Automatic chemical design using a continuous, data-driven representation of molecules

Rafa Gómez-Bombarelli, David Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy Hirzel, Ryan P. Adams, Alán Aspuru-Guzik
How to design molecules?

- Can predict properties of molecules (QSAR)
- Want a molecule with a certain property (inverse QSAR)
How to optimize molecules?

- Build a giant library of molecules and try them all
- Or: Local search based on small changes

Cornelius J. O’Connor, Henning S.G. Beckmanna and David R. Spring
Discrete optimization is basically impossible

- In more than 10 or 20 dimensions, search is too slow because no way to know which direction to go
- Gradient gives D hints for your D parameters
- need some version of \[ \frac{\partial \text{molecule}}{\partial \text{property}} \]
Discrete optimization is basically impossible

- In more than 10 or 20 dimensions, search is too slow because no way to know which direction to go.
- Gradient gives D hints for your D parameters.
- Need some version of \( \frac{\partial \text{molecule}}{\partial \text{property}} \)
Discrete optimization is basically impossible

• In more than 10 or 20 dimensions, search is too slow because no way to know which direction to go

• Gradient gives D hints for your D parameters

• need some version of $\frac{\partial \text{molecule}}{\partial \text{property}}$

Nguyen, Dosovitskiy, Yosinski, Brox, Clune
What is a molecule?

Graph | SMILES string

| ![Graph 1] | CCC[C@@H](O)CC\C=C\C\C=\C\C#CC#C\C=C\CO |
| ![Graph 2] | COC(=O)C(\C)=\C\C1\C(\C)(\C)[C@H]1C(=O)O[C@@H]2\C(C)=\C(=O)C2\CC=\CC=\C |
| ![Graph 3] | O1C=C[C@H]([C@H]1O2)c3c2cc(OC)c4c3OC(=O)C5=C4CCC(=O)5 |
| ![Graph 4] | OC[C@@H](O1)[C@@H](O)[C@H](O)[C@@H](O)[C@@H](O)[C@@H](O)1 |
• **Generating Sentences from a Continuous Space.**
“i want to talk to you.”
“i want to be with you.”
“i do n’t want to be with you.”
i do n’t want to be with you.
she did n’t want to be with him.

it made me want to cry.
no one had seen him since.
it made me feel uneasy.
no one had seen him.
the thought made me smile.
the pain was unbearable.
the crowd was silent.
the man called out.
the old man said.
the man asked.

he was silent for a long moment.
he was silent for a moment.
it was quiet for a moment.
it was dark and cold.
there was a pause.
it was my turn.
Repurposing text autoencoders

Can be trained on unlabeled data
Map of 220,000 Drugs
Map of 100,000 OLEDs
Random Organic LEDs

Variational autoencoder

Standard autoencoder
Molecules near aspirin
Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high. Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin.

No chemistry-specific design!
Gradient-based optimization

1. 
2. 
3. 
4. 

Property \( f(z) \)

Latent Space \( z \)

Most Probable Decoding \( \text{argmax } p(*)|z) \)
Proof of concept

But be careful what you wish for

Molecule 1

Molecule 2

Objective Values in Training Data

Molecule 1

Molecule 2

Objective Values

-15 -10 -5 0 5 10
Applications of small-molecule design

Organic LEDs, liquid crystals, organic solar cells, gas dielectrics, supercapacitors, batteries, electronic polymers, homogeneous catalysts, plastic additives, adhesives, sealants, 3D printing, paints and coatings, specialty fibers, biodegradable polymers, medical plastics, pesticides, small molecule drugs
1. **Twitter Plans Hundreds More Job Cuts as Soon as This Week** *(bloomberg.com)*
   198 points by bentlegen 4 hours ago | hide | 147 comments

2. **Microsoft Cognitive Services** *(projectoxford.ai)*
   61 points by igraviou 3 hours ago | hide | 13 comments

3. **Apple Introduces What It Calls an Easier to Use Portable Music Player** *(2001)*
   208 points by daschaefer 7 hours ago | hide | 153 comments

4. **Chrome Requiring Certificate Transparency in 2017** *(groups.google.com)*
   79 points by edmorley 4 hours ago | hide | 11 comments

5. **Delta functions [pdf]** *(berkeley.edu)*
   7 points by lisper 1 hour ago | hide | 1 comment

6. **Keras-based molecular autoencoder** *(github.com)*
   67 points by frisco 5 hours ago | hide | 24 comments

7. **Ask HN: What is your favorite internet rabbit hole?**
   763 points by karim 15 hours ago | hide | 382 comments

8. **Web Bloat Score Calculator** *(webbloatscore.com)*
   24 points by zdw 2 hours ago | hide | 2 comments

9. **IHaskell: A Haskell kernel for IPython** *(github.com)*
   19 points by sndean 2 hours ago | hide | discuss

10. **Introducing Initialized Capital** *(initialized.com)*
    489 points by ernestipark 13 hours ago | hide | 130 comments

11. **Social Media’s Dial-Up Ancestor: The Bulletin Board System** *(ieee.org)*
Future work
Future work

• Jointly train autoencoder and prediction model
Future work

• Jointly train autoencoder and prediction model
• Decode directly to graphs
Future work

• Jointly train autoencoder and prediction model
• Decode directly to graphs
• Decode directly to recipes for synthesis
Thanks!

Rafa Gómez-Bombarelli, Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre

Timothy Hirzel, Ryan P. Adams, Alán Aspuru-Guzik