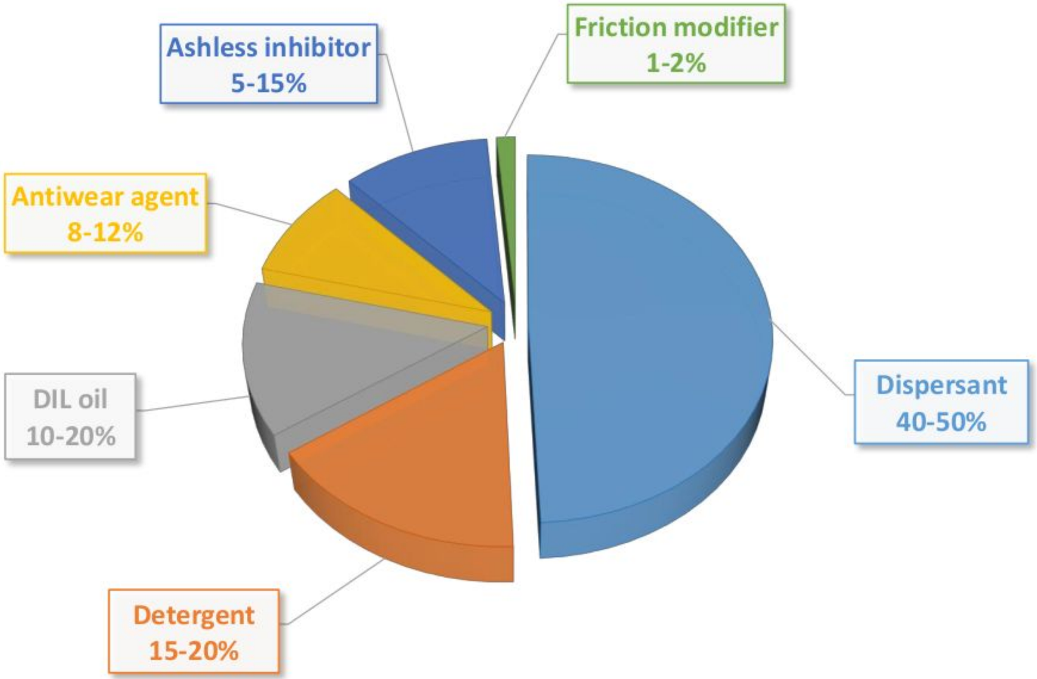
ICMAT
INSTITUTO DE CIENCIAS MATEMÁTICAS

Machine Learning for Molecular Design: a case study
Dispersant design

Roi Naveiro

SEIO - 2022

Dispersants in Lubricants



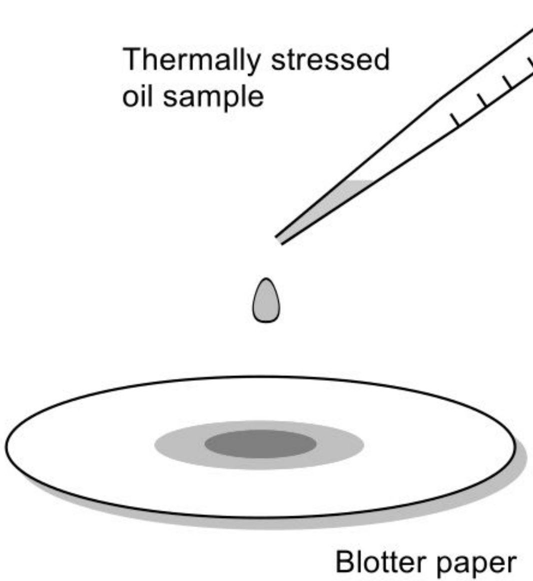
- Lubricants for combustion engines require formulated additive package (dispersants)

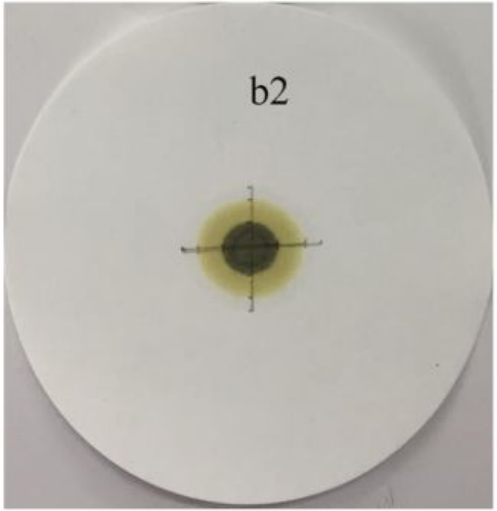
- Under harsh operating conditions of engines, soot is produced.

- Soot aggregation increases lubricant viscosity causing corrosion, deposit formation…

- Dispersants are molecules that adsorbs onto the surface of ultrafine carbon deposit precursors reducing their aggregation.

Measuring Dispersancy Efficacy - Blotter Spot

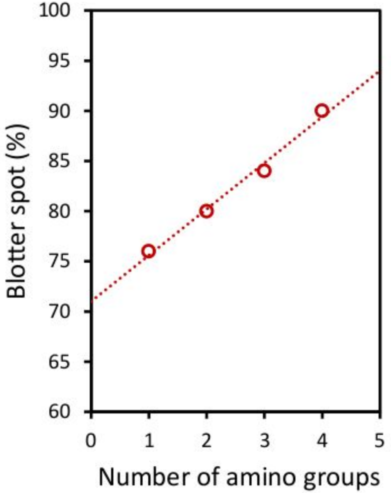


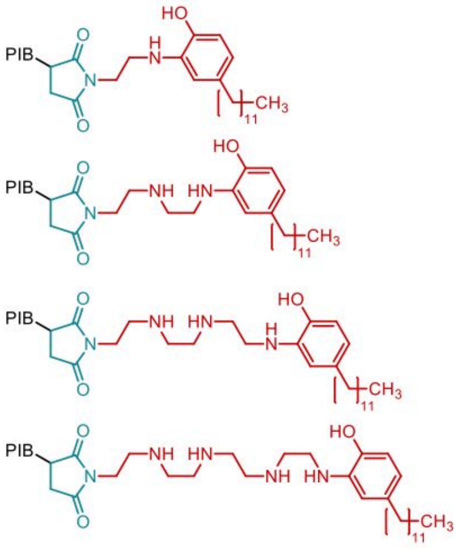


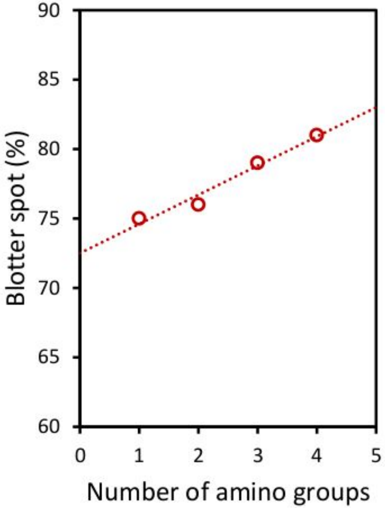
$$\text{Blotter Spot Dispersancy (\%)} = \frac{\text{diameter of black spot}}{\text{diameter of the total spot}} \text{ X } 100$$

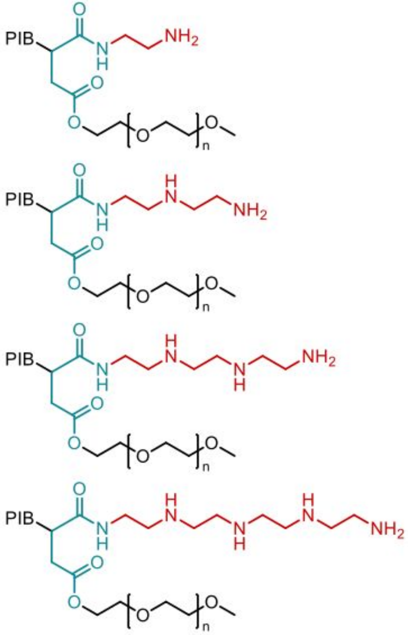
Dispersancy Estimation – Limits of Chemist Intuition

Within a family of substrate, predictable behaviors are appreciable.









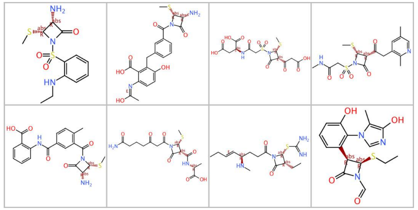
- However, the relationship between different families of substrates cannot be determined intuitively

Abdel Azim, A.-A. A. et al. *Int. J. Polym. Mater.* **2006**, *55*, 703
Abdel-Azim, A.-A. A. et al *Int. J. Polym. Mater.* **2007**, *57*, 114

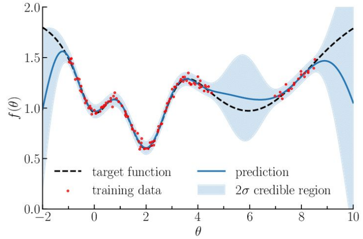
Goal - Find molecular structure with high blotter spot...

Solve black box optimization in chemical space (very limited number of evaluations!)

Data



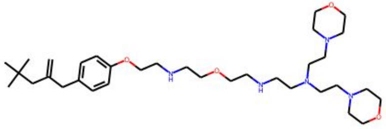
Probabilistic Model



Efficiently Explore Chemical Space

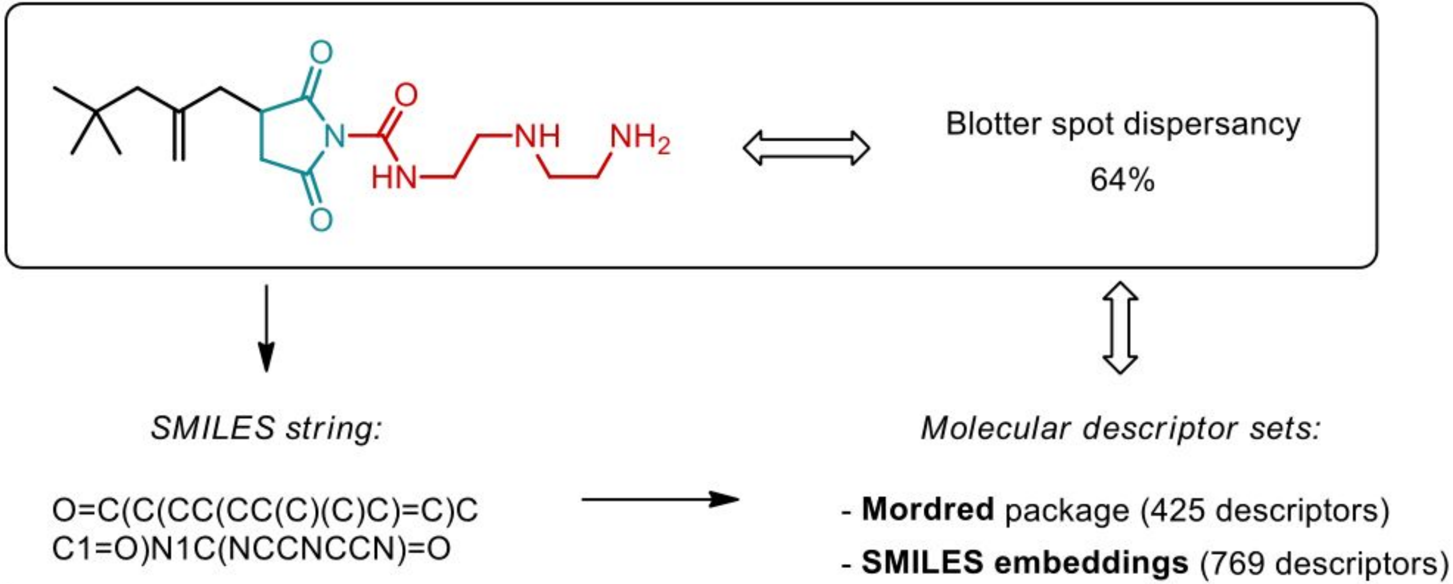
Síntesis & Evaluation

EU Maximization



Probabilistic Model for Dispersancy - Data and Molecular Representation

Dataset of 60 structures with associated Blotter Spot measure



Blotter spot dispersancy

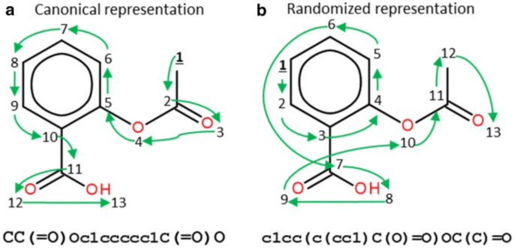
64%

SMILES string:

O=C(C(CC(CC(C)(C)C)=C)C
C1=O)N1C(NCCNCCN)=O

Molecular descriptor sets:

- **Mordred** package (425 descriptors)
- **SMILES embeddings** (769 descriptors)



CC(=O)Oc1ccccc1C(=O)O

c1cc(c(cc1)C(O)=O)OC(C)=O

Probabilistic Model for Dispersancy - The Model

- p >> N: sparsity inducing models
- Non linearity, interaction effects

- Bayesian Additive Regression Trees (BART) : sum-of-trees model + regularization prior

$$y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon; \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Posterior inference through MCMC

$$p((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma | \mathcal{D})$$

- Shallow trees capture varying (small) size interaction effects

- Natural way of performing variable selection (using variable importance measures)

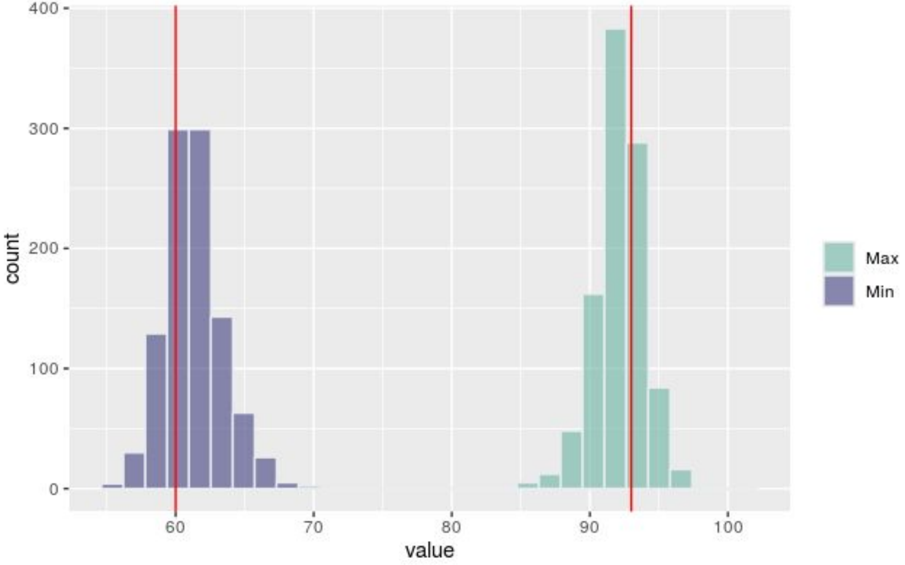
- Better predictive performance than: linear regression with horseshoe prior, GP.

Probabilistic Model for Dispersancy - Prediction

- Given new structure with descriptors x, we need to sample from the predictive distribution $p(y|x)$
- Sample

$$[T_j, M_j]_{j=1}^m, \sigma \sim p([T_j, M_j]_{j=1}^m, \sigma|\mathcal{D})$$

$$y \sim \mathcal{N}\left(\sum_{j=1}^m g(x; T_j, M_j), \sigma^2\right)$$



Posterior predictive samples from any function can be generated

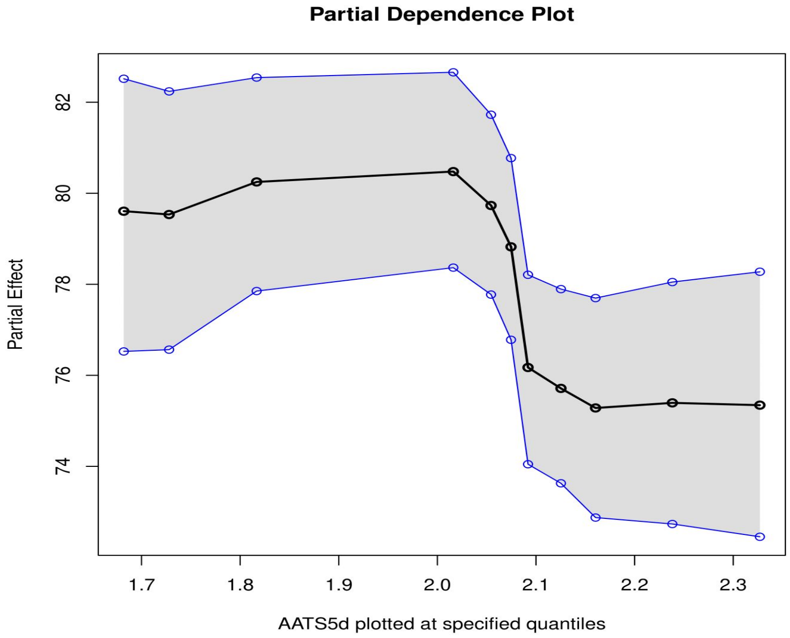
EU Optimization

- Idea: optimize expected utility to decide which structure to evaluate next
- Balance exploration vs exploitation
- Expected improvement: $\int \max\left(y - y^*, 0\right) \cdot p(y|x)dy$
- Probability of improvement: $\int \mathbb{I}(y > y^*) \cdot p(y|x)dy$
- MC estimation
- How do we find structures that maximize a given expected utility?
- Difficult… rely on chemists!

EU Optimization - Interpretability

- Chemist need to derive an **actionable hypothesis** from model output!

- Provide partial dependence of each covariate in output: $\mathbb{E}_{x_{-i}}\left[\sum_{j=1}^{m} g(x; T_j, M_j)\right]$



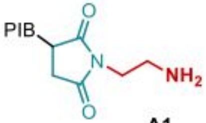
- But descriptors sometimes are difficult to interpret..

- In addition, some of the descriptors (neural embeddings) do not have interpretation!

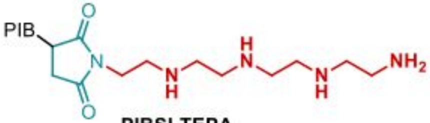
EU Optimization - Interpretability

Validation and chemical interpretation

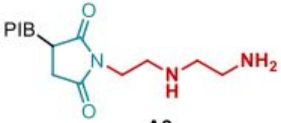
Density of amino groups in polar head



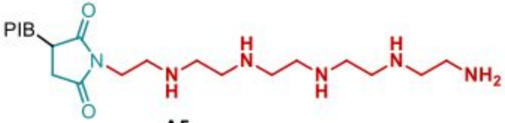
A1



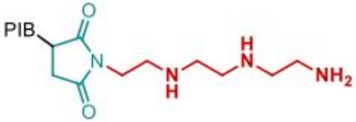
PIBSI-TEPA



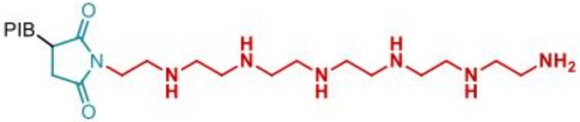
A2



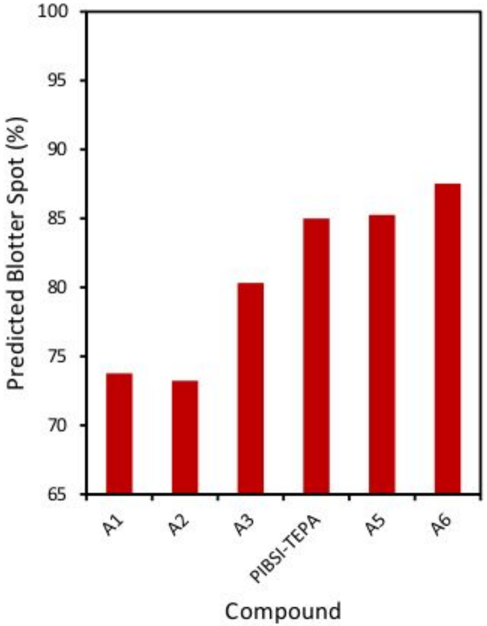
A5



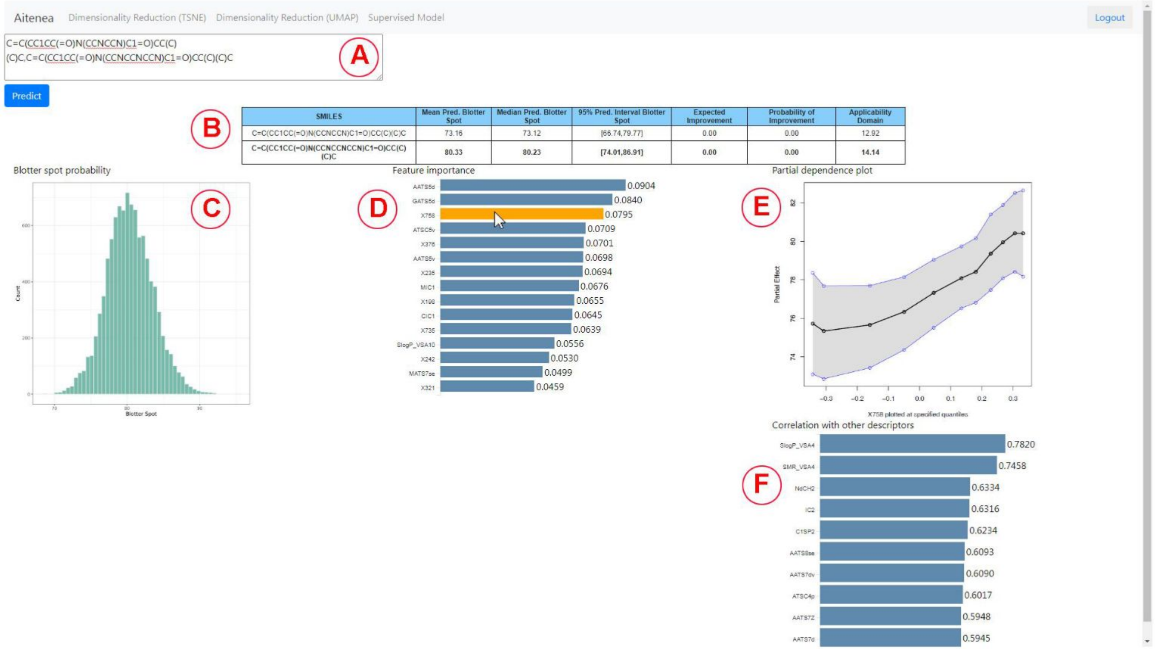
A3



A6



EU Optimization - Interpretability



C=C(CC1CC(=O)N(CCNCCN)C1=O)CC(C)(C)C,C=C(CC1CC(=O)N(CCNCCNCCN)C1=O)CC(C)(C)C

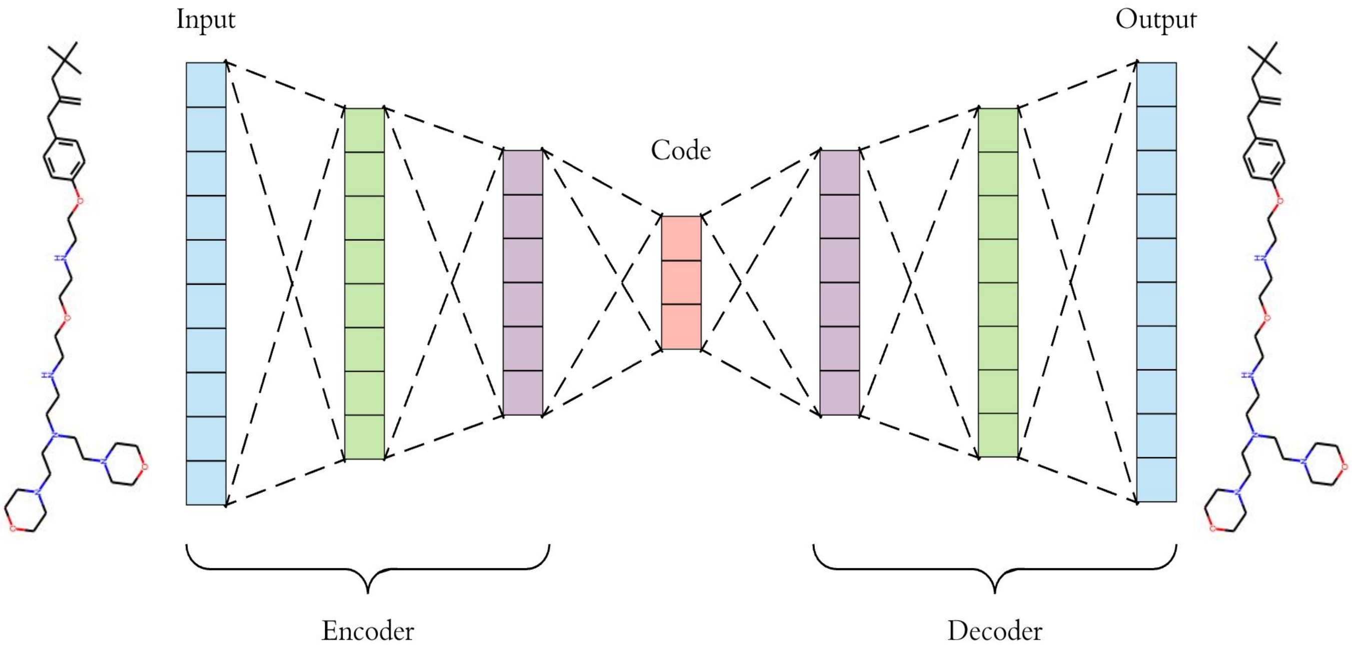
SMILES	Mean Pred. Blotter Spot	Median Pred. Blotter Spot	95% Pred. Interval Blotter Spot	Expected Improvement	Probability of Improvement	Applicability Domain
C=C(CC1CC(=O)N(CCNCCN)C1=O)CC(C)(C)C	73.16	73.12	[66.74,79.77]	0.00	0.00	12.92
C=C(CC1CC(=O)N(CCNCCNCCN)C1=O)CC(C)(C)C	**80.33**	**80.23**	**[74.01,86.91]**	**0.00**	**0.00**	**14.14**

- Other trends discovered these way, allowed chemists propose molecules with good expected improvement

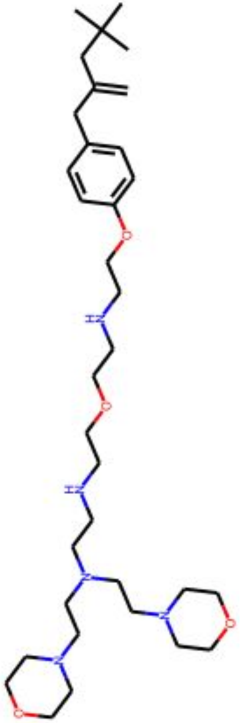
- Just one cycle of synthesis was enough for practical purposes

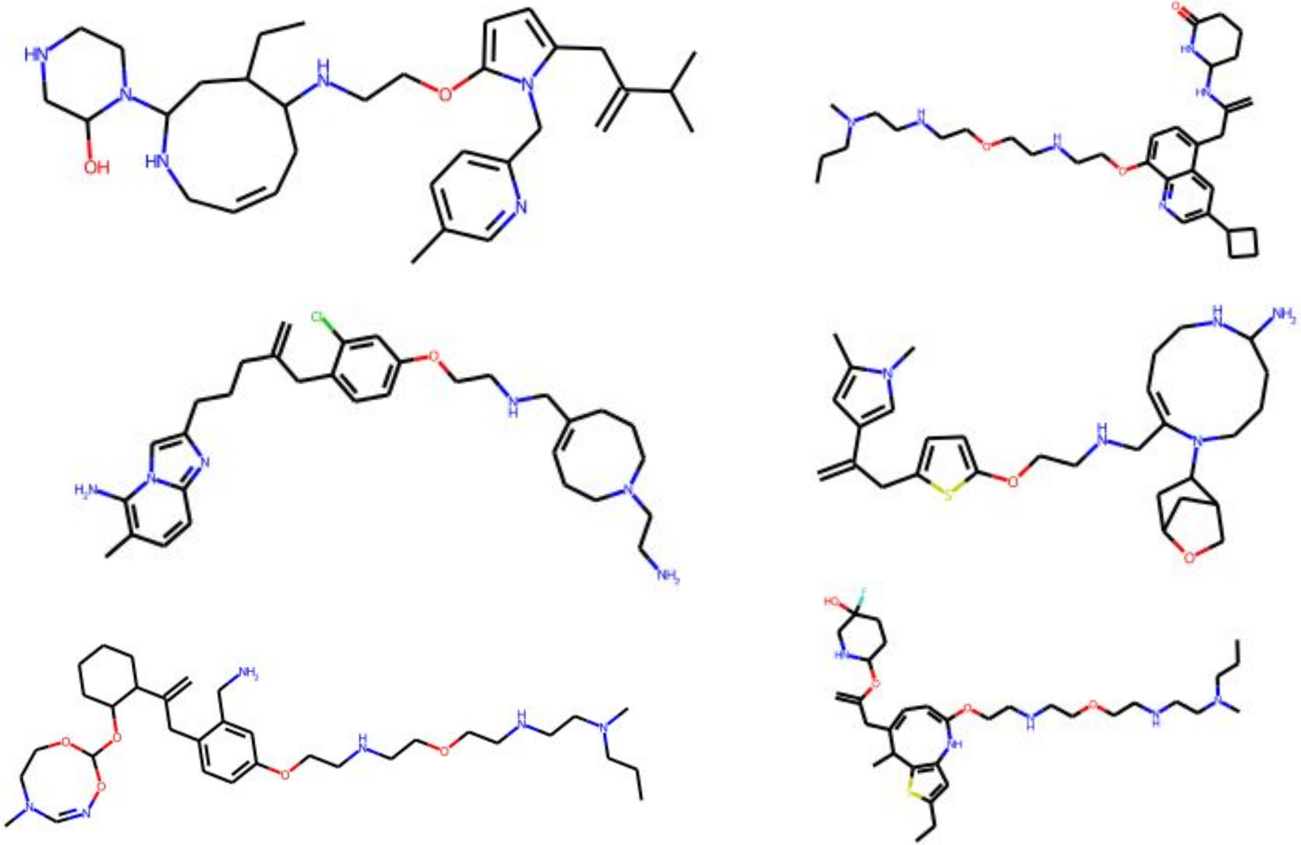
Molecular Generation on a Nutshell

- Goal: generate molecules that maximize Expected Utility
- Several approaches depending mainly on algorithm and molecular representation
- Deep Learning based (VAEs)



Generation Issues



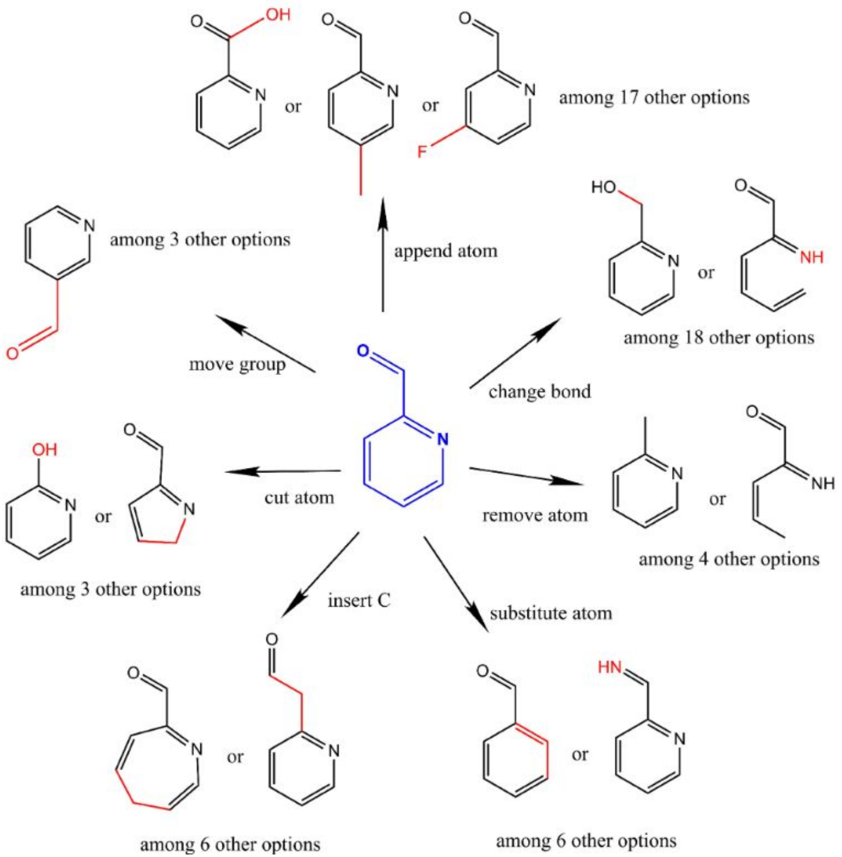


Discussion

- Statistical models can help accelerate molecular design
- Chemists need to interact with models. Interpretability is key (but very difficult)
- Removing humans from the process seems (almost) impossible. It would require automatic generation of new molecules
    - Multi-objective optimization
    - Small data regime
    - Structural constraints
    - Synthesizability
    - Uncertainty Quantification is key

Ongoing work

- Meta-heuristics for property optimization
- Genetic algorithms
- Iteratively mutate population of molecules (starting from a given one)



Acknowledgements

CONICET

UAM
Universidad Autónoma
de Madrid

FACULTAD DE
CIENCIAS
UNIVERSIDAD AUTÓNOMA DE MADRID

Altenea®
biotech

ICMAT
INSTITUTO DE CIENCIAS MATEMÁTICAS

REPSOL

Thanks!

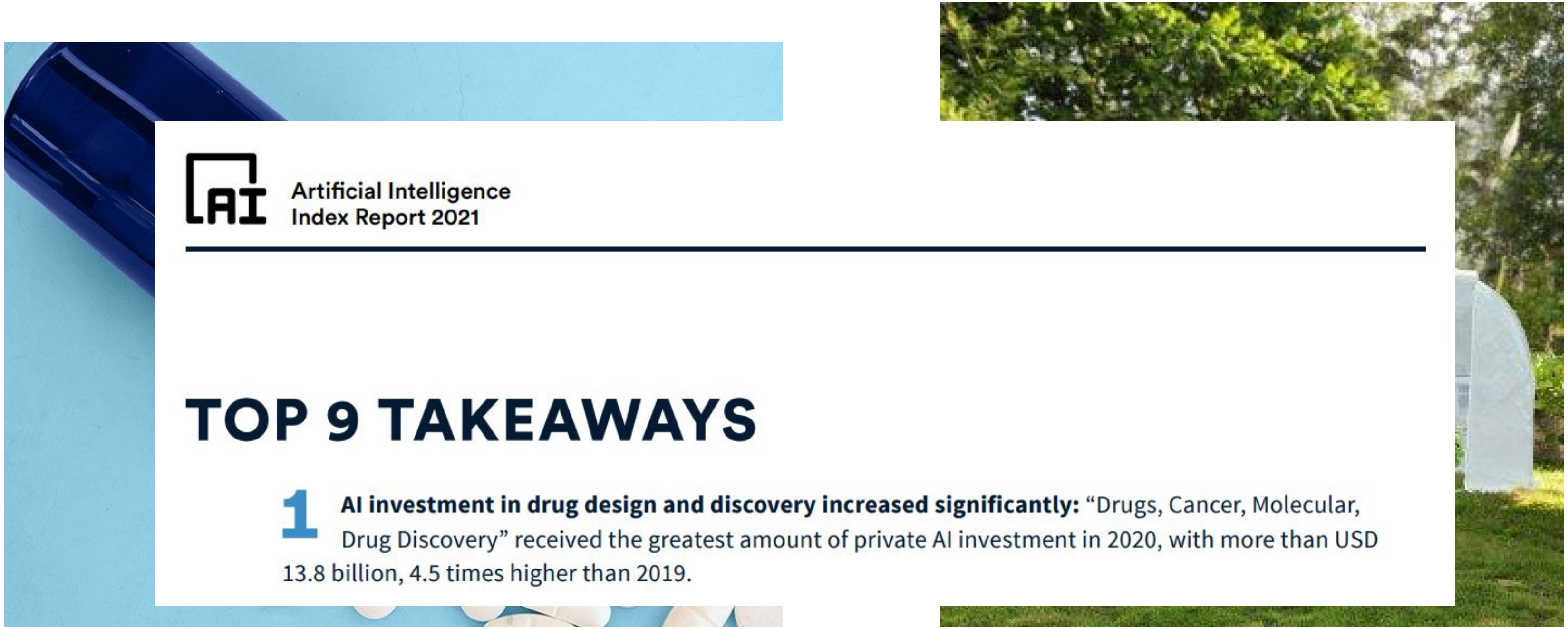
roi.naveiro@icmat.es

https://roinaveiro.github.io/

https://github.com/roinaveiro

Why?

Why?



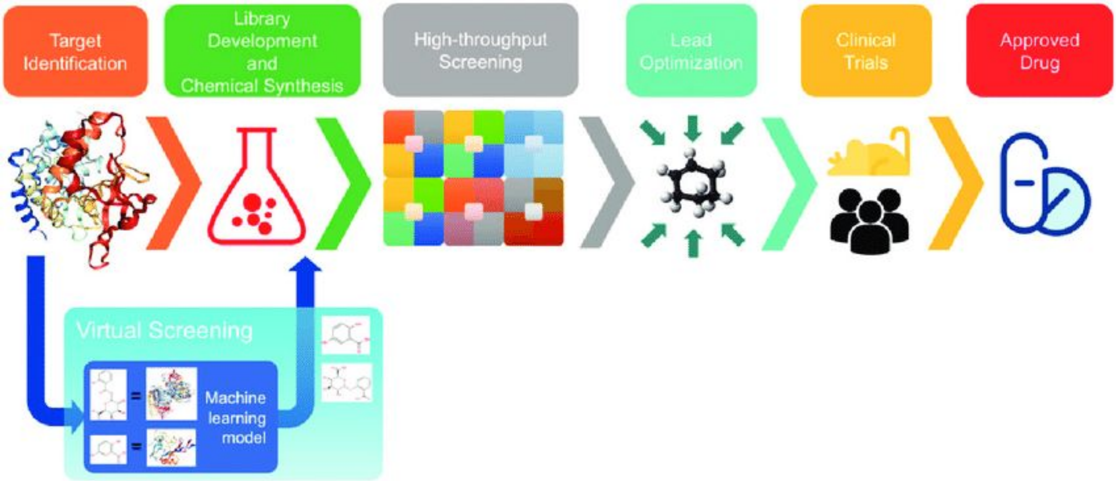
Artificial Intelligence
Index Report 2021

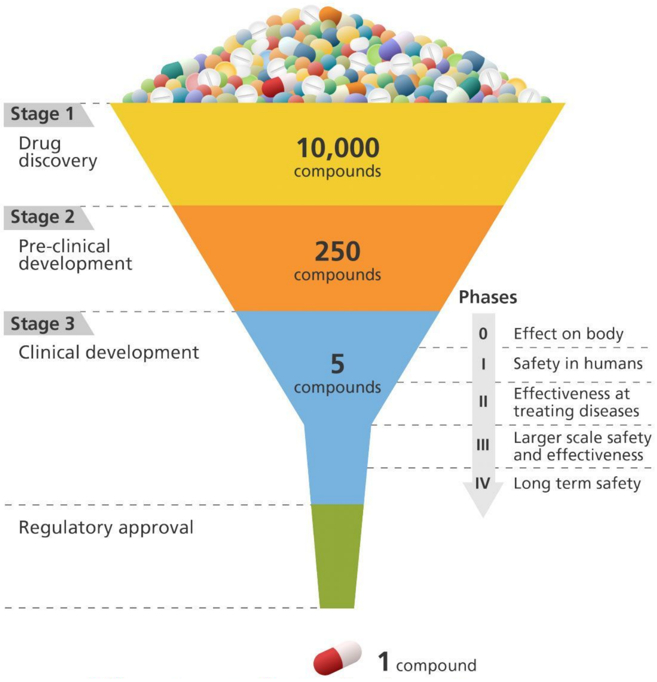
TOP 9 TAKEAWAYS

1 **AI investment in drug design and discovery increased significantly:** “Drugs, Cancer, Molecular, Drug Discovery” received the greatest amount of private AI investment in 2020, with more than USD 13.8 billion, 4.5 times higher than 2019.

The process of discovering new molecules

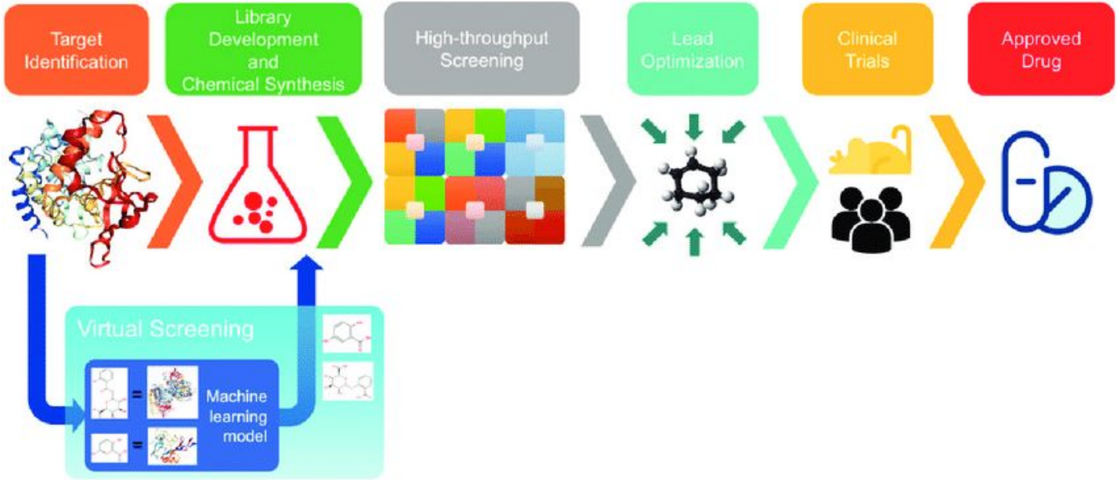
- Pharma: average time discovery - market, 13 years
- Outside pharma: 25 years
- Crucial 1st step: **generate pool of candidates**
- Daunting task (e.g. 10^{23} – 10^{60} drug-like molecules)





The old way and the soon-to-be-old way

- Old way
  - ○ Human experts propose, synthesize and test (*in vitro*)
- Soon-to-be-old way: high throughput virtual screening (HTVS)
  - ○ Predict properties through computational chemistry...
  - ○ ...leverage rapid **ML-based property predictions**



De novo molecular design

- Just existing molecules are explored
- Much time lost evaluating bad leads
- Traverse chemical space more “effectively”: reach **optimal molecules** with **less evaluations** than brute-force screening

> *“De novo molecular design is the process of **automatically proposing novel chemical structures** that **optimally satisfy desired properties**”*

Combinatorial, black-box, stochastic, multi-objective optimization with black-box constraints

Automatically proposing novel chemical structures

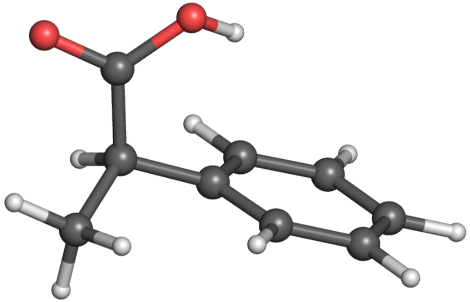
Two main ingredients

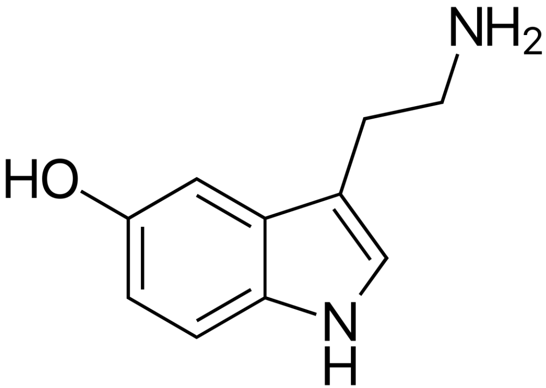
- Molecule representation
- Generative model

Representing molecules

Molecules are **3D QM objects** with: nuclei with defined positions surrounded by electrons described by complex wave-functions

- Digital encoding that serves as input to model
- **Uniqueness and invertibility**
- Trade-off: information lost vs complexity
  - 3D coord. representation (symmetries?)
  - More compact 2D (graph) representation
- 1D, 2D and 3D





1D representations - SMILEs

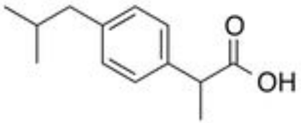
Simplified Molecular Input Line Entry System

Molecule as graph (bond length and conformational info is lost)

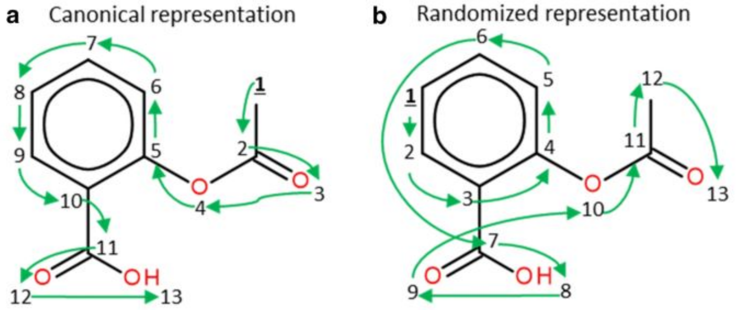
- Graph traversal
- Sequence of ASCII characters
- Non-unique → Canonical SMILES
- One-Hot-Encoding

- Leverage NLP techniques

- SMILE-based methods struggle to generate **valid** molecules
- Valid = valency rules
- Learn spurious grammar rules



Ibuprofen
CC(C)Cc1ccc(cc1)C(C)C(O)=O



CC(=O)Oc1ccccc1C(=O)O

c1cc(c(cc1)C(O)=O)OC(C)=O

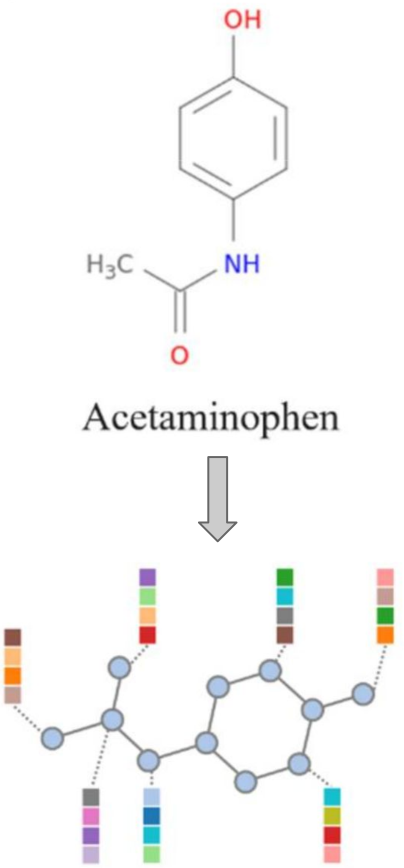
2D representations

- Nodes represent atoms
- Edges represent bonds
- Nodes/Edges have associated features (atom number, bond type, etc.)

- Capture connectivity!
- Symmetry invariant representation

- More difficult to generate than sequences
- Taylored algorithms that work with graphs (composing transformations on graphs, symmetries?)

- Graph Neural Nets!



3D representations

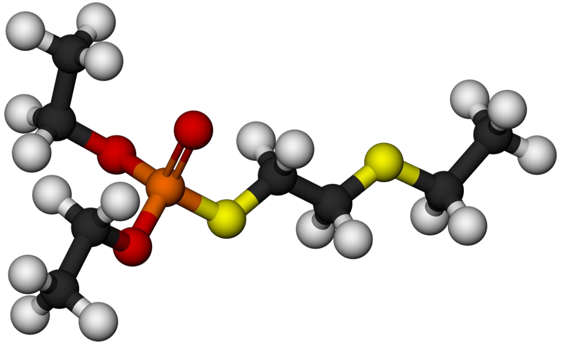
- 3D point clouds

$$\mathcal{M} = \{x_i, r_i\}_{i=1}^{p}$$ where x_i are features and r_i are coordinates.

- Minimal information lost (conformational preferences, bond lengths, etc.)

- Symmetries?
- Too many degrees of freedom

- Generation: sequentially choose pair of atoms, relative position, bond length and angles



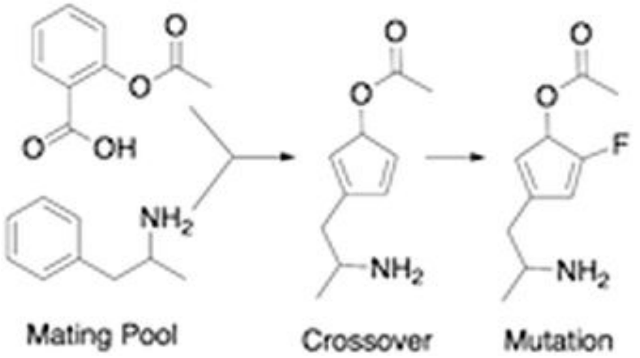
How to generate molecules?

Myriad of different ways. A useful distinction:

- Gradient-free methods
- Gradient-based methods

Gradient Free Methods

- Graph-based genetic algorithms
  - Mutations and crossover on a pool of candidates
  - Elitist natural selection rule
- Yoshikawa et. al. propose using SMILES
  - Population of SMILES
  - Grammatical Evolution
- Many more...

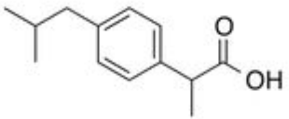


Gradient Based Methods

- Recurrent Neural Networks
- (Variational) Autoencoders
- Normalizing Flows
- Generative Adversarial Networks (GANs)

Recurrent Neural Networks

- Work on sequences (SMILES)



Ibuprofen
CC(C)Cc1ccc(cc1)C(C)C(O)=O

- Goal: given training sequences → learn to generate new sequences that resemble those of training.

- Sequence: $S_{1:T} = (S_1, \ldots, S_T)$ where $S_i \in \mathcal{V}$

- Training: maximum likelihood, equiv to minimize loss function:

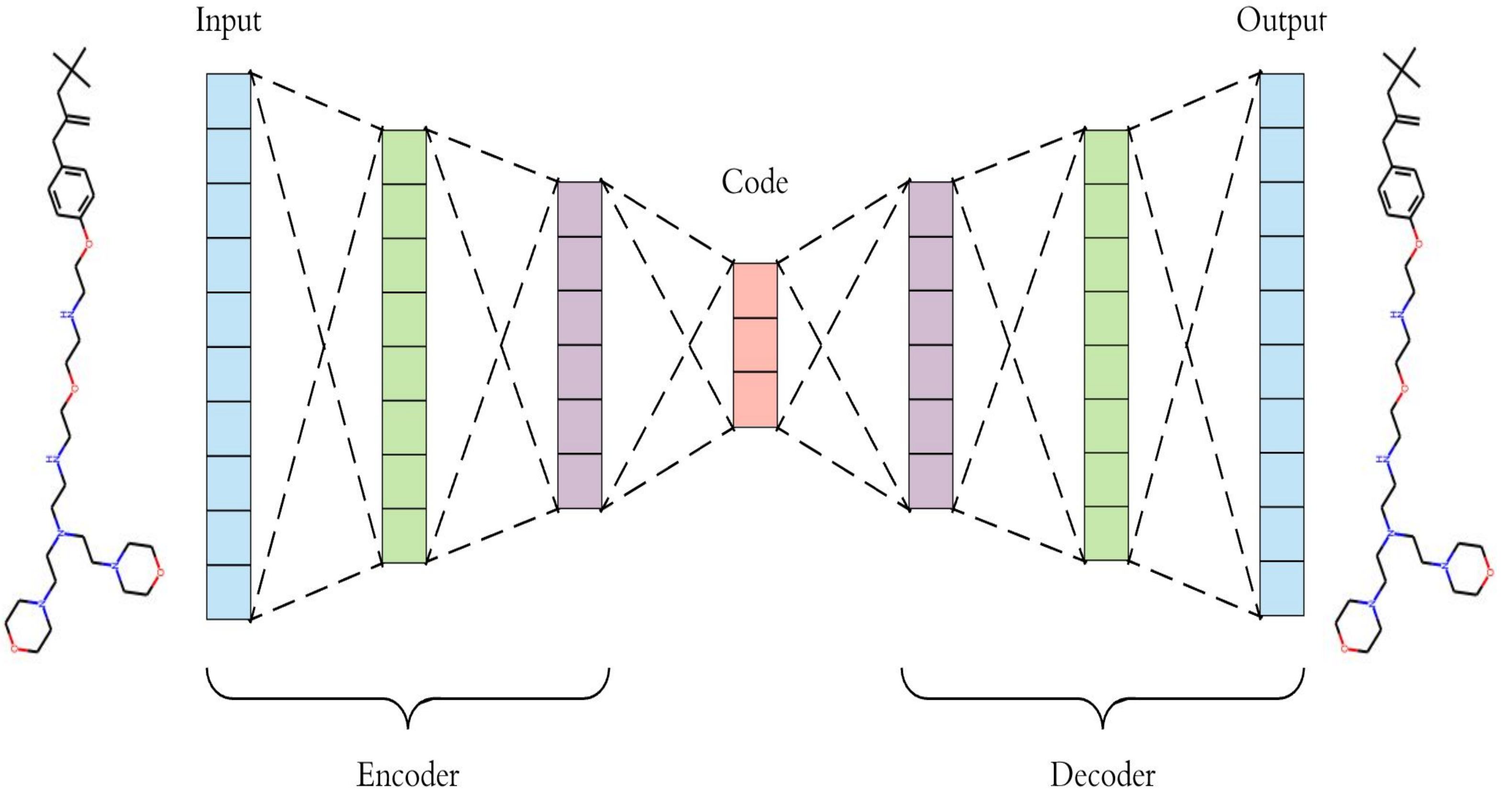
$$L^{MLE} = -\sum_{s \in \mathcal{T}} \sum_{t=2}^{T} \log \pi_\theta(s_t | S_{1:T-1})$$

- Generation: sequentially sample from multinomial dist.

- Thermal rescaling

$$\hat{p}_i \propto \exp(\frac{p_i}{T})$$

(Variational) Autoencoders



Variational Autoencoders

- Goal: learn probabilistic latent variable model for data generation

$$z \sim p(z)$$
$$x \sim p_\theta(x|z)$$

- We want to maximize $p(x) = \int p_\theta(x|z)p(z)dz$; instead maximize

$$\log p(\boldsymbol{x}) \geq \mathrm{E}_{z \sim q_\phi(z|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}|z)p(z)}{q_\phi(\mathbf{z}\,|\,\boldsymbol{x})}\right]$$

- RHS is equal to

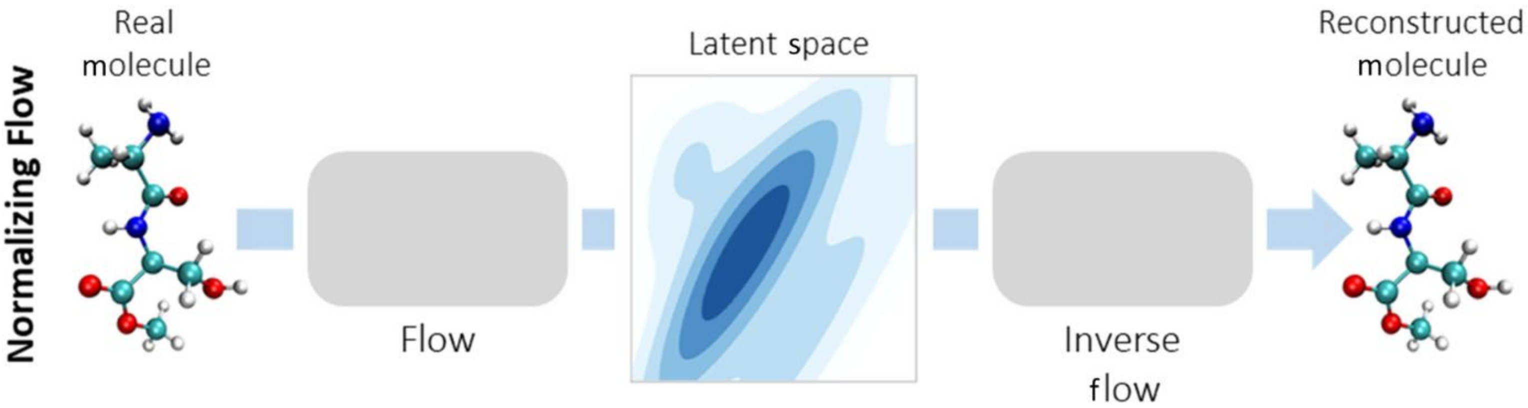
$$\mathbb{E}_{z \sim q_\phi(z|x)}\left[\log p_{\theta(x|z)}\right] - D_{KL}[q_\phi(z|x), p(z)]$$

Variational Autoencoders

- Typically: $p(z)$ independent standard normal dist. and $q_\phi(z|x)$ factorized multivar. normal
- Mean and variance functions of encoder parameterized through CNN.
- Decoder normally RNN

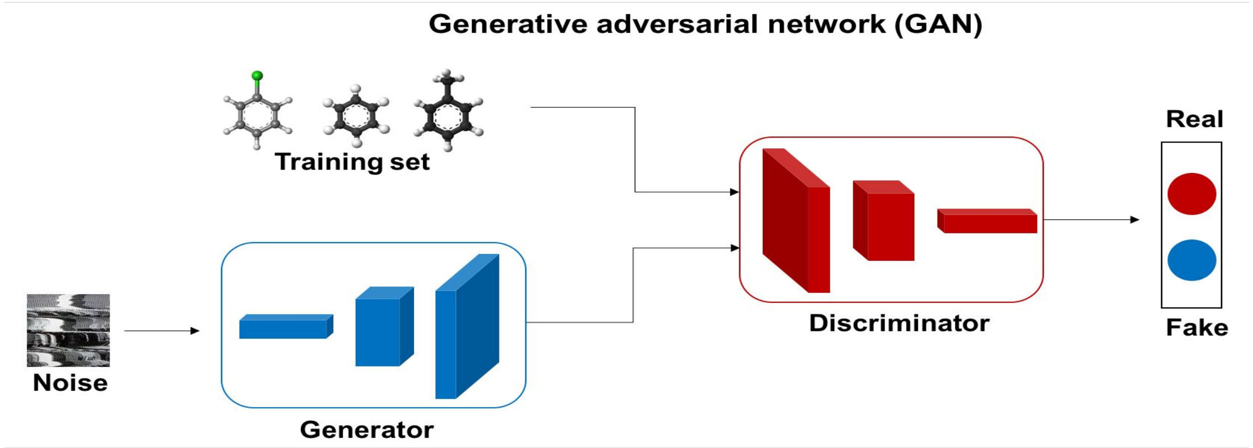
- Training
  - Encode each training sample x into z
  - Decode z into x’
  - Minimize loss function
- Generation
  - Get point in latent space z
  - Decode z sampling $x \sim p_\theta(x|z)$

Normalizing Flows



- Learn series of parametric bijective transformations of probability distributions
- Allows (easy) calculation of exact likelihood.
- Deep NN with bijective layers

Generative Adversarial Networks



- Generator: generate molecule from Gaussian noise
- Discriminator: distinguish real from fake molecules
- Train to compete against each other

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \in p_d(x)}\left[\log D(x)\right] + \mathbb{E}_{z \in p_z(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$

Recall that...

"De novo molecular design is the process of ***automatically proposing novel chemical structures*** *that* ***optimally satisfy desired properties****"*

Combinatorial, black-box, stochastic, multi-objective optimization with black-box constraints

Generate molecules that optimally satisfy desired properties

- Goal: learn valid molecules with **desirable properties**
- Infeasible to measure properties experimentally for every generated molecule...
- Infeasible to use computational chemistry to compute properties...
- **Prediction**: quantitative structure-activity relationship (QSAR)
- Done usually in separate datasets
- Many models depending on property, representation, etc.
  - Molecular Descriptors
  - SMILEs
  - Graphs

Using properties to guide generation

1. Reinforcement Learning coupled with sequence generator

- A time t, state is $(s_0, \ldots, s_t)$
- Action is next token $a_t = s_{t+1}$
- After taking action, a reward R_t is perceived
- Goal, learn policy $\pi_\theta(a|s)$

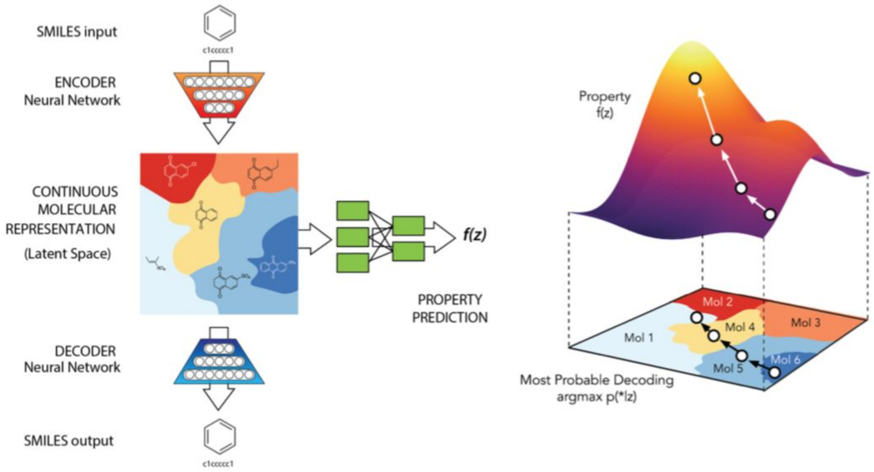
$$\mathbf{max}_\theta\, \mathbb{E}[\textstyle\sum_{i=1}^{T} R_i | s_0, \theta]$$

- The only non-zero reward is $R_T$which is equal to the property prediction

Using properties to guide generation

2. Optimization with VAE

   - Learn map from latent space to property (e.g. through GP)
   - Optimize that map (gradient ascent, bayesian optimization, etc.)



Issues/Thoughts

- Multi-objective optimization
  - Many properties to be optimized (depending even on different stakeholders!)
  - Drug discovery: **high binding affinity to biological target**, low toxicity, solubility, synthetically accessible, stability, economical costs!
  - Commonly: predict properties independently and combine predictions in loss function.
  - Also, hold properties constant implicitly through structural constraints.
  - **Decision theory**: **multi-attribute utilities** to incorporate different objectives for different stakeholders into the generative process

Issues/Thoughts

- Uncertainty quantification
  - Models rely on predictions to generate promising molecules
  - Accuracy of these models is key
  - In small data regimes... models tend to be less accurate.
  - Incorporate uncertainty quantification into generative process! (**Bayesian inference**)
  - Exploration vs exploitation (**Bayesian optimization**)
  - **Bayesian decision theory**

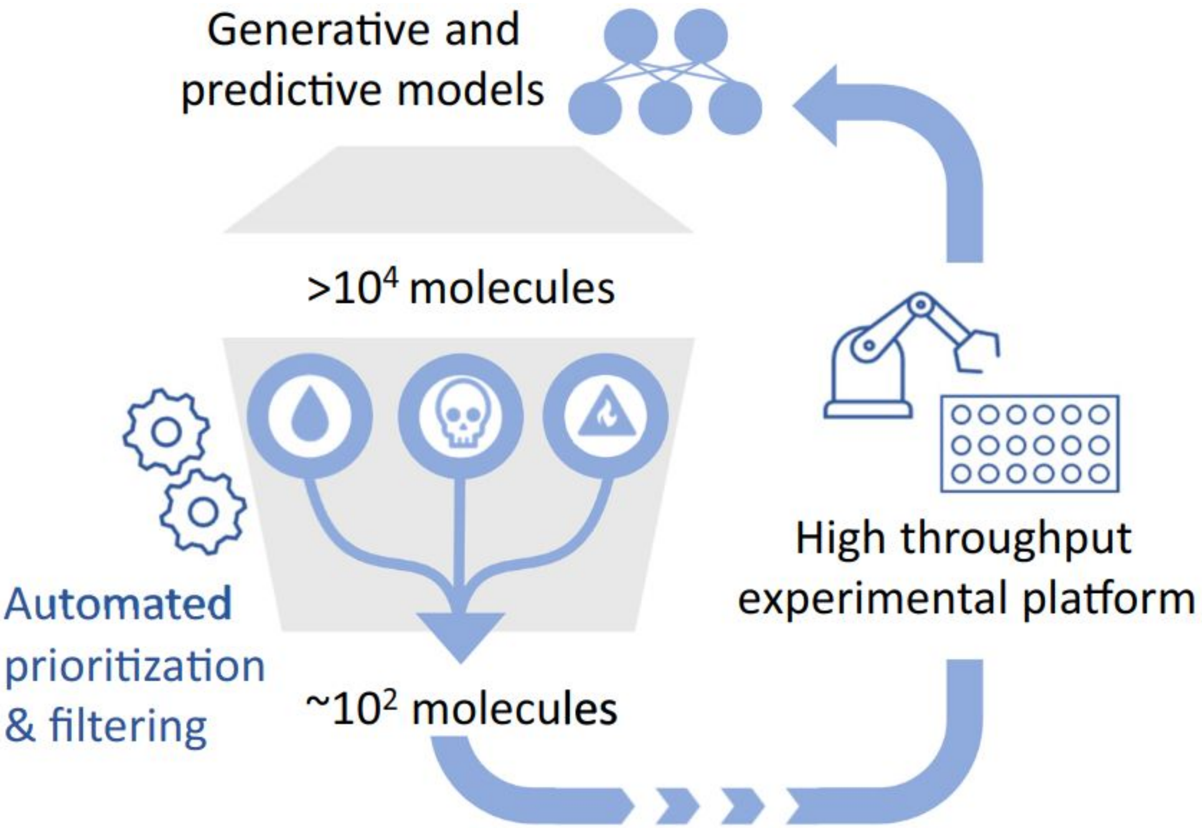
Issues/Thoughts

- Synthesizability
  - Generated molecules must be easy to synthesize
  - This concept is hard to define!
  - Methods to automatically evaluate synthesizability without human intervention
  - Rather than molecules, generate synthetic pathways (learn reactions)

Other relevant fields

- Graph based deep learning
- Geometric deep learning
- Combinatorial black-box optimization
- Heuristic search algorithms
- Reinforcement Learning

The dream - Closing the loop


Generative and predictive models
>10^4 molecules
Automated prioritization & filtering
~10^2 molecules
High throughput experimental platform

The reality?

- More likely: computer-aided molecular design
- Interpretability
    - Prediction is not enough, we need understanding (?).
    - Chemist need to derive an actionable hypothesis from model output.
    - If chemist sees, e.g. structural elements responsible for toxicity, she might have ideas on how to modify molecule to diminish toxicity
    - Interpretable representations: molecular descriptors...?
    - Interpretable methods to determine **causality between structure presence and property** (**causal inference, counterfactual inference**)

References

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., & Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. Wiley Interdisciplinary Reviews: Computational Molecular Science, e1608.

Meyers, J., Fabian, B., & Brown, N. (2021). De novo molecular design and generative models. Drug Discovery Today, 26(11), 2707-2715.

Elton, D. C., Boukouvalas, Z., Fuge, M. D., & Chung, P. W. (2019). Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering, 4(4), 828-849.

Gallego, V., Naveiro, R., Roca, C., Ríos Insua, D., & Campillo, N. E. (2021). AI in drug development: a multidisciplinary perspective. Molecular Diversity, 25(3), 1461-1479.

Yoshikawa, N., Terayama, K., Sumita, M., Homma, T., Oono, K., & Tsuda, K. (2018). Population-based de novo molecule generation, using grammatical evolution. Chemistry Letters, 47(11), 1431-1434.

Unconstrained generation

- Goal: learn general distribution of molecules in chemical space
- Evaluated based on chemical validity, novelty, uniqueness