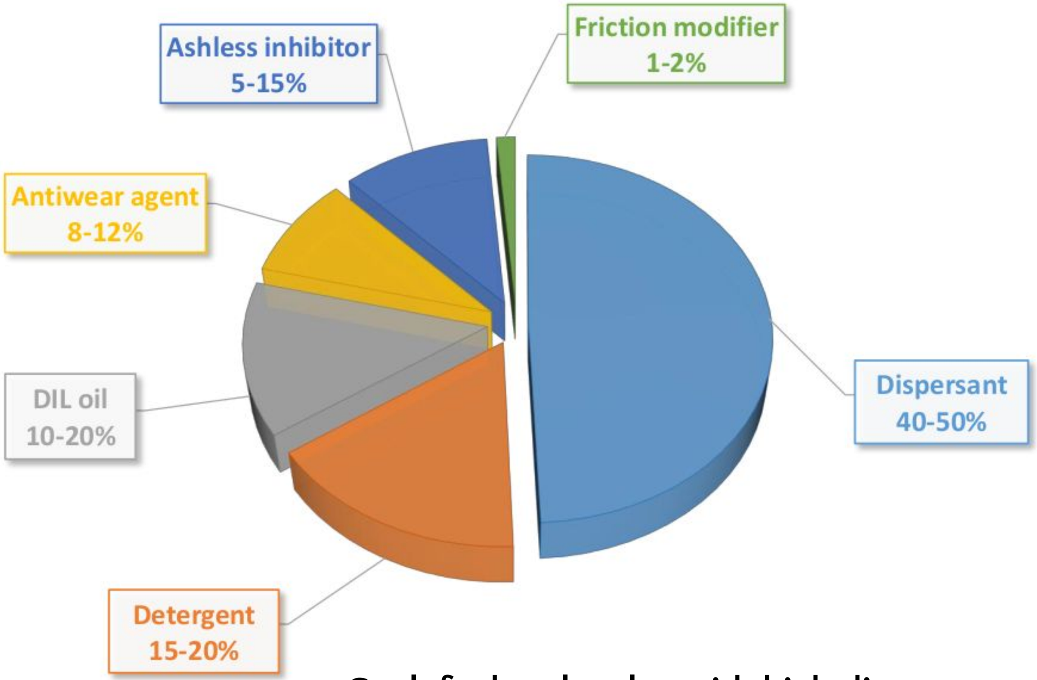
ICMAT
INSTITUTO DE CIENCIAS MATEMÁTICAS

Machine Learning for Molecular Design: a case study in dispersant design

Roi Naveiro

SEIO - 2022

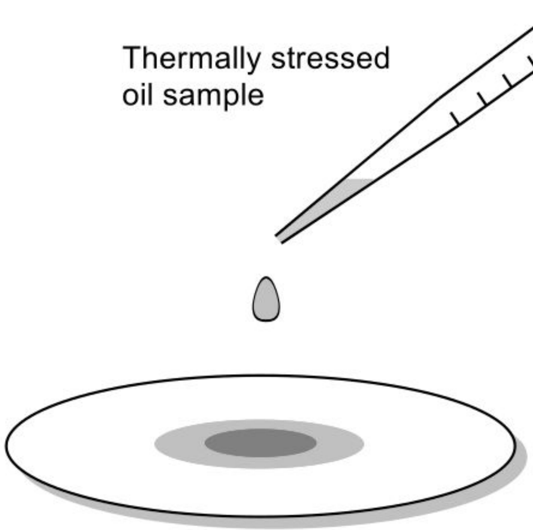
Dispersants in Lubricants

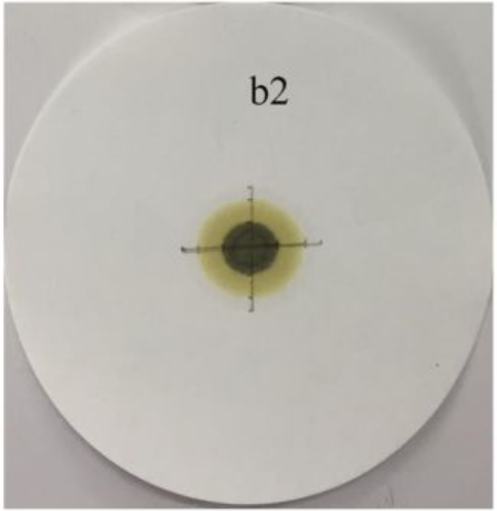


- Lubricants for combustion engines require formulated additive package (dispersants)
- Under harsh operating conditions of engines, soot is produced.
- Soot aggregation increases lubricant viscosity causing corrosion, deposit formation…
- Dispersants are molecules that adsorbs onto the surface of ultrafine carbon deposit precursors reducing their aggregation.

Goal: find molecules with high dispersancy efficacy

Measuring Dispersancy Efficacy - Blotter Spot

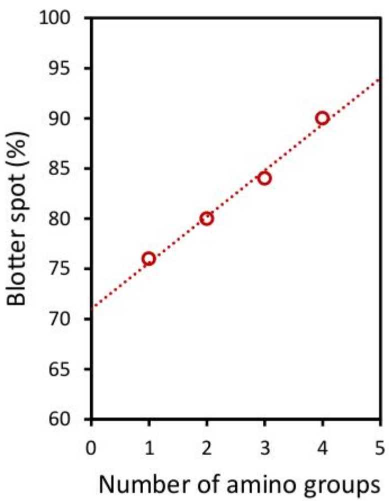


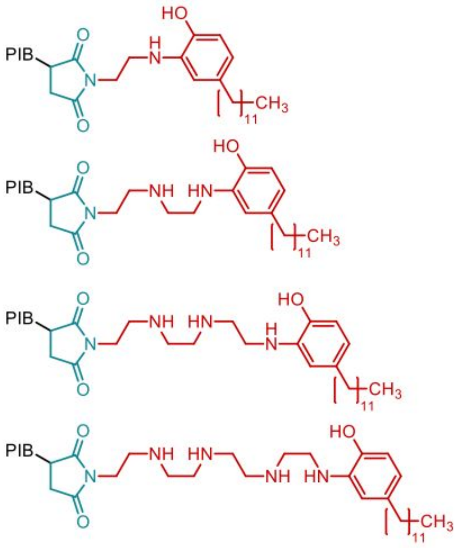


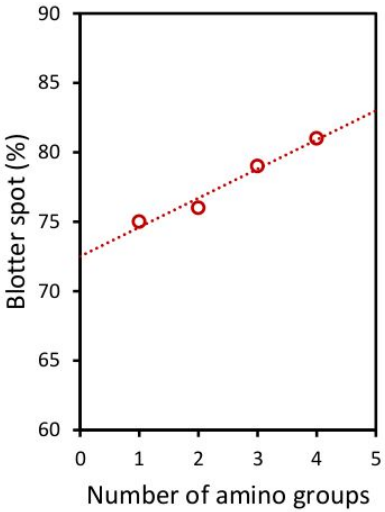
$$\text{Blotter Spot Dispersancy (\%)} = \frac{\text{diameter of black spot}}{\text{diameter of the total spot}} \text{ X } 100$$

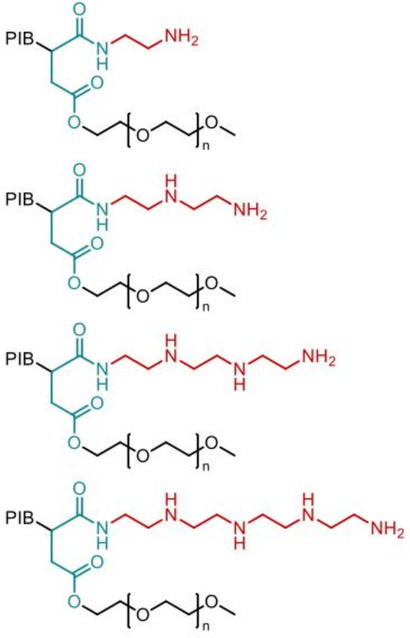
Dispersancy Estimation – Limits of Chemist Intuition

Within a family of substrate, predictable behaviors are appreciable.







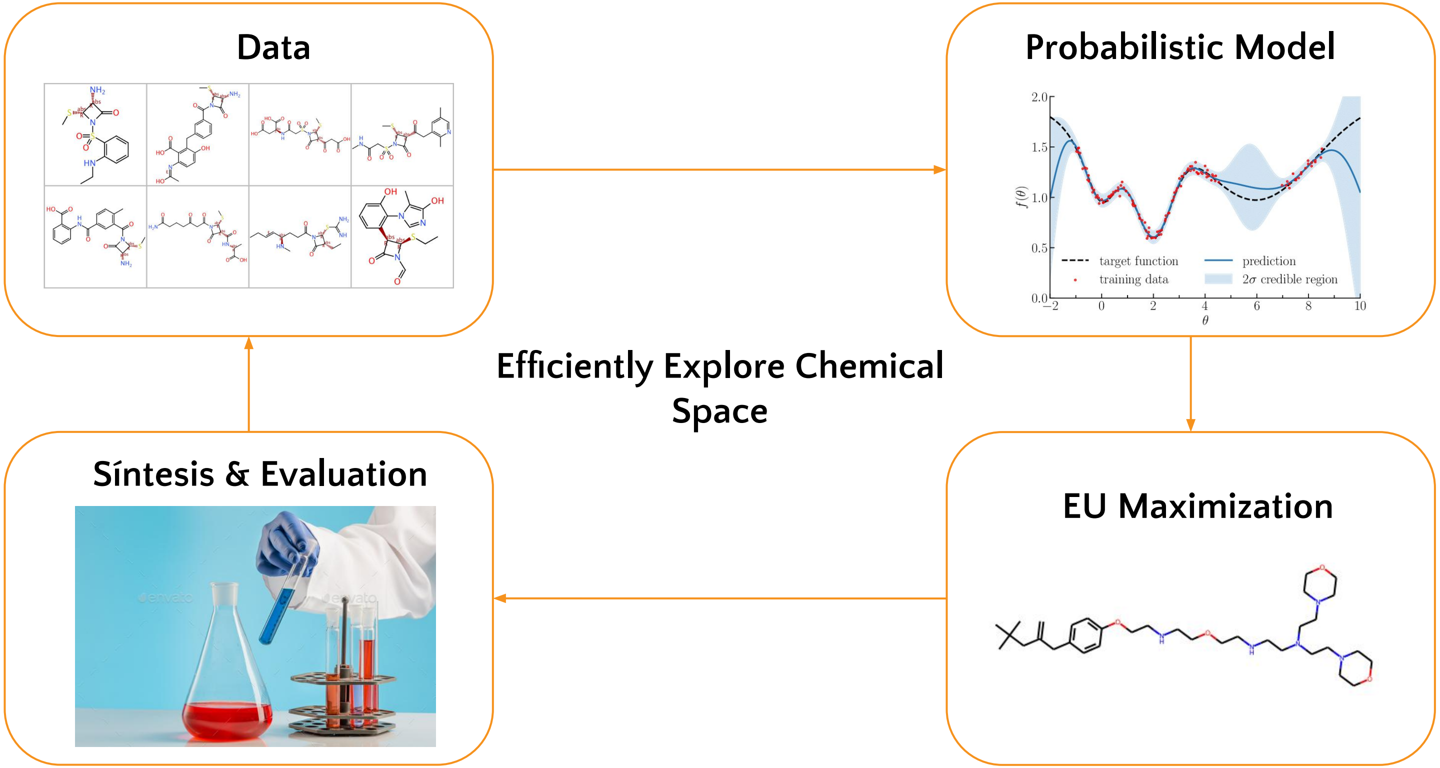


- However, the relationship between different families of substrates cannot be determined intuitively

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

Leverage data to find molecular structure with high blotter spot…

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



Data

Probabilistic Model

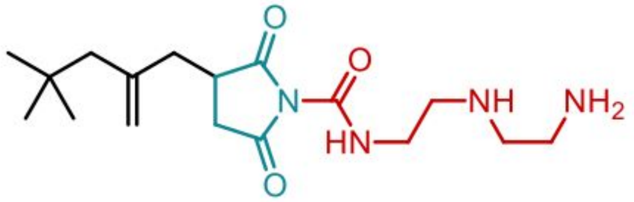
Efficiently Explore Chemical Space

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%

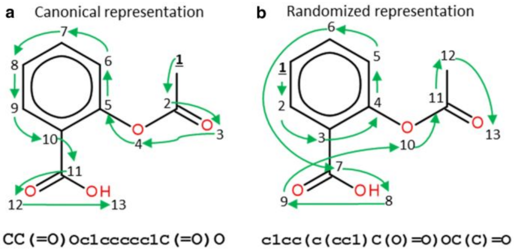
SMILES string:

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)


a Canonical representation
b Randomized representation
O
OH
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)

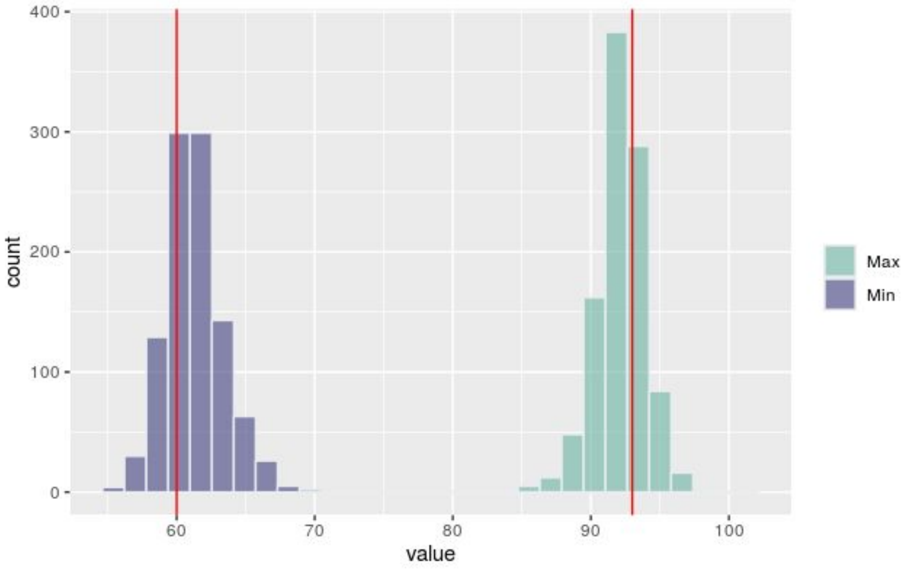
- Better predictive performance than: linear regression with horseshoe prior, GP.

Probabilistic Model for Dispersancy - Prediction

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

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

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

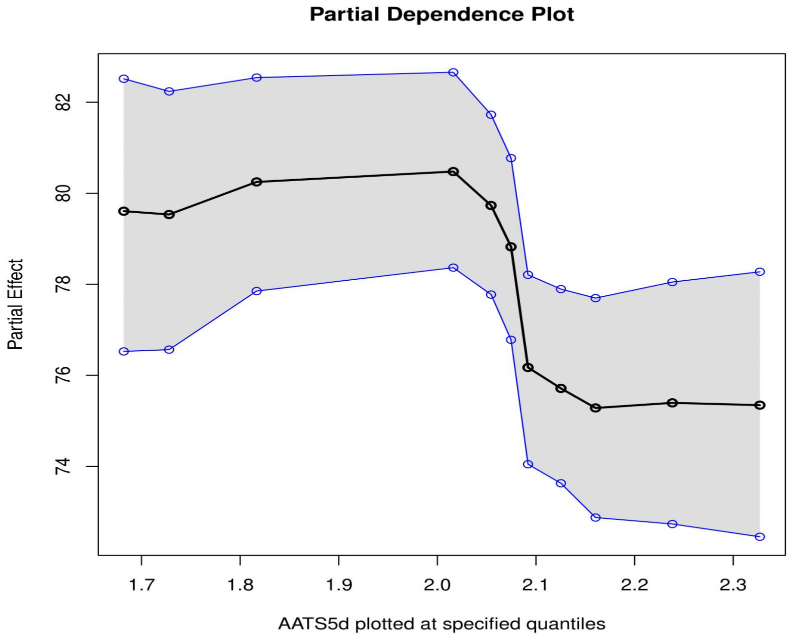


EU Optimization

- 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: $\mathbb{E}_{x_{-i}}\left[\sum_{j=1}^{m} g(x; T_j, M_j)\right]$

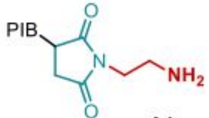


- But descriptors sometimes are difficult to interpret..
- In addition, some of the descriptors (neural embeddings) do not have interpretation!

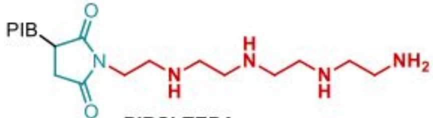
EU Optimization - Interpretability

Validation and chemical interpretation

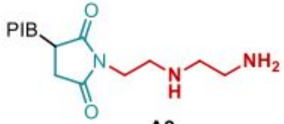
Density of amino groups in polar head



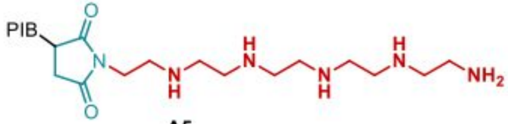
A1



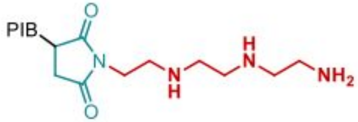
PIBSI-TEPA



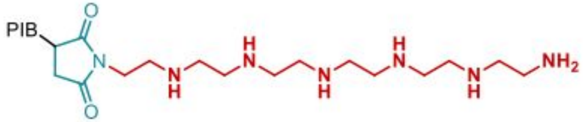
A2



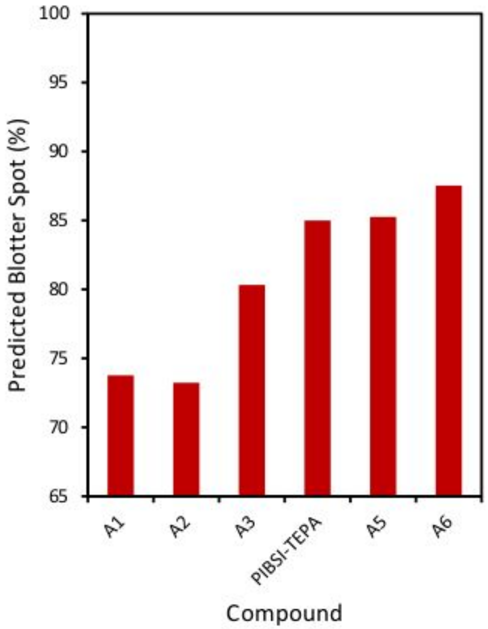
A5



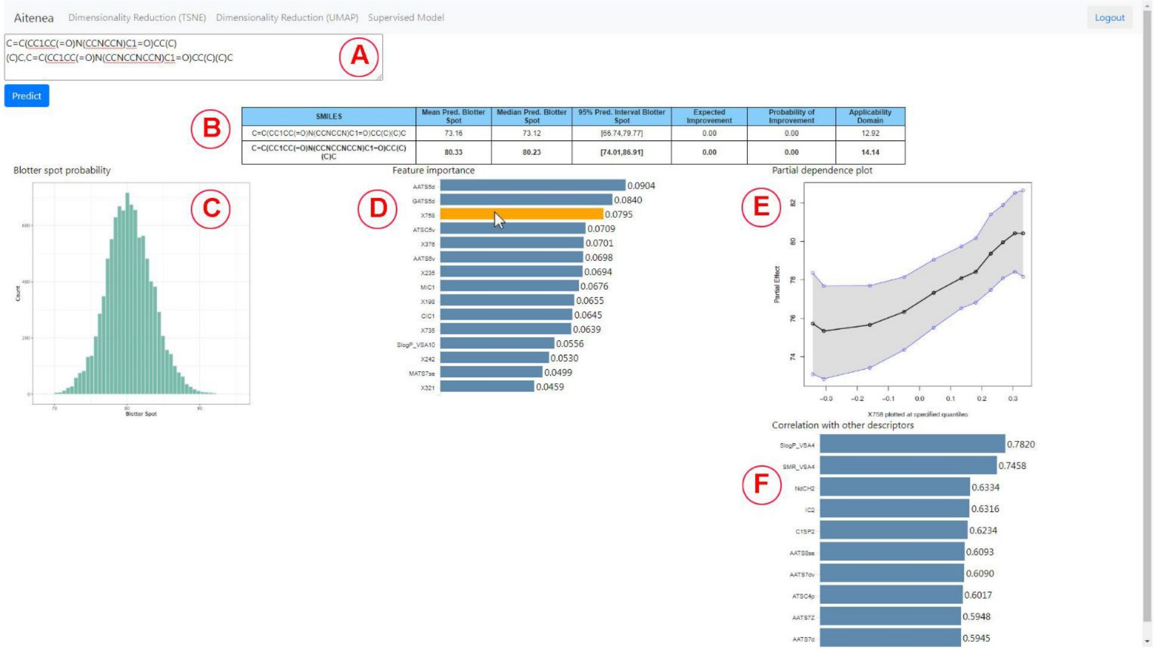
A3



A6



EU Optimization - Interpretability


Aitenea
Dimensionality Reduction (TSNE)
Dimensionality Reduction (UMAP)
Supervised Model
Logout
A
Predict
B
C
D
E
F

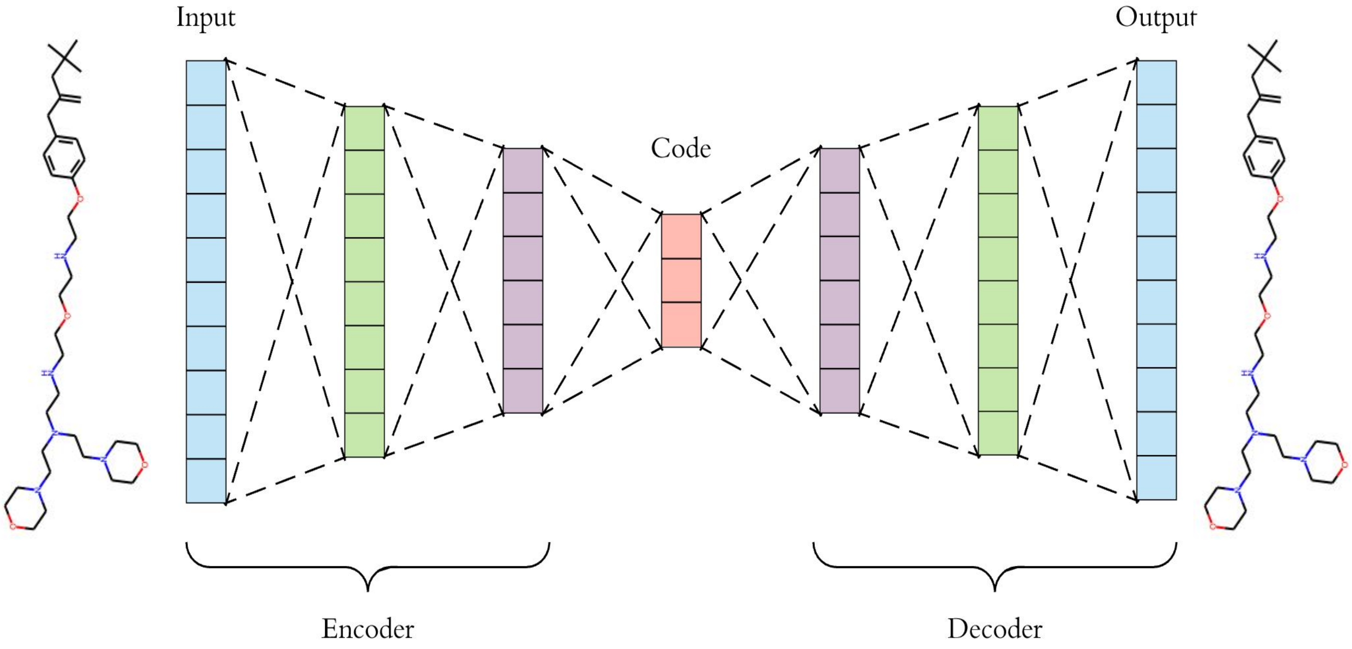
C=C(CC1CC(=O)N(CCNCCN)C1=O)CC(C)(C)C,C=C(CC1CC(=O)N(CCNCCNCCN)C1=O)CC(C)(C)C

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**

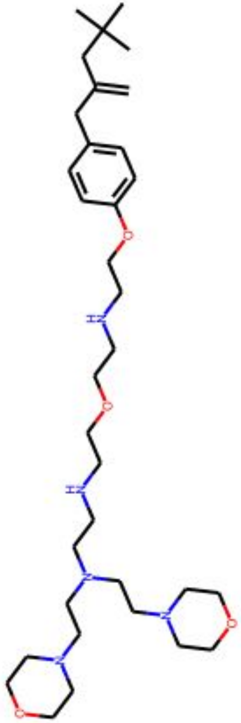
- Other trends discovered these way, allowed chemists propose molecules with good expected improvement
- Just one cycle of synthesis was enough for practical purposes...

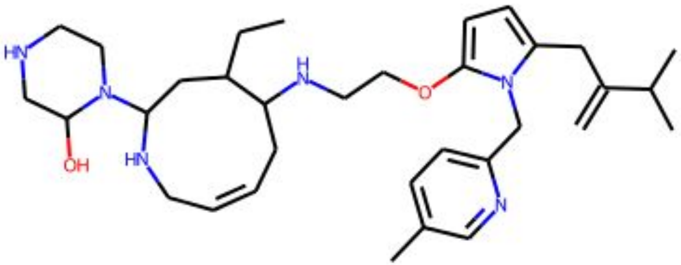
Molecular Generation on a Nutshell

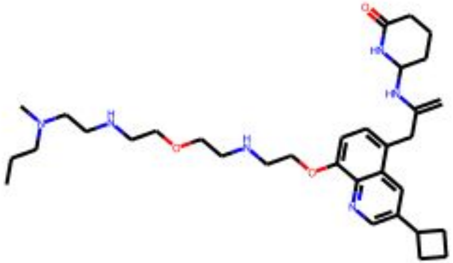
- Goal: generate molecules that maximize Expected Utility
- Several approaches depending mainly on algorithm and molecular representation
- Deep Learning based (VAEs)

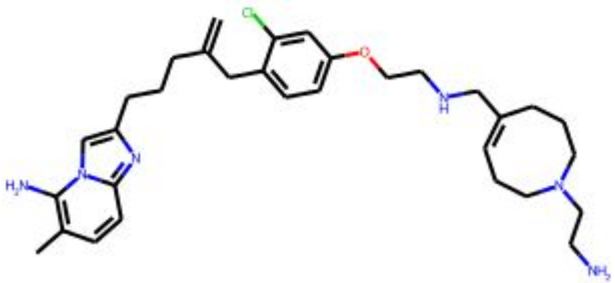


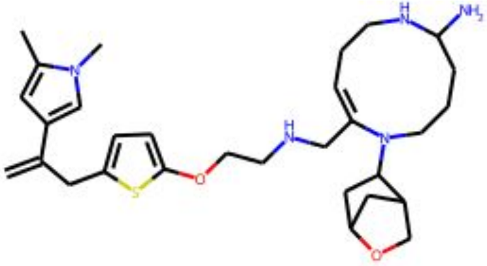
Generation Issues

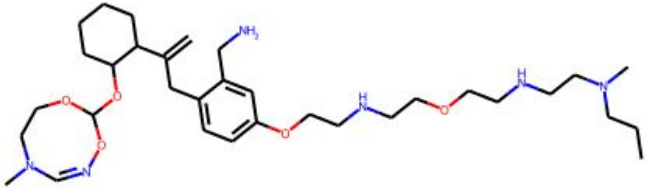


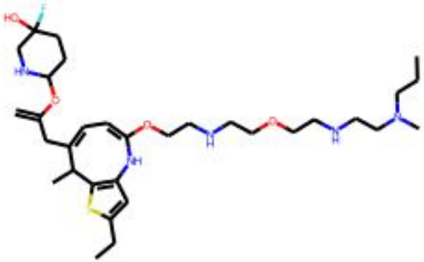










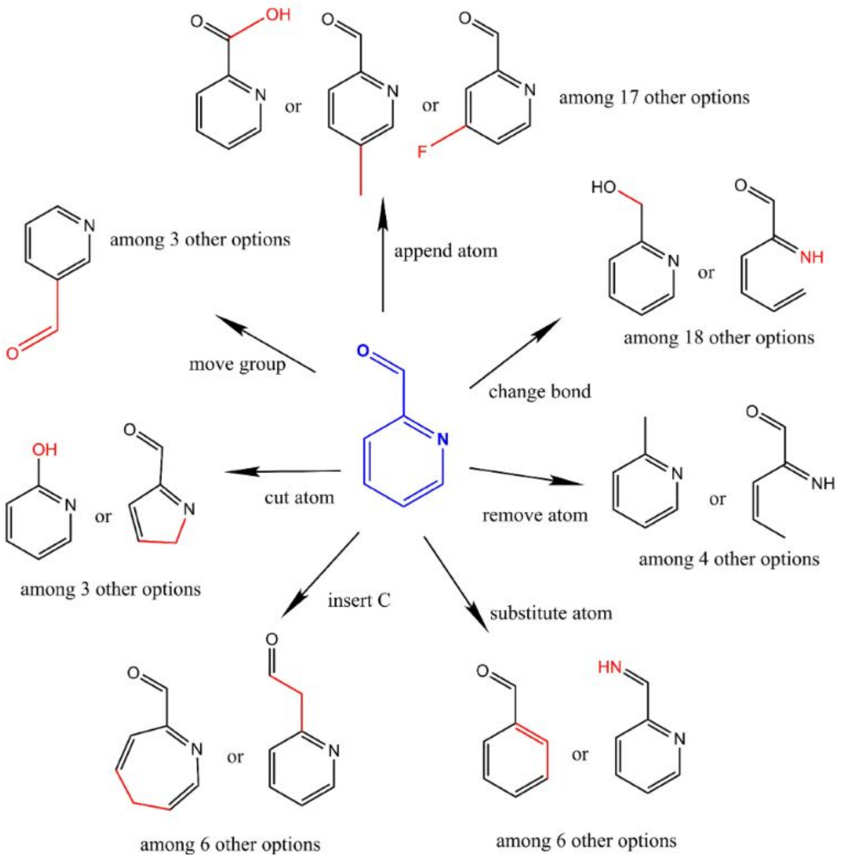


Discussion

- 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