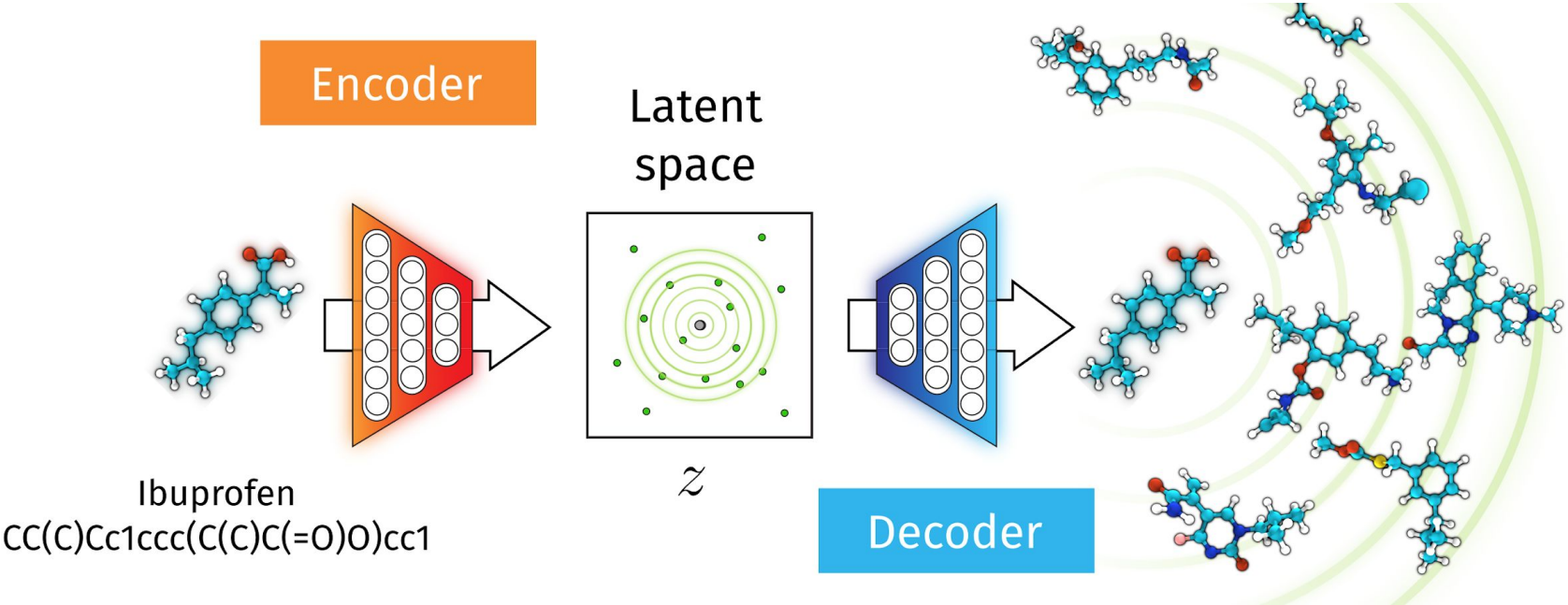
Learning a continuous and reversible representation for molecules



"Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules" ACS Cent. Sci., 2018
Rafa Bombarelli*, Jennifer N. Wei*, David Duvenaud*, José Miguel Hernandez-Lobato*, **Benjamín Sanchez-Lengeling**, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alan Aspuru-Guzik

Variational Autoencoders (VAE): Sampling

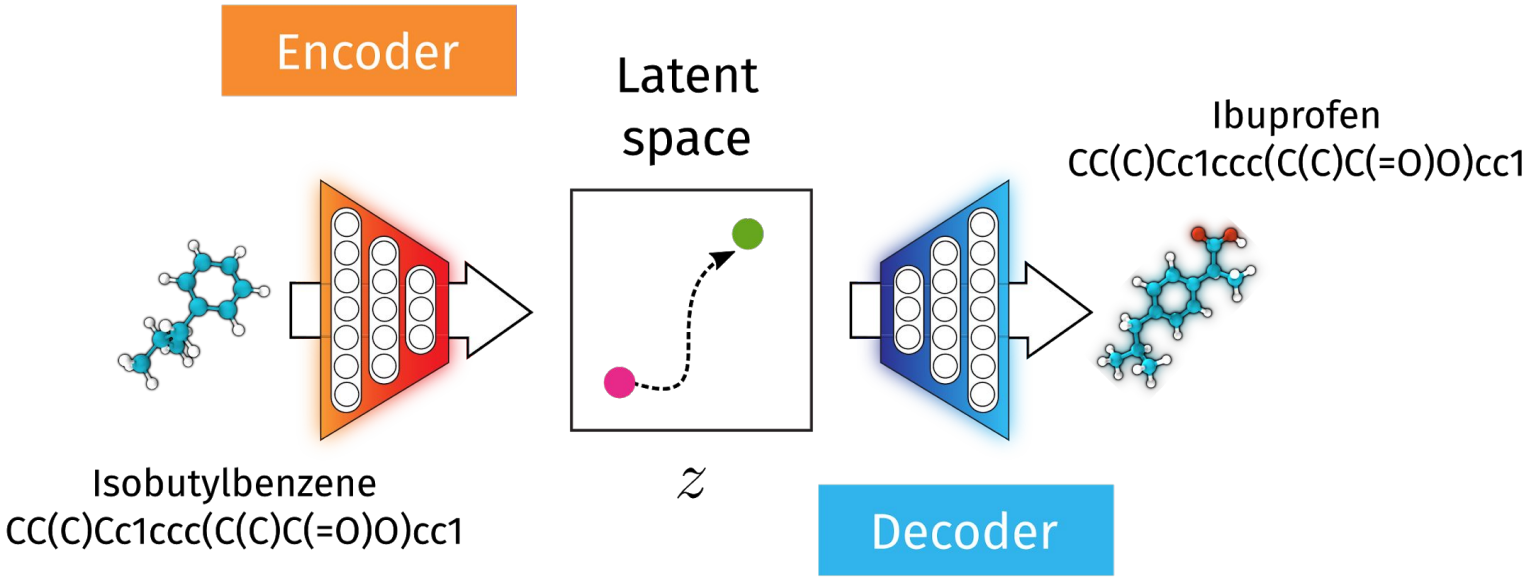
Decoded latent vectors become molecules



"Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules" ACS Cent. Sci., 2018
Rafa Bombarelli*, Jennifer N. Wei*, David Duvenaud*, José Miguel Hernandez-Lobato*, **Benjamín Sanchez-Lengeling**, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alan Aspuru-Guzik

Variational Autoencoders (VAE): Optimizing

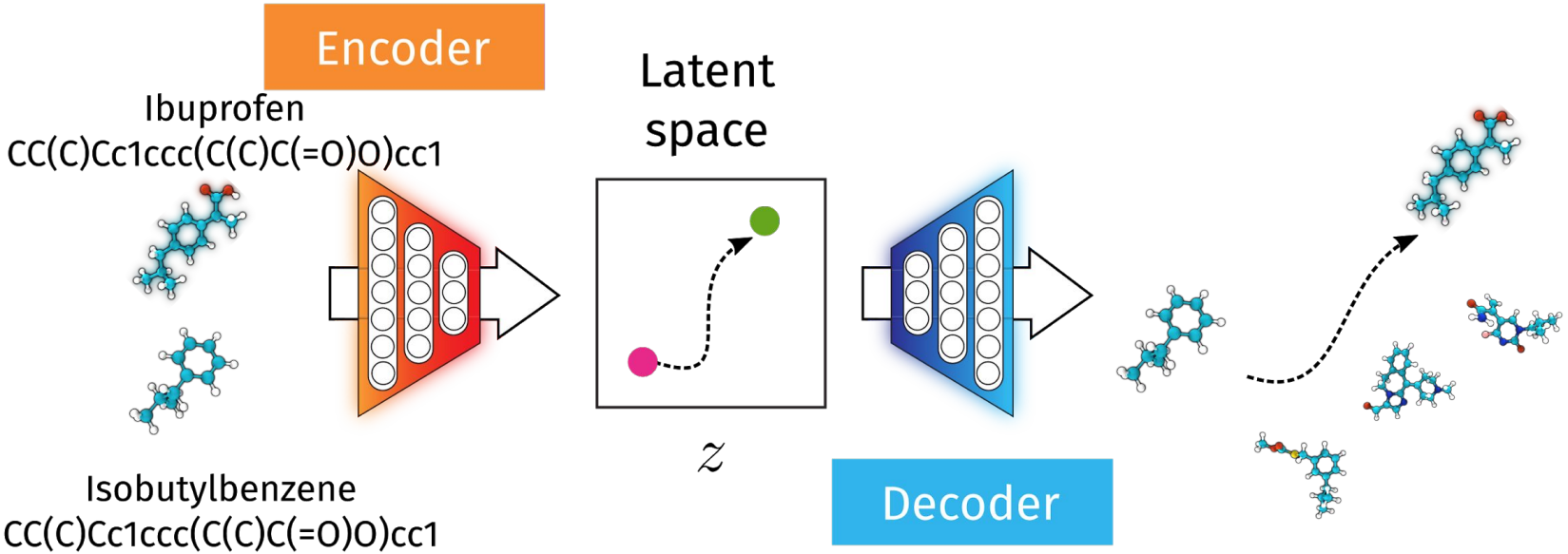
Optimizing in the latent space



"Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules" ACS Cent. Sci., 2018
Rafa Bombarelli*, Jennifer N. Wei*, David Duvenaud*, José Miguel Hernandez-Lobato*, **Benjamín Sanchez-Lengeling**, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alan Aspuru-Guzik

Variational Autoencoders (VAE): Interpolating

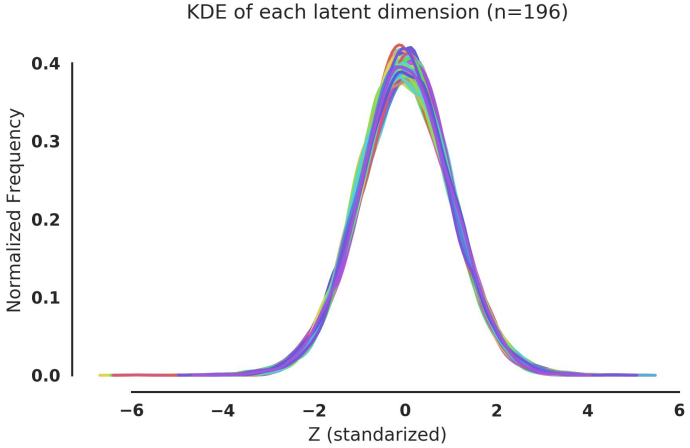
Connecting two latent vectors by smooth paths

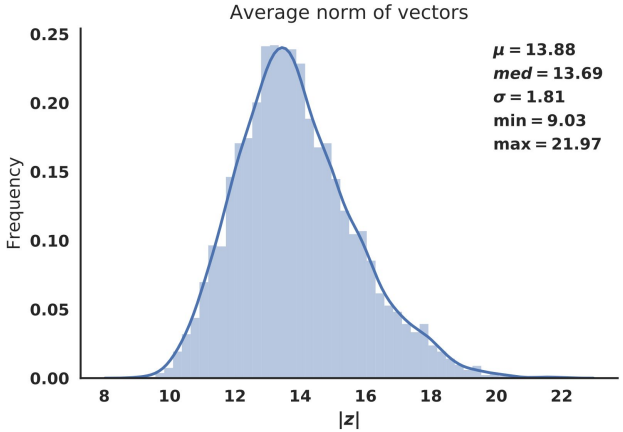


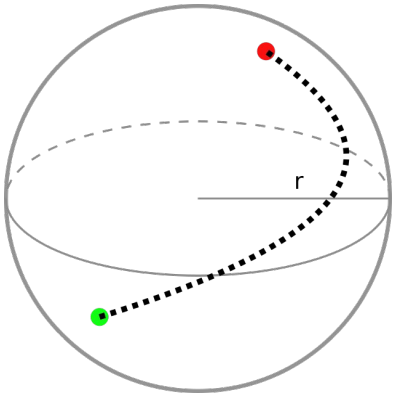
"Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules" ACS Cent. Sci., 2018
Rafa Bombarelli*, Jennifer N. Wei*, David Duvenaud*, José Miguel Hernandez-Lobato*, **Benjamín Sanchez-Lengeling**, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alan Aspuru-Guzik

Geometry of latent space

High dimensional spaces are not intuitive, our latent space is like a hyper annulus

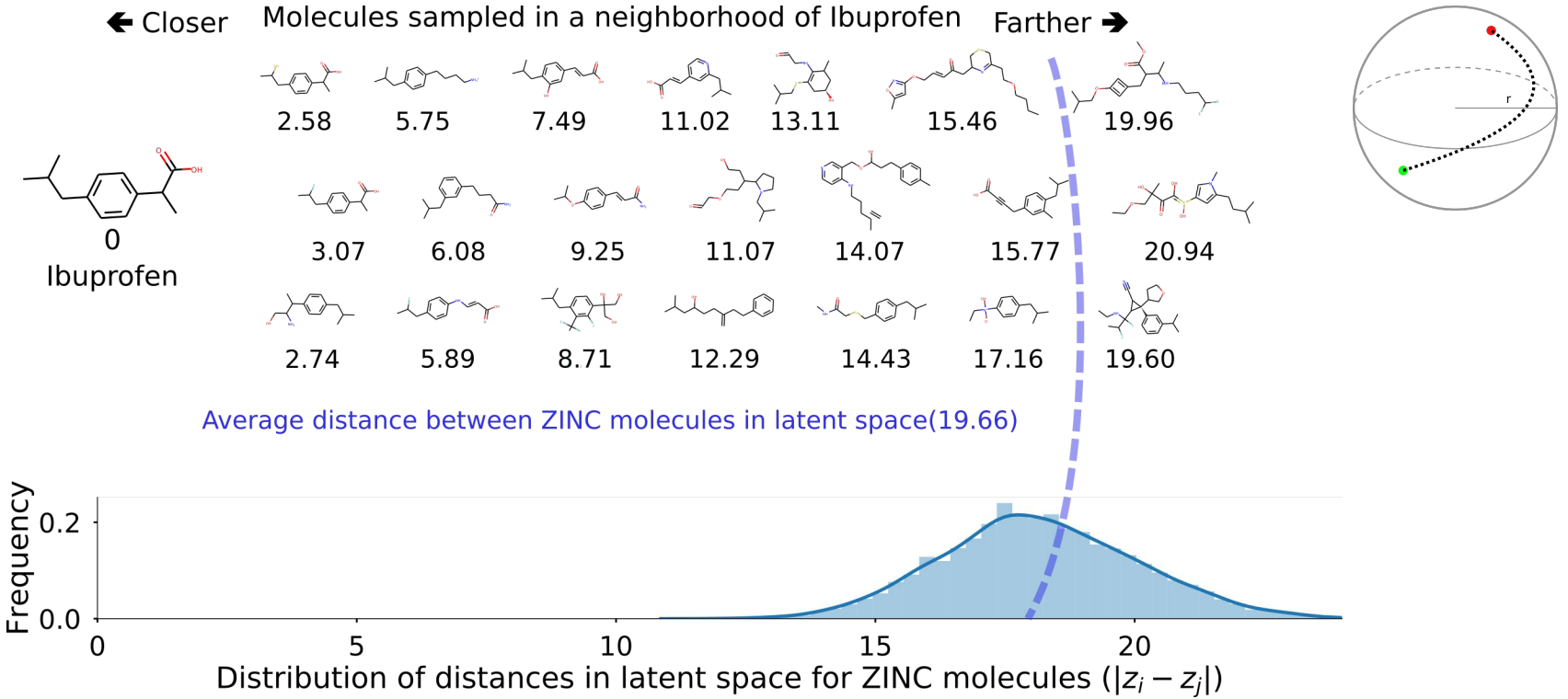






Local structure in the manifold of latent space

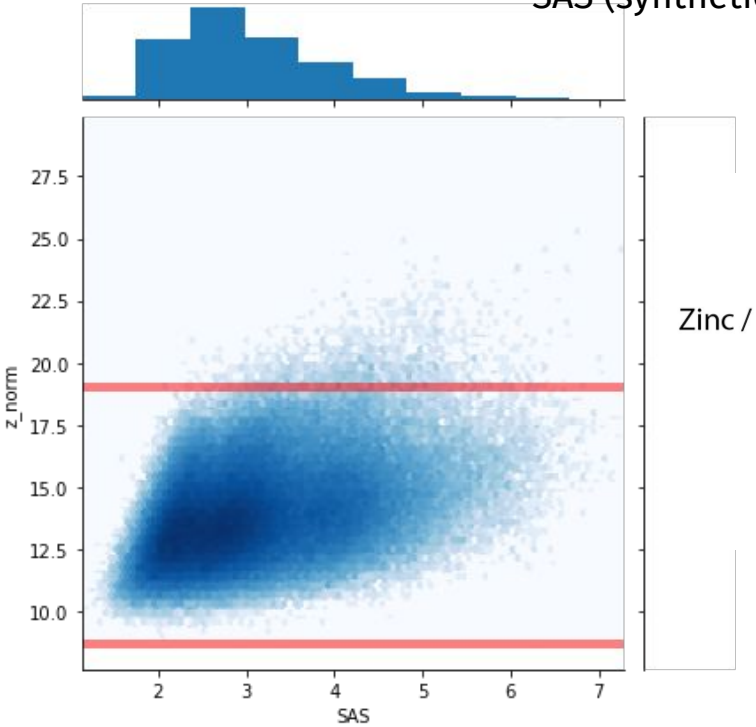
In a neighborhood of a molecule we find small local changes

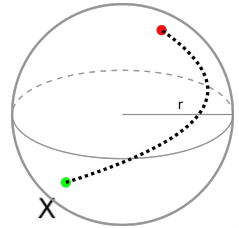


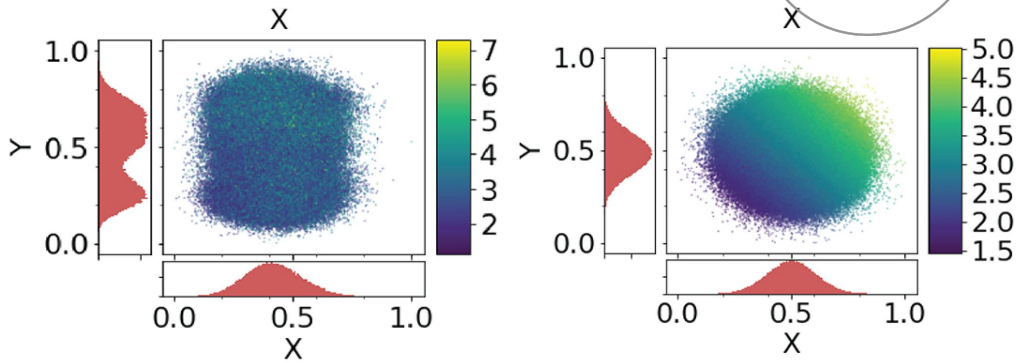
Global structure in the manifold of latent space

If we wish to optimize, having organized structure is advantageous

SAS (synthetic accessibility score)







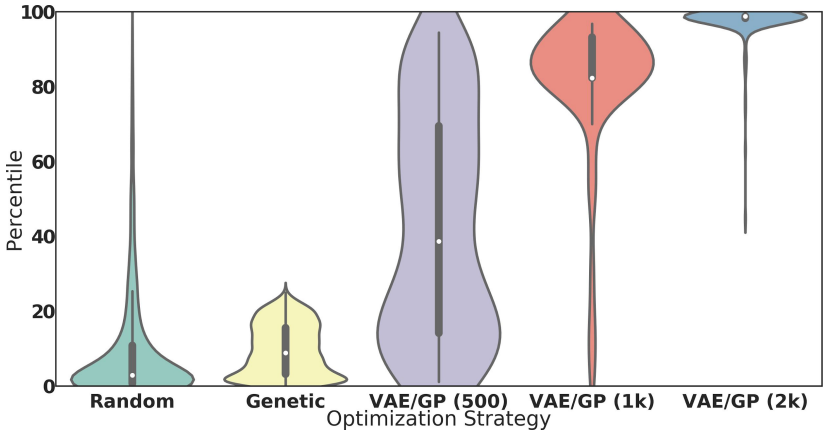
Latent space visualized via 2D PCA (X,Y)

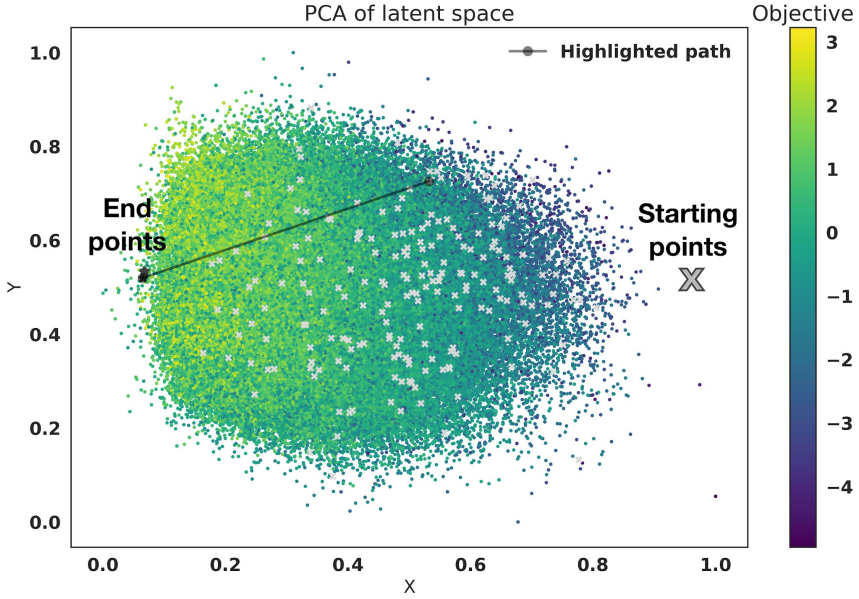
Optimization in latent space

Constrained bayesian optimization and local search at the end

Objective:
5 QED - SAS
(druglike and easy to synthesize)

QED in [0,1]
SAS in [1,6]



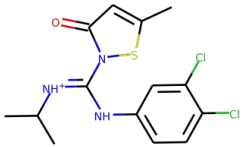


Optimization in latent space

Tracing the path between start and end points

Objective:
5 QED - SAS
(druglike and easy to synthesize)

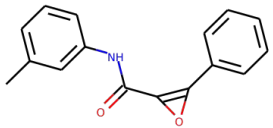
Start



(0.65,3.56,8.06%)

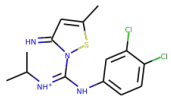
(QED, SAS, Percentile)

Finish

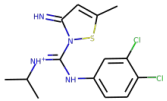


(0.89,2.09,98.23%)

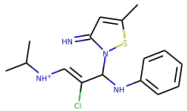
Intermediate path



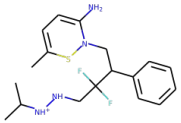
(0.57,3.67,9.21%)



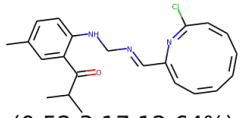
(0.57,3.67,9.21%)



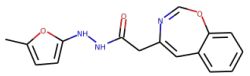
(0.74,4.46,10.16%)



(0.49,4.57,1.35%)



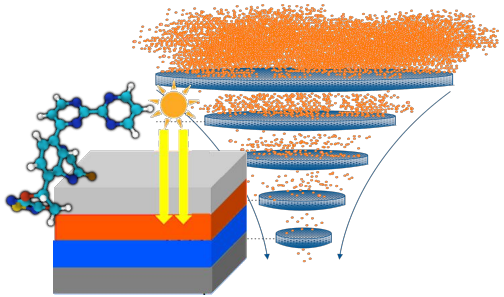
(0.52,3.17,12.64%)



(0.84,3.42,54.03%)

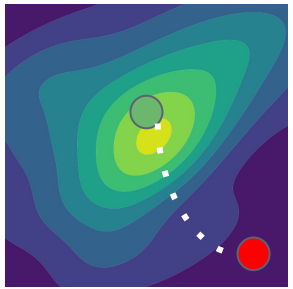
How can we find molecules according to functionality?

High throughput virtual screening (HTVS)



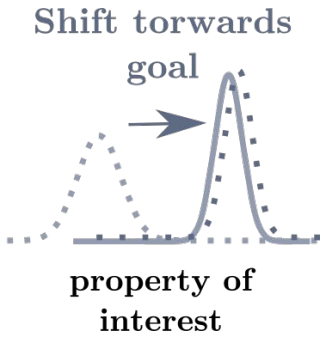
Quantum chemistry, Gaussian Process prediction and molecular structure interpretation.

Explicit optimization



Variational Autoencoders, exploring and optimizing in latent space

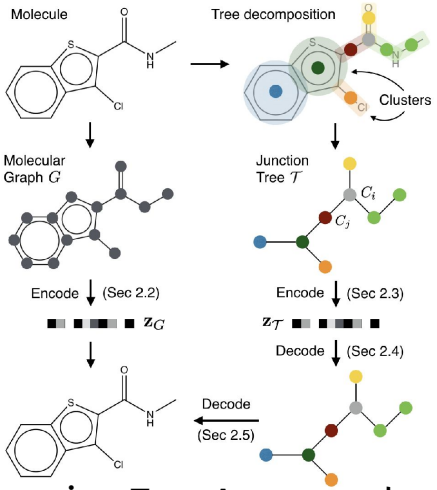
Implicit optimization



Reinforcement learning and generative adversarial networks

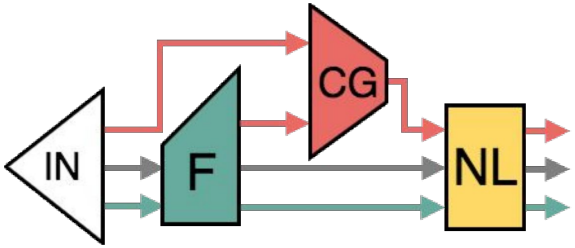
Looking to the future: representations and algorithms

Representations that capture hierarchical structure and symmetries of molecules



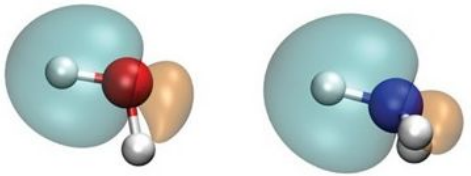
Junction Tree Autoencoder

Jin, W., Barzilay, R. & Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. arXiv: 1802.04364. (2018).



Tensor field networks

Thomas, N. *et al.* Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv: 1802.08219* (2018).

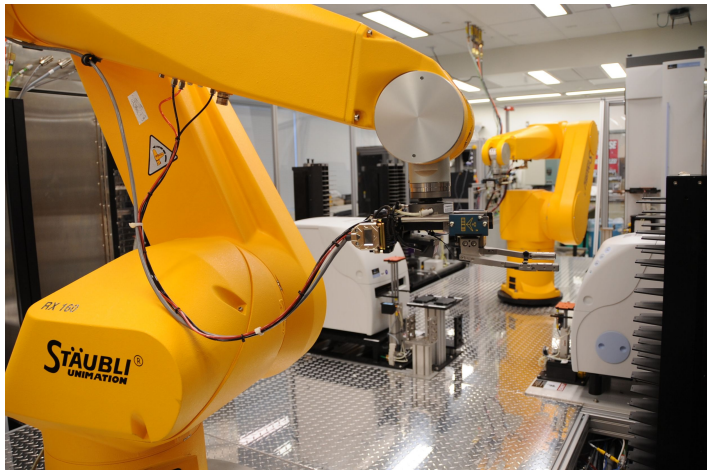


Incorporating more electronic structure (orbitals, wavefunctions)

Looking to the future: Larger, high quality datasets

Automation might pave the way for high quality datasets

- Unlabeled datasets: upto 166B (GBD-17)
  - **ZINC ~ 980M molecules**
- Experimental datasets: upto 437k (PBCBA)
- Simulated data sets:
  - Semi-empirical upto 91M, most are less than 1M
  - High quality quantum chemistry: upto 134k (qm9)



One challenge is on generalizability:

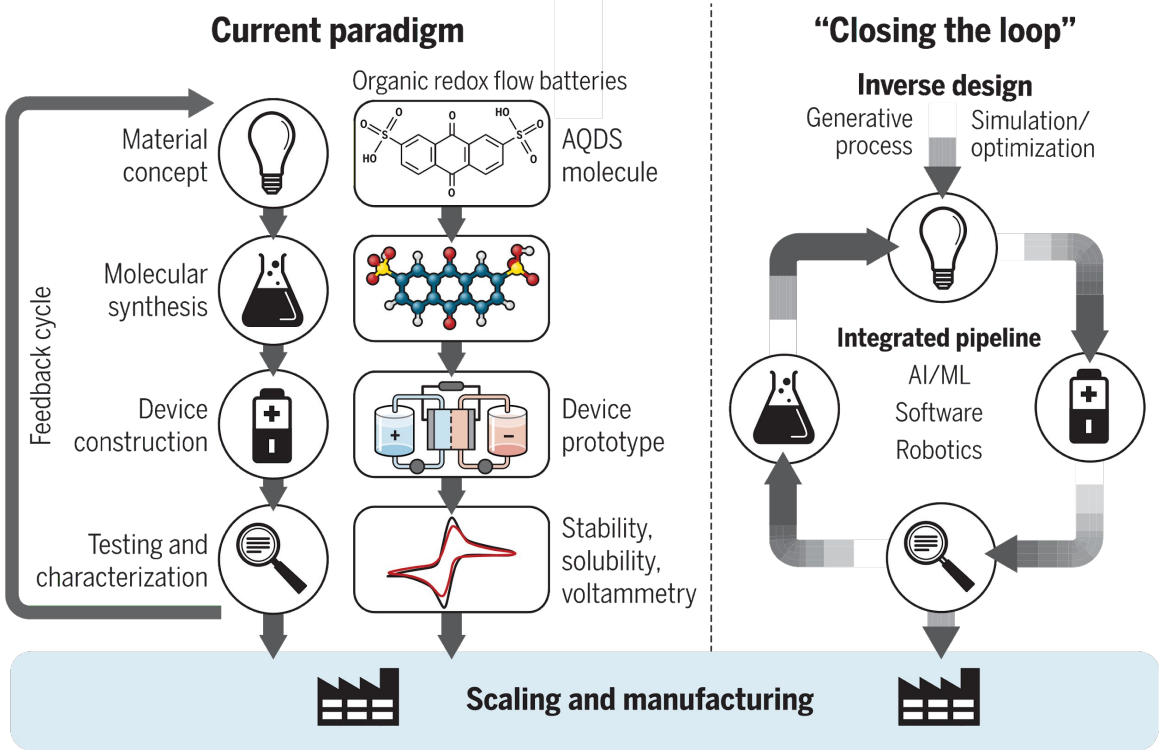
We can accurately predict electronic properties for small molecules (qm9). *

Remains to be seen for the rest of molecular space.

* Neural Message Passing for Quantum Chemistry, 2017
Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, George E. Dahl.

Looking to the future: closing the loop

Inverse design is one component within a broader material challenge.



Machine Learning for chemistry of smell

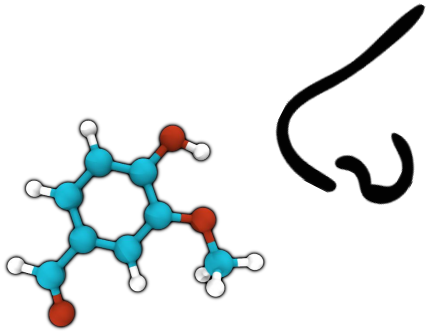
Nosebrain @ Google Brain

Alex Wiltschko

Jennifer Wei

Brian Lee

Carey



More new this thursday at
https://ai.googleblog.com/

- Large high quality dataset
- Fast and accurate prediction (GNN)
- Interpretability tools (graph attribution and embeddings)
- Easy to test (just take a whiff)
- Generation seems plausible (small organic molecules)

Many thanks!

Alán Aspuru-Guzik

A2G2

Adrian Jinich
Jennifer Wei
Daniel Tabor
Dennis Sheberla
Aniket Zinzuwadia
Loic Roch
Florian Hase
Luis Martin
Jhonathan Romero

Aniket ZInzuwadia
Rafa Bombarelli
Dmitri Rappaport
Tim Hirzel
Steven Lopez
Semion Saikin
Teresa Tamayo
Gabriel Guimaraes
Carlos Ouiteral

Outside A2G2:
Alex Wiltschko
Dario Perea
Christoph Brabec group
Insilico Medicine
Afshan Mohajeri

And many others!

Thanks to AISIS ‘19

Also check out RIIAA v3 in 2020!

RIIAA MEX

International Meeting on Artificial Intelligence and its Applications

riiaa.org

Any questions?

(and thanks for hearing me !)

Also email:

beangoben@gmail.com

bmsanchez@google.com