Altenea® biotech

ICMAT
INSTITUTO DE CIENCIAS MATEMÁTICAS

Some ML applications in Online Marketing and Molecule Design

Roi Naveiro

From academia to industry

With **a bit** of creativity, **some** knowledge and **a lot** of effort; you can do incredible stuff...

From academia to industry

With **a bit** of creativity, **some** knowledge and **a lot** of effort; you can do incredible stuff...

...if you know some math/stats/programming!

What are we gonna see

- Online Marketing
- Demo 1
- Molecular Design
- Demo 2

Online Marketing - The problem

- Xeerpa collects data from social loggings
    - Likes on facebook
    - Posts in Twitter
    - Photos in IG...
- **Goal (at large):** process this information and analyse it to improve marketing decisions
- Many things to be done!
- We will see how to process information coming from:
    - Likes
    - Images

Online Marketing - Information coming from (Facebook) likes

- Facebook defines many categories such as: IPAs, Veggie Food, Soccer, Rock ‘n’ Roll
- Every category contains many Facebook pages (that users could like)

$$q = (q_1, \ldots, q_T) \text{ where } q_i = 1$$

if page belongs to category and 0 otherwise

- Similarly, users are represented as vector d.

- Goal: score every user in every category

Online Marketing - Scoring people based on Facebook likes

"A common problem in Information Retrieval (IR) is the following: given a corpus of documents, each of them represented by a sequences of words, how to find the more relevant documents to a given query. This problem reduces to assigning a score to a (query, document) pair."

This is the same! Words are Facebook pages, Users are documents, Categories are queries

Online Marketing - Scoring people based on Facebook likes

- IR assigns a number for each word in each document, that weights the importance of a word in a document
- Assign a weight to each like
- Two thoughts:
  - If there is no like, what should be the weight?
  - Should a like to Real Madrid be as important as a like to Cultural y Deportiva Leonesa?
- TF-IDF (as in IR)

Online Marketing - Scoring people based on Facebook likes

- Term Frequency
  - 1 if like is present 0 otherwise
  - 1 / (Number of likes)

- Inverse document frequency (how much info a like provides?)

$$\log\left(\frac{N}{n_t+1}\right)+1$$

- tf-idf = tf * idf

Online Marketing - Scoring people based on Facebook likes

- Each user is a vector

$$v(d) \text{ where } v(d)_i \text{ tf-idf of the i-th like}$$

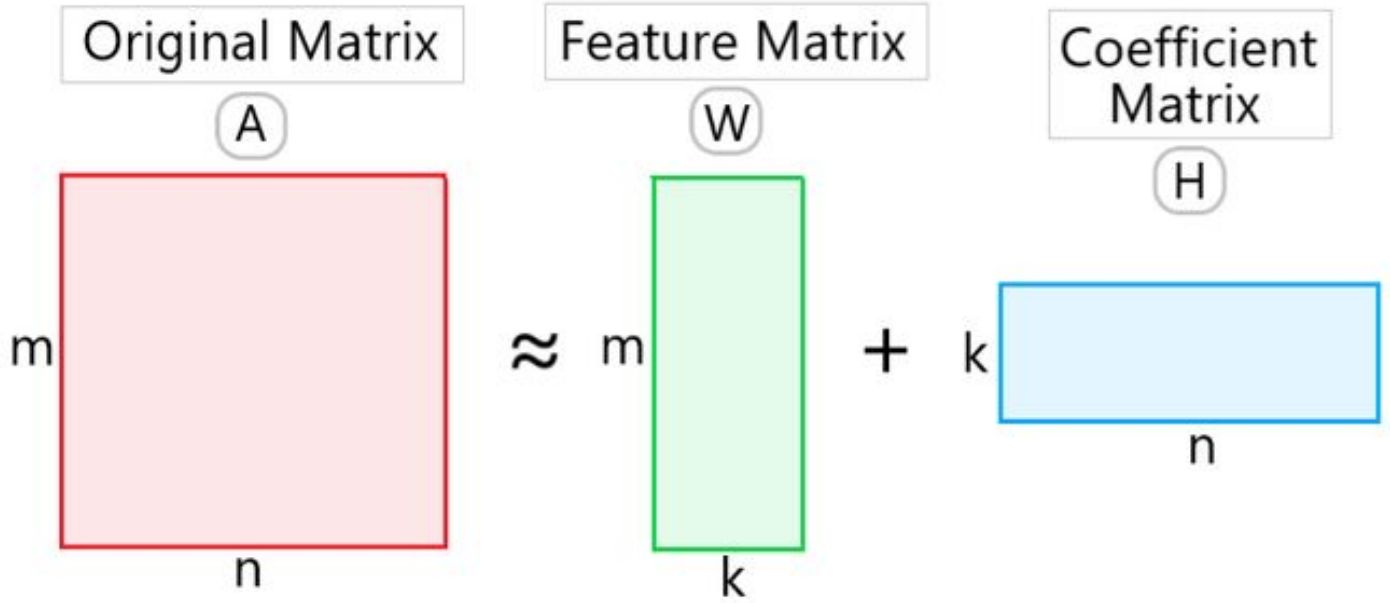
- Same for categories q!
- A common score

$$\text{score}(d, q) = \frac{v(d) \cdot v(q)}{|v(d)||v(q)|}$$

- Lives in [0,1]
- User with no likes in category will have 0
- User liking all pages in category (and with no other likes) will have 1

Online Marketing - Community detection

Detect communities of similar users

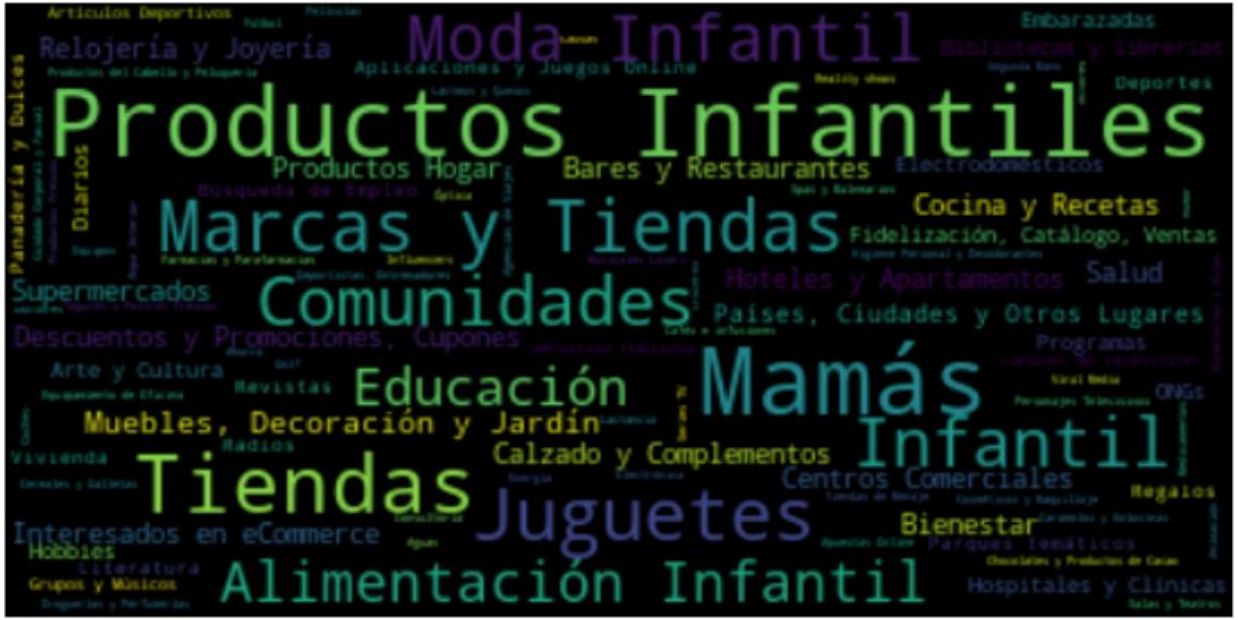


Minimize reconstruction error

$$\|A - WH\|^2 \sum_{i=1}^{n} = \sum_{j=1}^{n} (X_{ij} - [WH]_{ij})^2$$

Online Marketing - Community detection

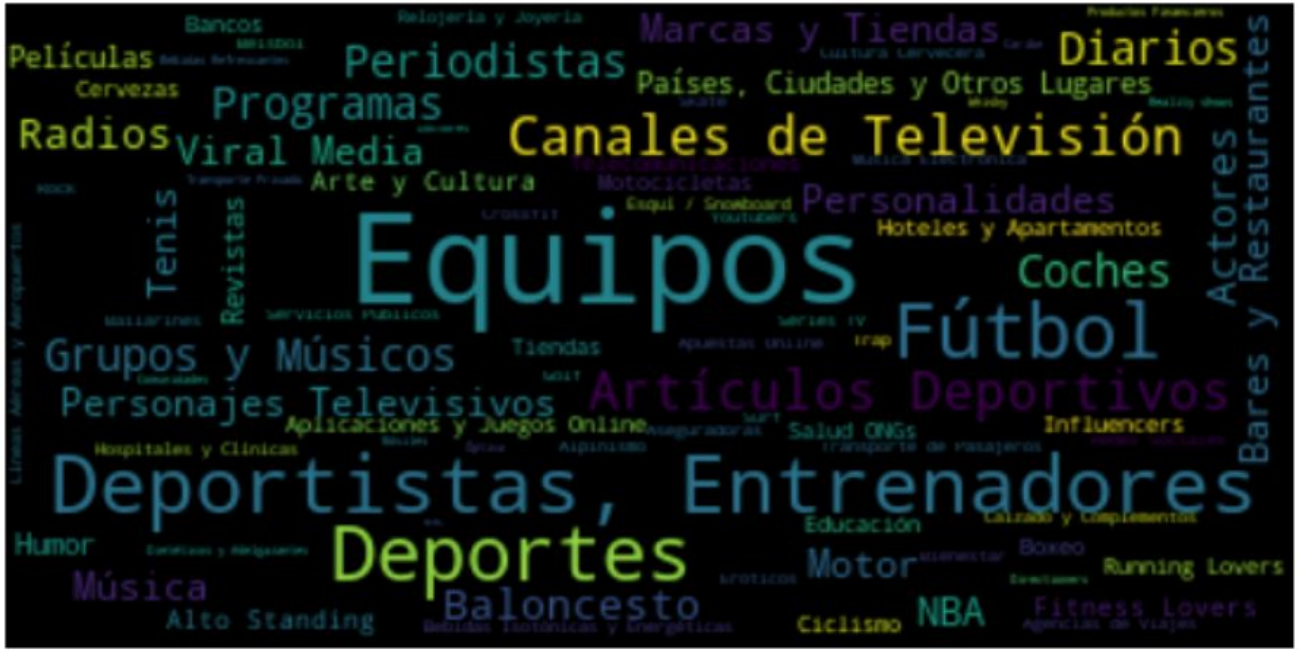
Women between 30 and 50



Online Marketing - Community detection

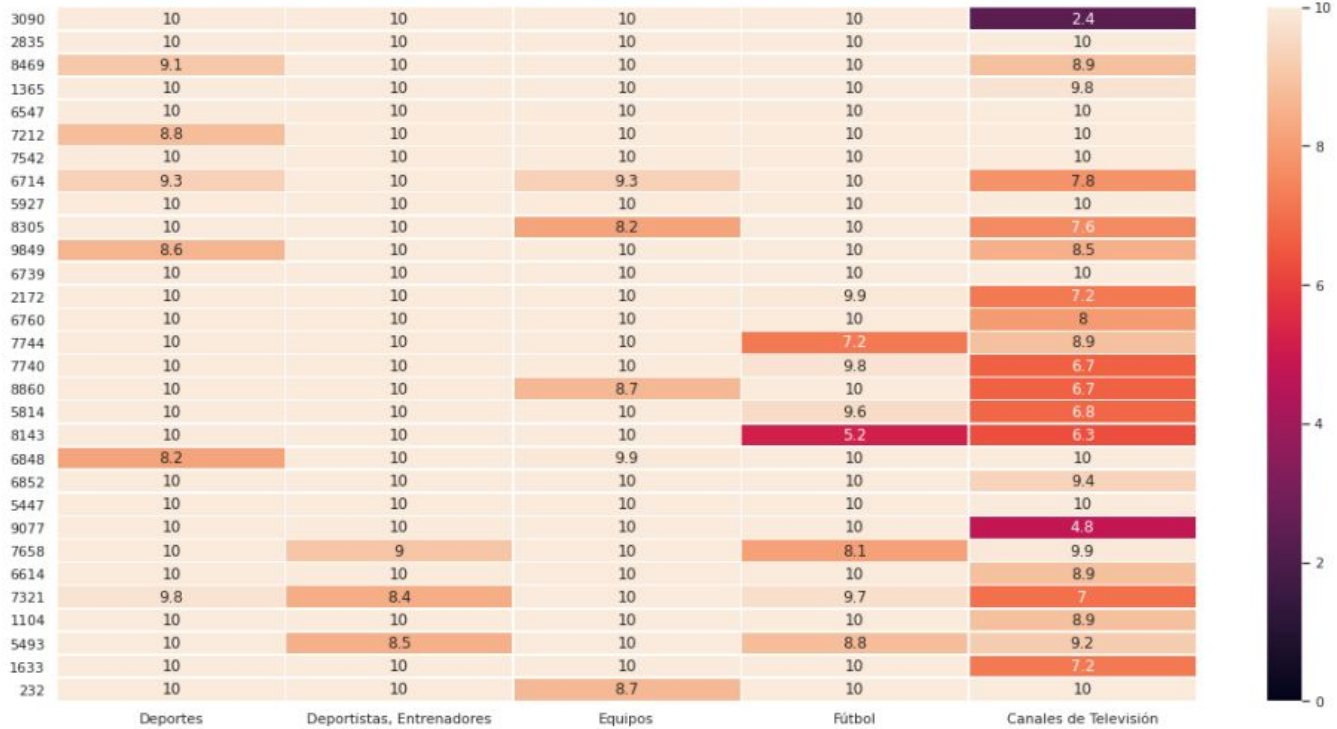
Online Marketing - Community detection

Men between 30 and 50



Online Marketing - Community detection

Men between 30 and 50

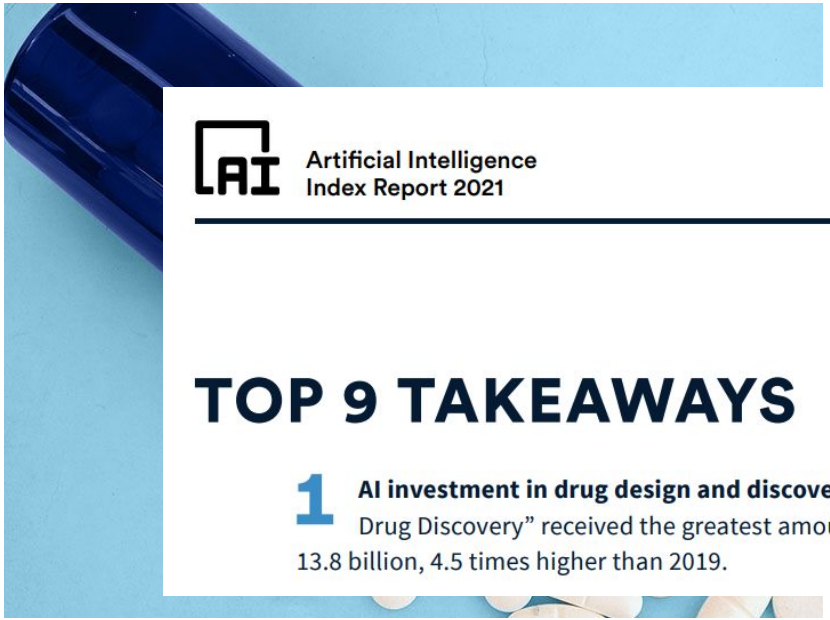


Online Marketing - Scoring based on images!

- How to do score users in categories based on their IG images?
- We need to associate each image to a category or group of categories
- This has to be done automatically!
- **Demo**

Molecular design - Why?

Molecular design. Why?



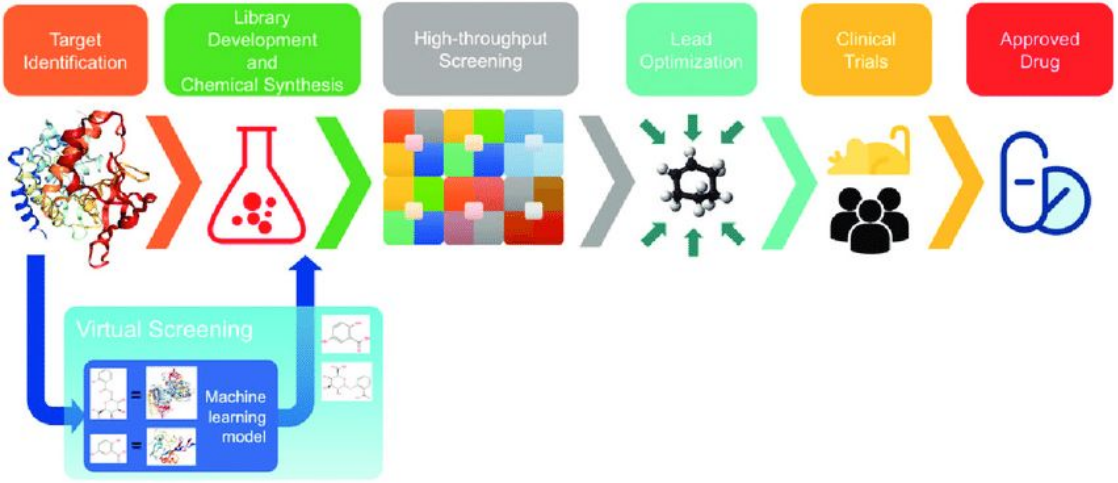
Artificial Intelligence
Index Report 2021

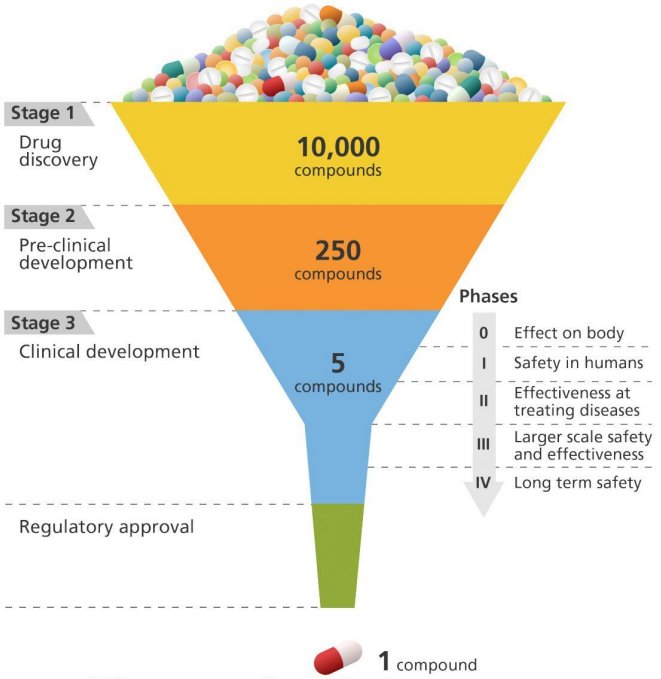
TOP 9 TAKEAWAYS

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.

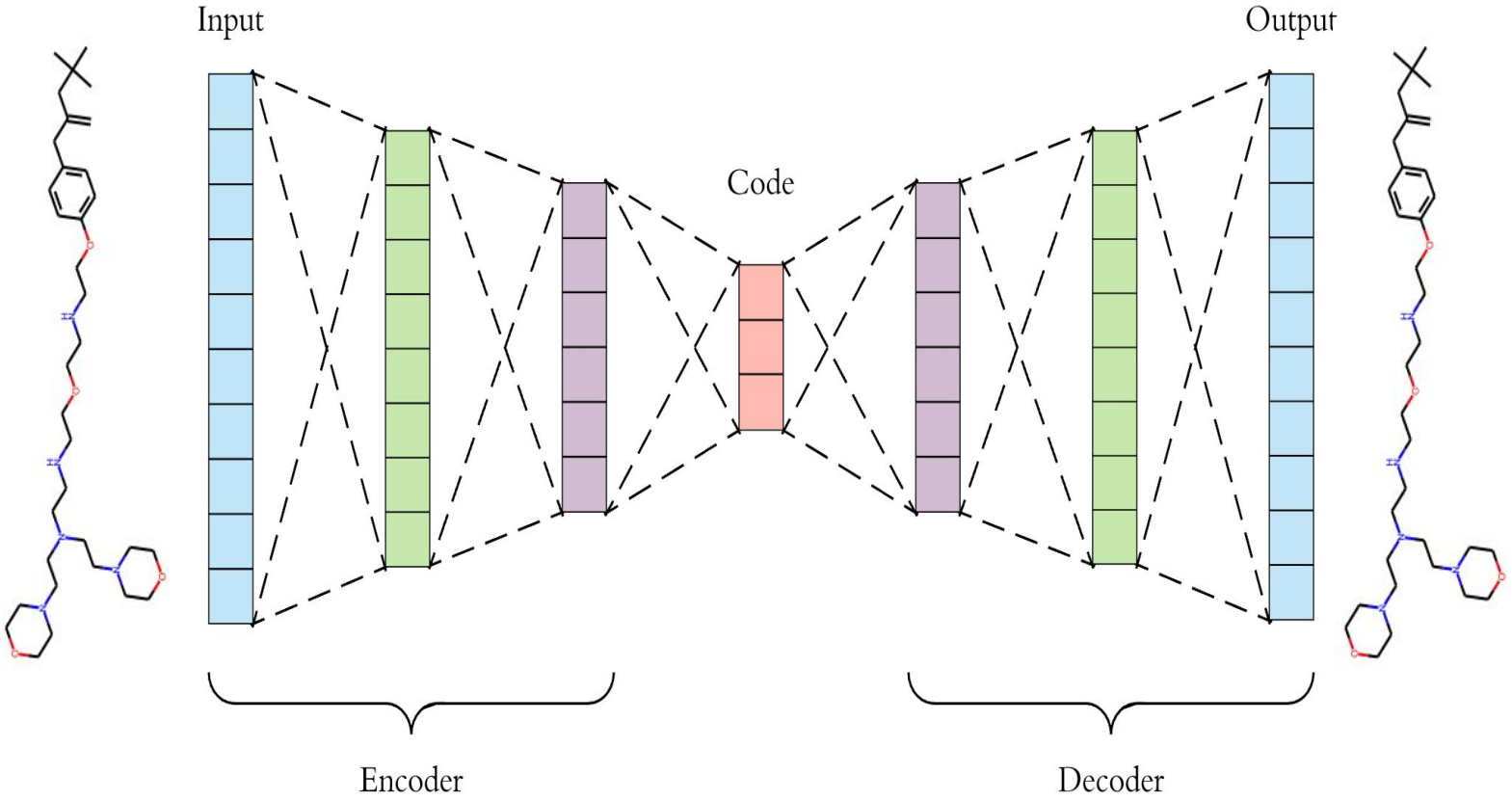
The process of discovering new molecules

- 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)



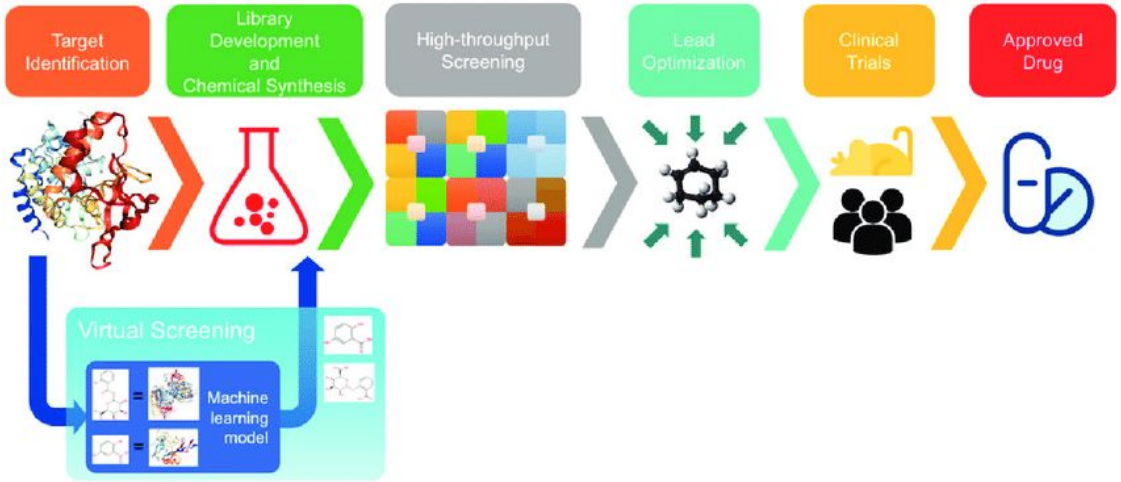


(Variational) Autoencoders



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

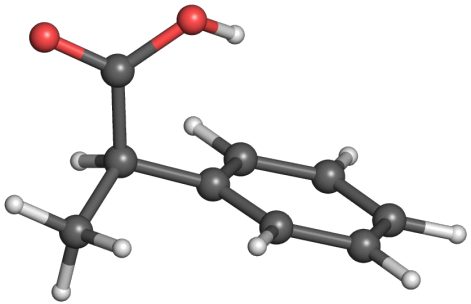
Two main ingredients

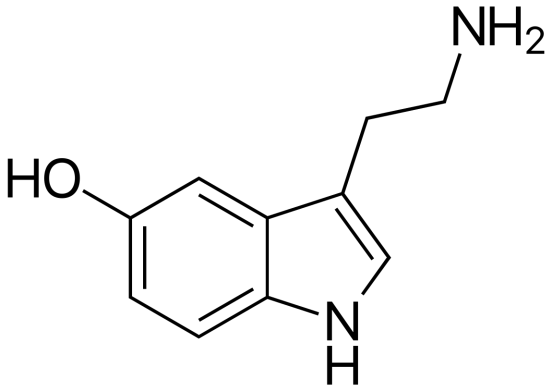
- Molecule representation
- Generative model

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

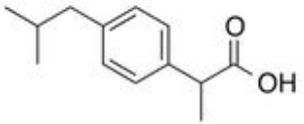




1D representations - SMILEs

Simplified Molecular Input Line Entry System

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

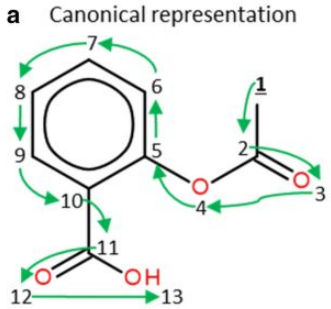


Ibuprofen
`CC(C)Cc1ccc(cc1)C(C)C(O)=O`

- 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



`CC(=O)Oc1ccccc1C(=O)O`

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

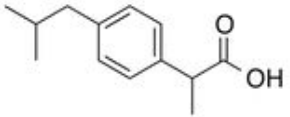
How to generate molecules?

Myriad of different ways. A useful distinction:

- Gradient-free methods
- Gradient-based methods

Recurrent Neural Networks

- Work on sequences (SMILES)
- Goal: given training sequences → learn to generate new sequences that resemble those of training.



Ibuprofen
CC(C)Cc1ccc(cc1)C(C)C(0)=0

- Sequence: $S_{1:T} = (S_1, \ldots, S_T)$ where $S_i \in \mathcal{V}$
- Training: maximum likelihood, equiv to minimize loss function:

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

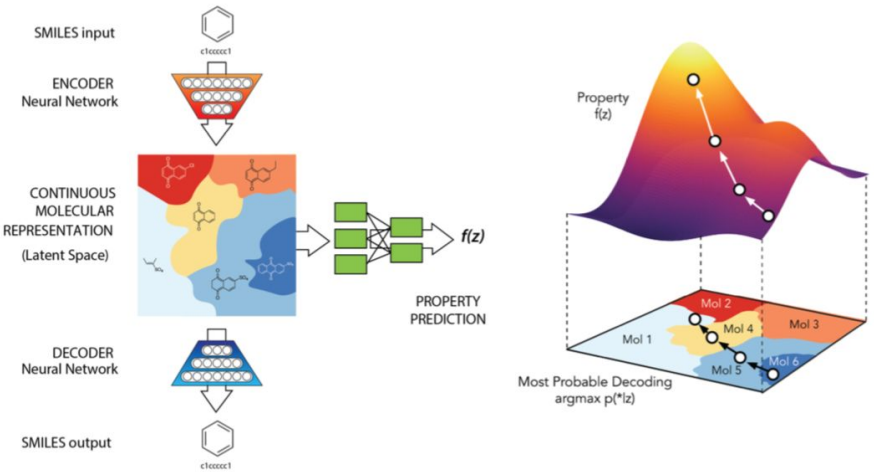
- Generation: sequentially sample from multinomial dist.
- Thermal rescaling

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

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.)



Let’s generate some molecules!

Demo 2

Thanks!

roi.naveiro@icmat.es

https://roinaveiro.github.io/

https://github.com/roinaveiro