DE NOVO DRUG DESIGN – DO WE REALLY WANT TO BE “ORIGINAL”?

A real-world case study on colchicine-site tubulin binders.

MAXIM SHEVELEV, DRAGOS HORVATH (dhorvath@unistra.fr), GILLES MARCOU AND ALEXANDRE VARNEK

Laboratoire de Chemoinformatique, UMR7140 CNR-U. Strasbourg, 4 rue Blaise Pascal, 67000 Strasbourg, France

Tub in Train European Joint Doctorate

Université de Strasbourg

Chémoinformatique STRASBOURG

cnrs

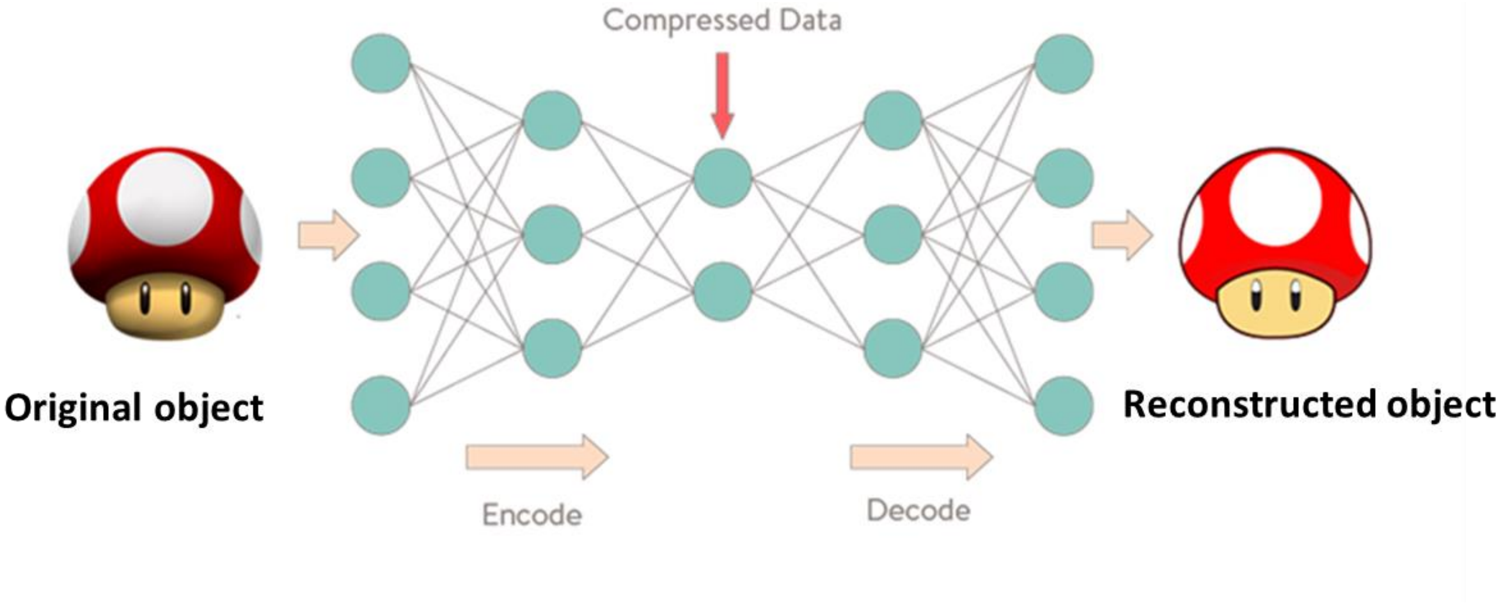
CAN WE DO BETTER THAN “CLASSICAL” VIRTUAL SCREENING, BY USING AI-DRIVEN *DE NOVO* DRUG DESIGN TOOLS?

OUTLINE:

- Generative Models as a Chemical Space Navigator: Inverse QSAR
- Tubulin…
- QSAR Model and Virtual Screening
- “Inverting” Virtual Screening Hits
- Conclusions…

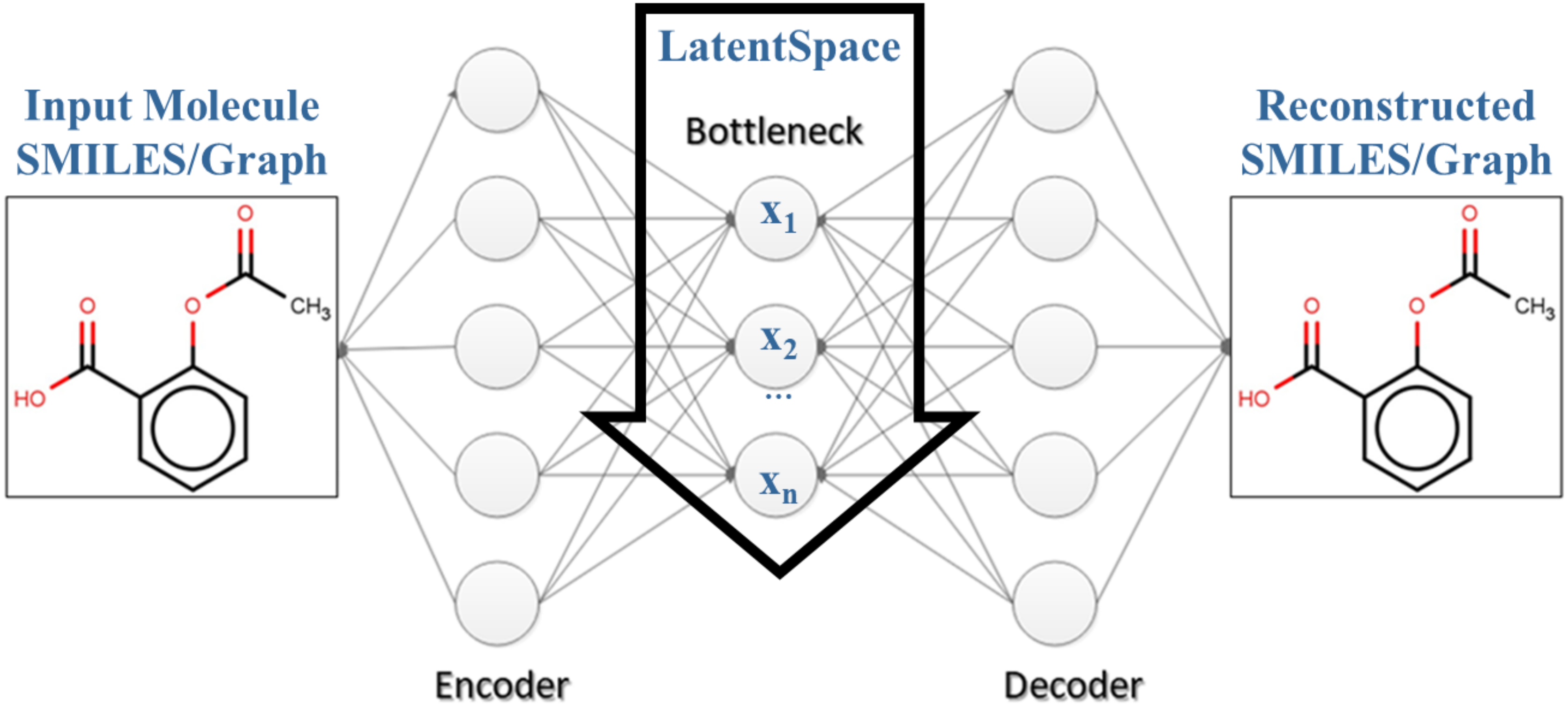
THE AUTOENCODER PARADIGM

- Inspired from Image or Language processing, AutoEncoders are Deep Neural Networks, producing an efficient dense representation of the input by performing specific compression of learned data.
- The “latent” states of Bottleneck Neurons fully characterize the object!

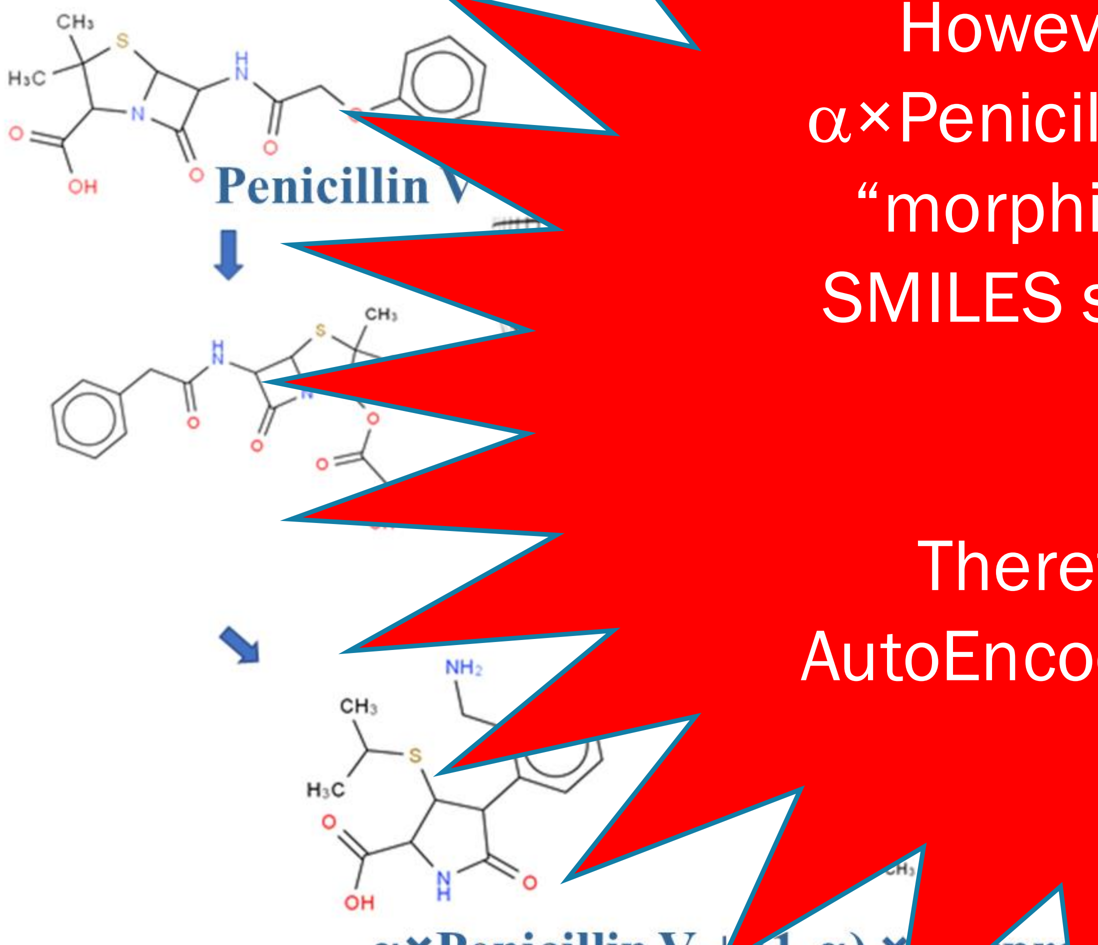


THE AUTOENCODER PARADIGM

- Inspired from Image or Language processing, AutoEncoders are Deep Neural Network producing an efficient dense representation of the input, by performing specific compression of learned data.
- The “**latent**” states of Bottleneck Neurons fully characterize the object!
- It is reversible: provide any latent vector (x_1, x_2, …, x_n) and the Decoder will return a chemical structure associated to those coordinates…



MOVING IN LATENT ECT S IN S RUCTURE SPACE



Penicillin V

α×Penicillin V + (1-α) × upr

albeit it is nothing but el!

pically return ILES

molecules?

theory

mical space, hyped libraries of Bigger compounds.

et actice we have no clue how any structures are covered…

However, performing the α×Penicillin + (1-α)×Ibuprofen "morphing" trick directly on SMILES strings is harrowingly difficult.

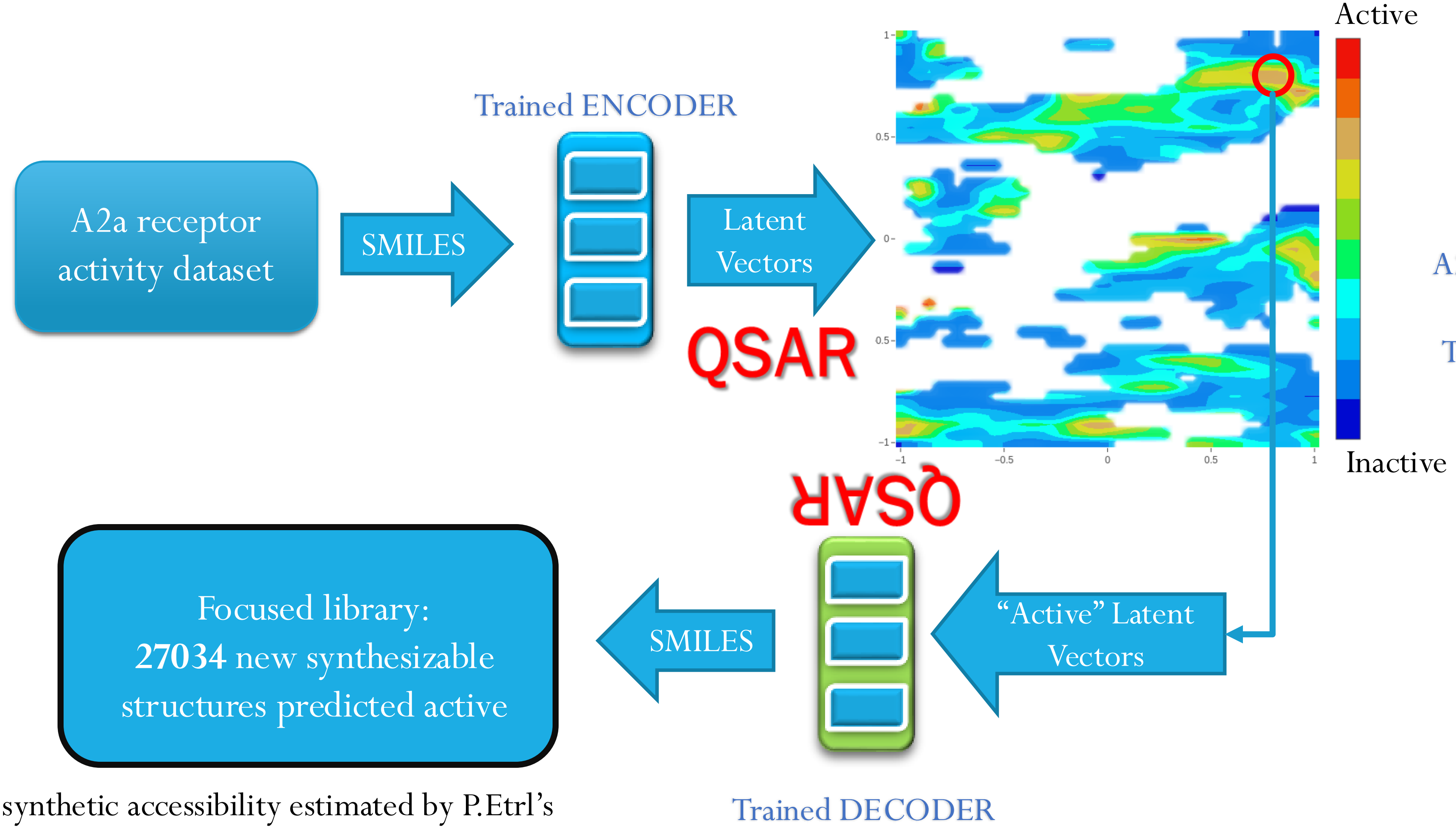
Therefore, like it or not, AutoEncoders are here to stay!

B. Sattarov *et al. J. Chem. Inf. Model.*, 2019, 59(3), 1182-1196

NOW WE KNOW HOW TO MOVE – BUT WHERE SHALL WE GO ?

- To this purpose, we need a QSAR model to annotate the points (x_1, x_2, ..., x_n) of latent space by the predicted property P=QSAR(x_1, x_2, ..., x_n) of therein residing compounds...
- But, wait... are latent vectors eligible molecular descriptors for QSAR models?
  - Beware, they may be atom numbering-dependent!
  - Ok, they contain all the chemical information needed... but it's obscurely encrypted in (x_1, x_2, ..., x_n)
- In spite of this, and surprisingly, latent vectors were shown to support robust QSAR models!

NOW WE KNOW HOW TO MOVE – BUT WHERE SHALL WE GO ?



A2a receptor activity dataset

SMILES

Trained ENCODER

Latent Vectors

QSAR

Active

Inactive

“Active” Latent Vectors

QSAR

Trained DECODER

SMILES

Focused library:
27034 new synthesizable structures predicted active

Latent Space-based QSAR Model:

A2a Activity Landscape by Generative Topographic Mapping

*synthetic accessibility estimated by P.Etrl’s SAscore (2009) model

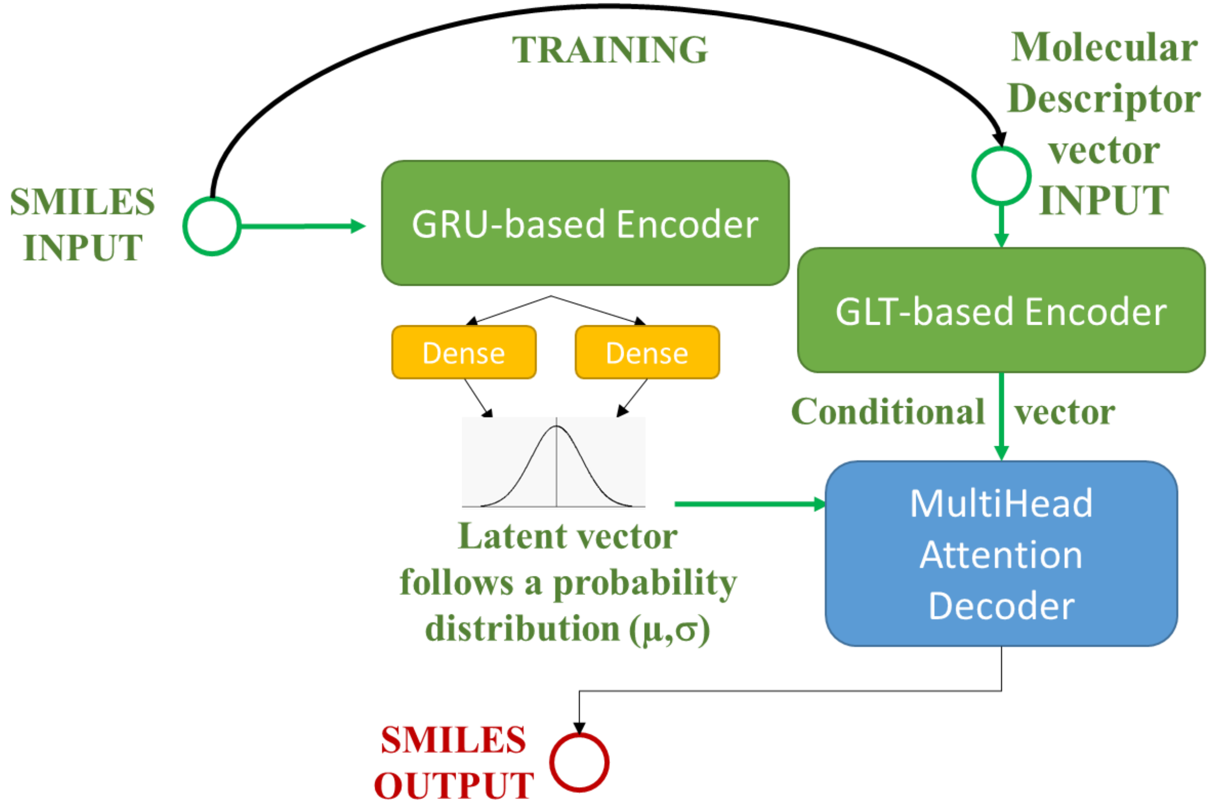
B. Sattarov et al. J. Chem. Inf. Model., 2019, 59(3), 1182-1196

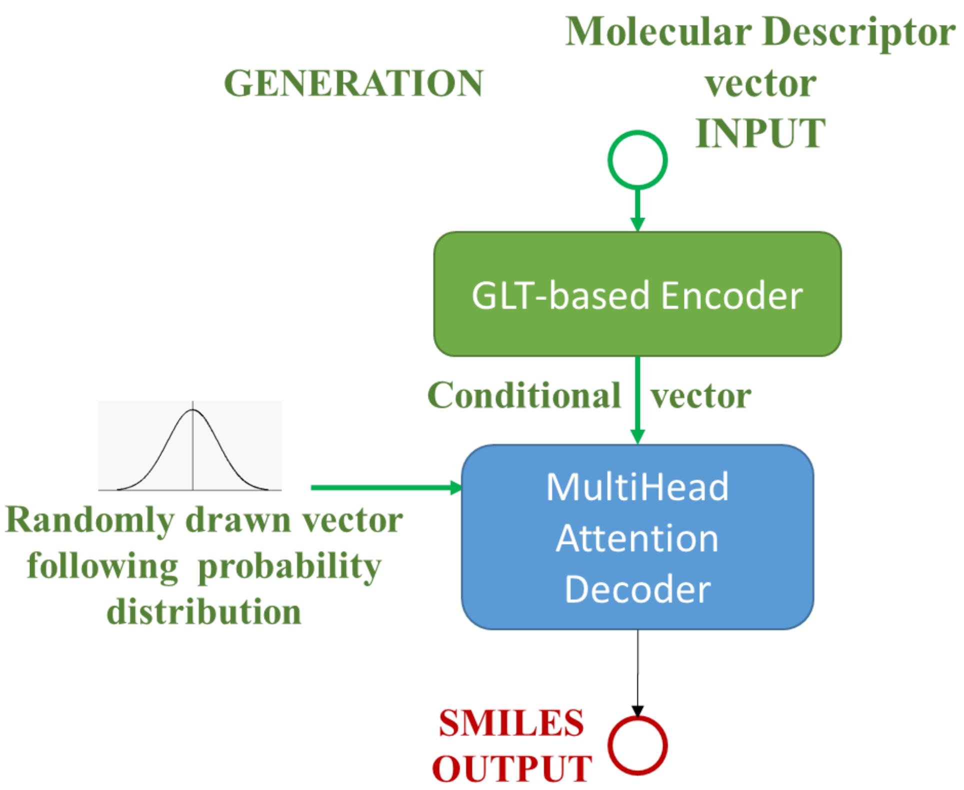
NOW WE KNOW HOW TO MOVE – BUT WHERE SHALL WE GO ?

- To this purpose, we need a QSAR model to annotate the points (x_1, x_2, ..., x_n) of latent space by the predicted property P=QSAR(x_1, x_2, ..., x_n) of therein residing compounds...
- But, wait... are latent vectors eligible molecular descriptors for QSAR models?
  - Beware, they may be atom numbering-dependent!
  - Ok, they contain all the chemical information needed... but it is obscurely encrypted in (x_1, x_2, ..., x_n)
- In spite of this, and surprisingly, latent vectors were shown to support robust QSAR models!
- Inverse Latent Space QSAR is trivial with Encoder Technology...
- **However, classical molecular descriptors cannot be simply dismissed!**

ATTENTION-BASED CONDITIONAL VARIATIONAL AUTOENCODER (ACoVAE)

- This architecture supports working with a descriptor space other than the latent vector space. Descriptor vectors serve as “conditions” to modulate decoder output.


TRAINING
SMILES INPUT
GRU-based Encoder
Molecular Descriptor vector INPUT
GLT-based Encoder
Dense
Dense
Conditional vector
Latent vector follows a probability distribution (μ,σ)
MultiHead Attention Decoder
SMILES OUTPUT


GENERATION
Molecular Descriptor vector INPUT
GLT-based Encoder
Conditional vector
Randomly drawn vector following probability distribution
MultiHead Attention Decoder
SMILES OUTPUT

Bort W, et.al JCIM 2022;62(22):5471-84. doi: 10.1021/acs.jcim.2c01086.

ACoVAE -DRIVEN INVERSE QSAR AS A *DE NOVO* DESIGN TOOL: WORKFLOW

- Fetch a structure-activity data set and build some QSAR model based on information-rich molecular descriptors: $A=QSAR(D_1, D_2, ..., D_n)$
- Train an ACoVAE model based on a significant sample of drug-like compounds (ChEMBL), employing the used QSAR descriptors.
- [A] Find descriptor vector values D* maximizing the predicted activity: $max(A)=QSAR(D^*_1, D^*_2, ..., D^*_n)$.
  - ...or, alternatively, ...

Spoiler

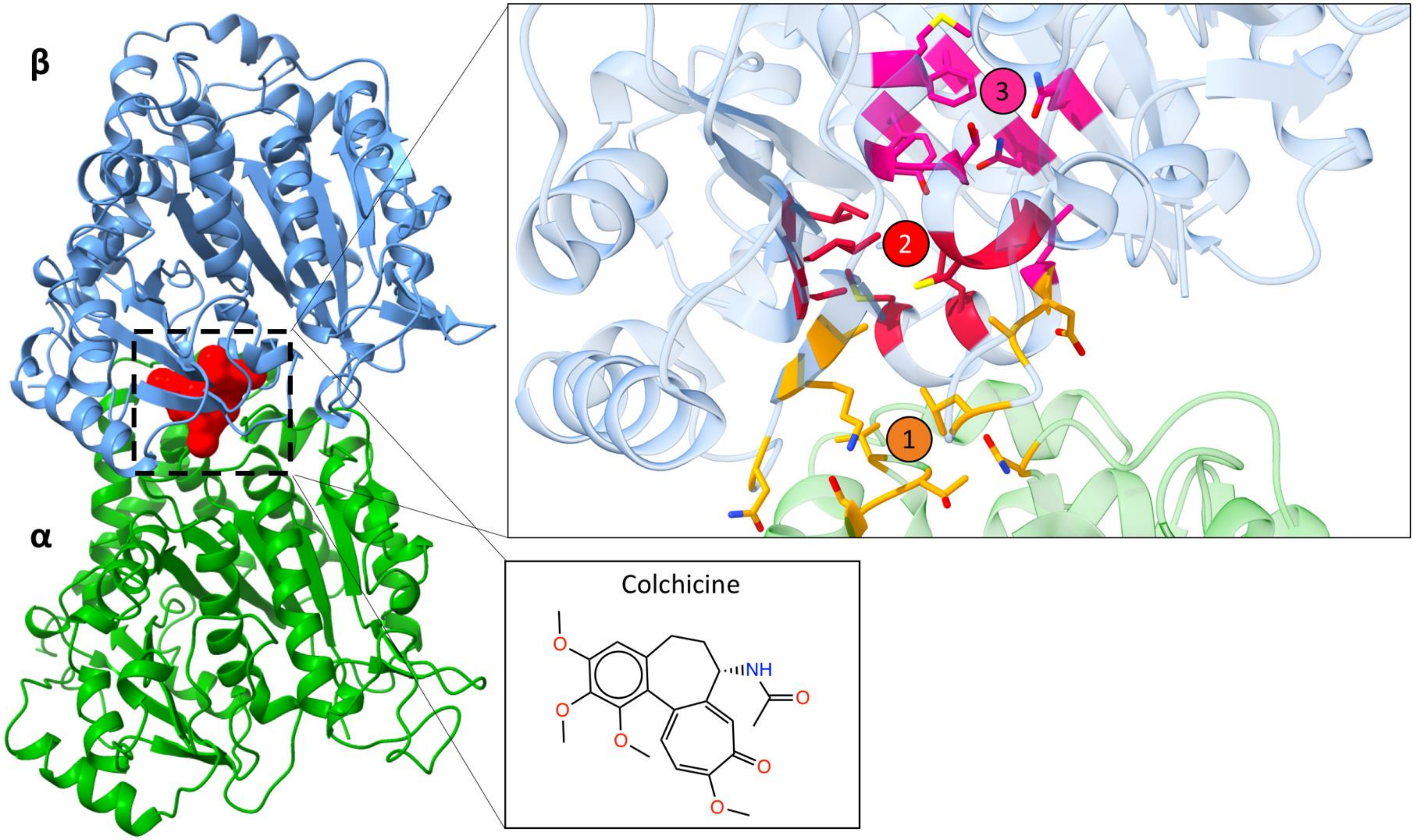
- [B] Use the QSAR model to virtually screen some large compound library, and select the virtual hits predicted to have high A values. Assume optimal descriptors D* to be the actual descriptors of these virtual hit “seeds”.
- Pass the D* values to the ACoVAE decoder, in order to generate (several) novel structures matching given descriptor values
- Evaluate novel structures (predicted A, docking scores, feasibility, drug-likeness, etc...)

OUTLINE

- Generative Models as a Chemical Space Navigator: Inverse QSAR
- Tubulin…
- QSAR Model and Virtual Screening
- “Inverting” Virtual Screening Hits
- Conclusions…

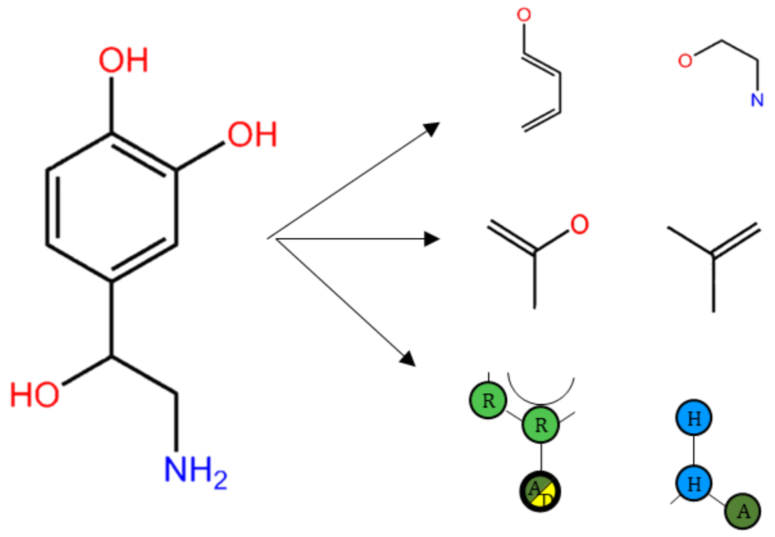
THE CHALLENGE: FINDING ORIGINAL INHIBITORS OF THE COLCHICINE SITE OF TUBULIN

- Ligand binding at the colchicine site inhibits microtubule formation by obstructing the “curved-to-straight” conformational shift in tubulin, herewith exerting a cytotoxic effect targeted at rapidly replicating cancer cells
- Colchicine is a natural product with a very specific scaffold



STRUCTURE-ACTIVITY DATA

- López-López* *et al.* annotated compounds with reported cytotoxicity pIC_{50} values on HeLa cells by their putative action mechanisms. 379 of these are likely acting as colchicine site binders.
- Compound structures were standardized, and 95 different ISIDA fragment count descriptor sets were generated, where fragmentation schemes differed in terms of :
  - topology (sequences, atom-centered fragments)
  - Size (minimal, maximal)
  - atom labels (symbol, pharmacophore, force field type)



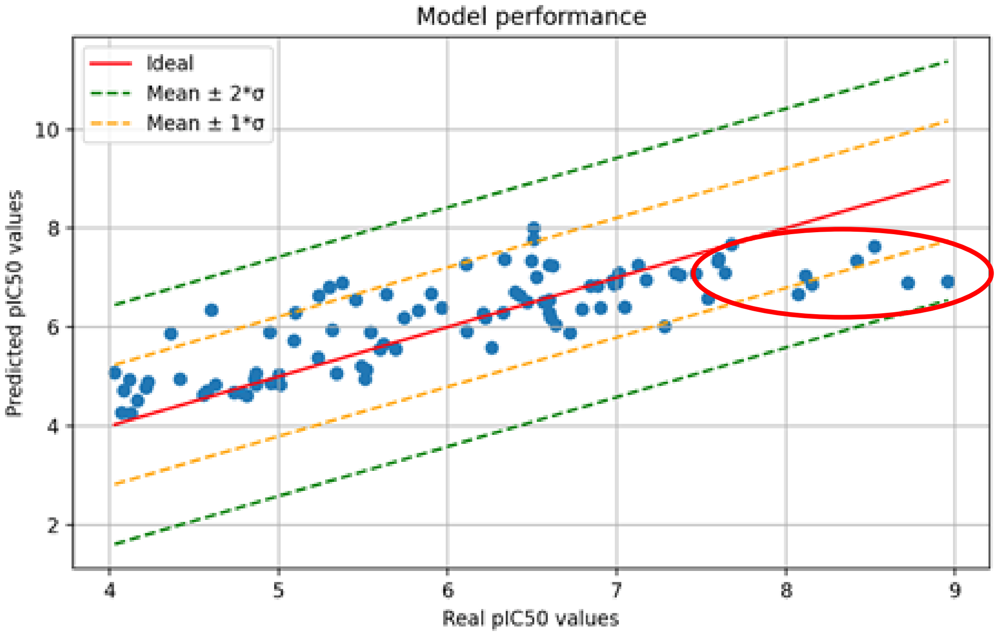
* López-López E, Cerda-García-Rojas CM, Medina-Franco JL. Molecular Informatics. 2023;42(1). doi: 10.1002/minf.202200166

OUTLINE

- Generative Models as a Chemical Space Navigator: Inverse QSAR
- Tubulin…
- QSAR Model and Virtual Screening
- "Inverting" Virtual Screening Hits
- Conclusions…

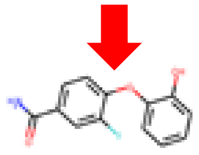
QSAR MODEL

- An evolutionary model-building procedure identified the best descriptor set and optimal model hyperparameters for the Random Forest Regressor model
  - fitness was the mean coefficient of determination Q^2 over a 12-times repeated 3-fold cross-validation scheme
- Best descriptor set: atom pair counts at topological distances ranging from 1 to 5, with atoms rendered by their CVFF force field types (ISIDA notation: IA-FF-P-2-6).
- “Bounding Box”-based Applicability Domain (AD)
- $Q^2 = 0.63$



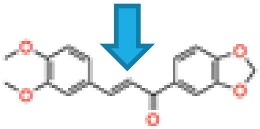
VIRTUAL SCREENING OF THE ENAMINE PHENOTYPIC LIBRARY

- The 5760 compounds underwent standardization, IA-FF-P-2-6 descriptor calculation, AD compliance check and, if compliant, prediction of their HeLa cytotoxicity pIC_{50}.
- The 15 compounds of highest predicted were defined as “seed” molecules for *de novo* generation.



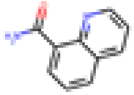
S-1

7.35



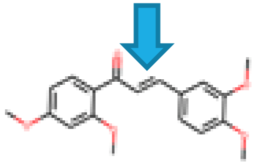
S-2

7.24



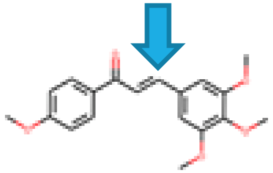
S-3

7.16



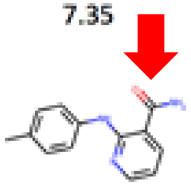
S-4

7.09



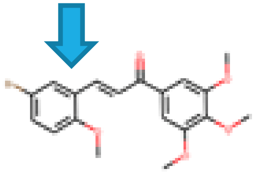
S-5

7.09



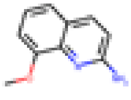
S-6

7.06



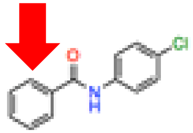
S-7

7.00



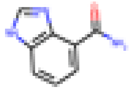
S-8

6.97



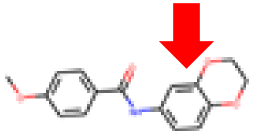
S-9

6.83



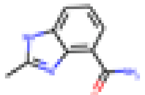
S-10

6.76



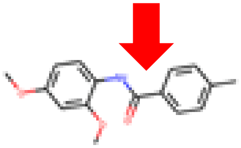
S-11

6.75



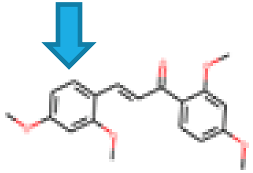
S-12

6.74



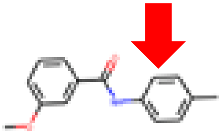
S-13

6.67



S-14

6.65



S-15

6.64

Atom pair counts enable Scaffold Hopping!

OUTLINE

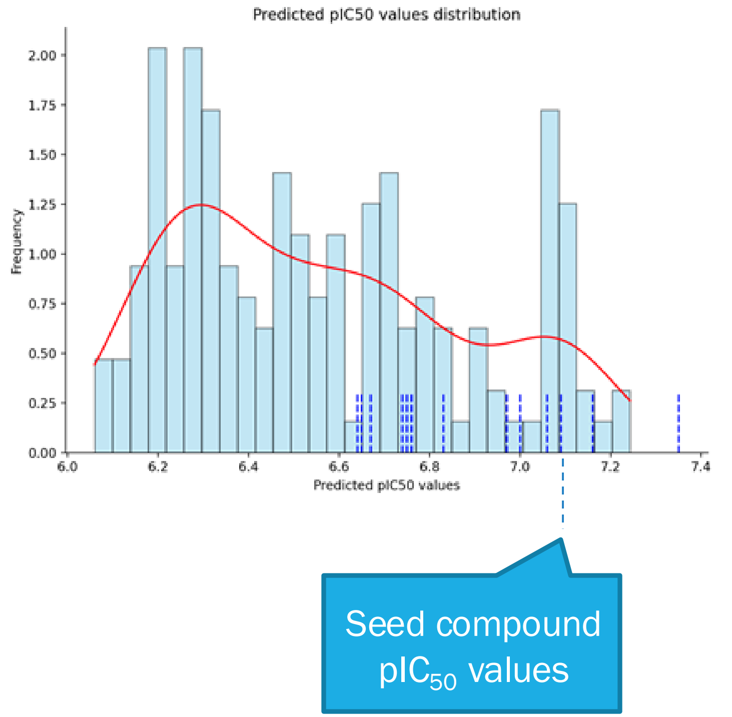
- Generative Models as a Chemical Space Navigator: Inverse QSAR
- Tubulin…
- QSAR Model and Virtual Screening
- “Inverting” Virtual Screening Hits
- Conclusions…

ACoVAE TRAINING

- The ACoVAE model was pre-trained on the ChEMBL database (v. 26, 1,M molecules). The molecules were standardized and IA-FF-P-2-6 ISIDA fragment descriptors were calculated, resulting in a descriptor vector with 2901 fragment features for each molecule.
- The input SMILES strings were limited to a maximum length of 100 characters, and the latent space was a 64-dimensional hypersphere.
- The internal dimension of the trans-former model was set to 256, with 4 layers and 8 heads in the multi-head attention mechanism
- The model was trained with two components of the loss function:
  - reconstruction loss, computed as the sparse categorical cross-entropy between the input and the output
  - Kullback-Leibler divergence between the learned latent distribution and the prior distribution.
- The AdaBelief optimizer was used to fit model parameters. The model was trained for 200 epochs with a batch size of 512..

COMPOUND GENERATION

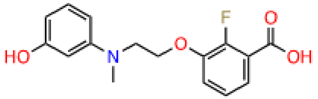
- The 15 descriptor vectors of selected seed compounds were all subjected to the ACoVAE tool, with 500 random latent space samples
  - → 7500 generated SMILES, out of which 6623 were syntactically valid
- Output SMILES strings were standardized and subjected to duplicate removal.
  - → 782 unique structures
- These 782 were subjected to the QSAR-driven virtual screening - ISIDA descriptor calculation, AD compliance check, HeLa pIC_{50} value prediction.
  - → 163 structures inside AD, with predicted pIC_{50} values

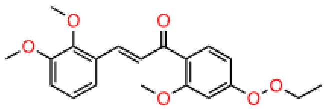


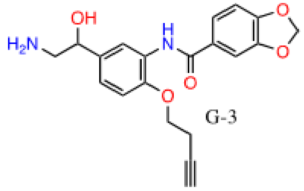
Seed compound pIC_{50} values

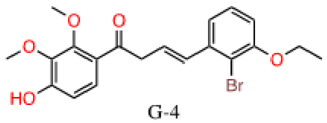
FURTHER VALIDATION OF GENERATED COMPOUNDS

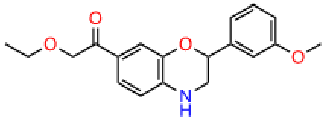
- AutoDock GPU was shown to properly re-dock colchicine in the binding site with an RMSD value of 1.10 Å.
- The 163 generated compounds were docked into the active site prepared from the 402B PDB file, with many achieving docking scores superior to the one of colchicine.
- Top 20 docked molecules were selected for detailed analysis.

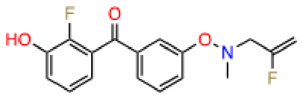

G-1

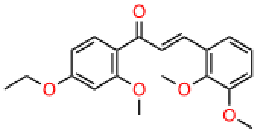

G-2

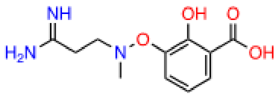

G-3

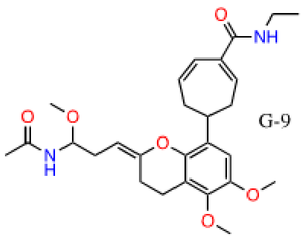

G-4

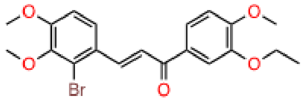

G-5

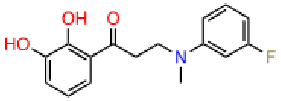

G-6

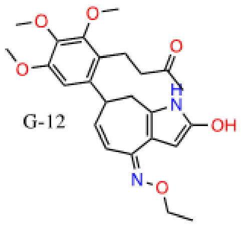

G-7

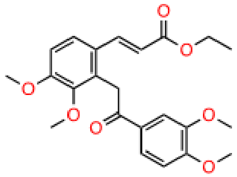

G-8

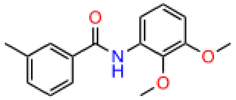

G-9

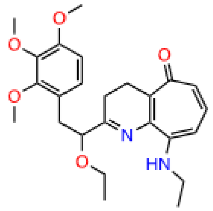

G-10

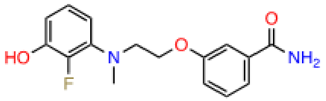

G-11

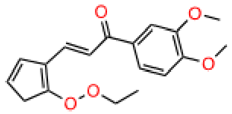

G-12

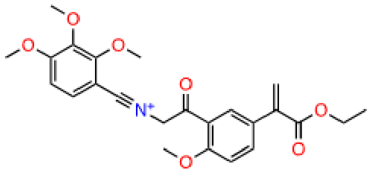

G-13

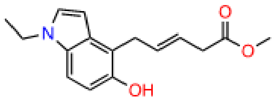

G-14

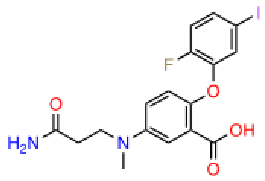

G-15


G-16

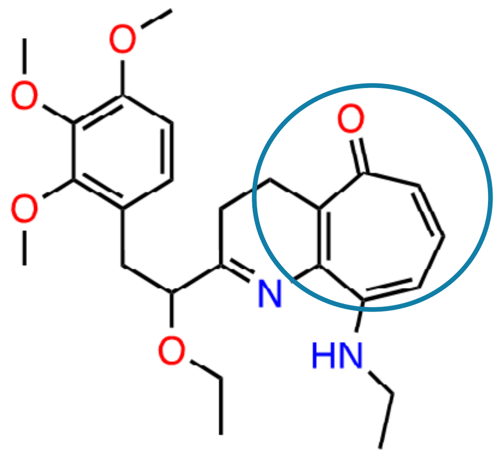

G-17


G-18

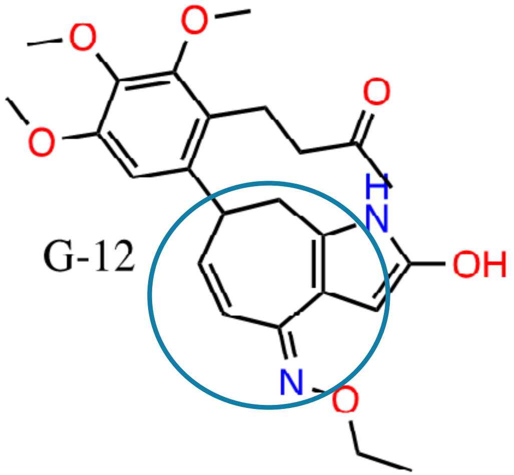

G-19


G-20

INTRIGUING – AI PARTLY "REDISCOVERED" THE COLCHICINE MOTIVE !



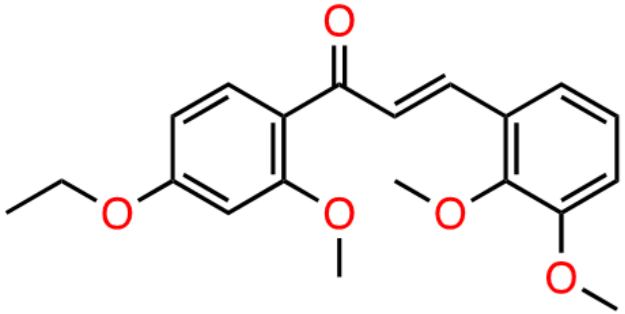
G-15



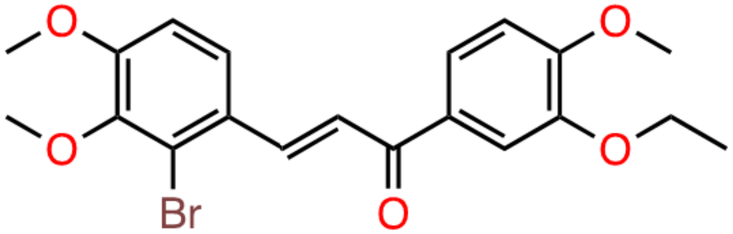
G-12

- … although the tropone ring was absent from all the seed compounds (yet present in ChEMBL – general ACoVAE training)
- It was present in the QSAR training set, but those compounds are not "special" to the ACoVAE. Nonetheless, descriptors of the seed compounds were sufficient to "suggest" generation of (rather uncommon) tropone rings!

TECHNICALLY NOVEL – BUT HARDLY A SCOOP !



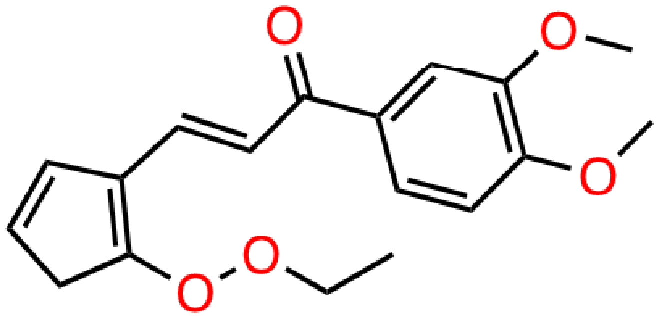
G-7



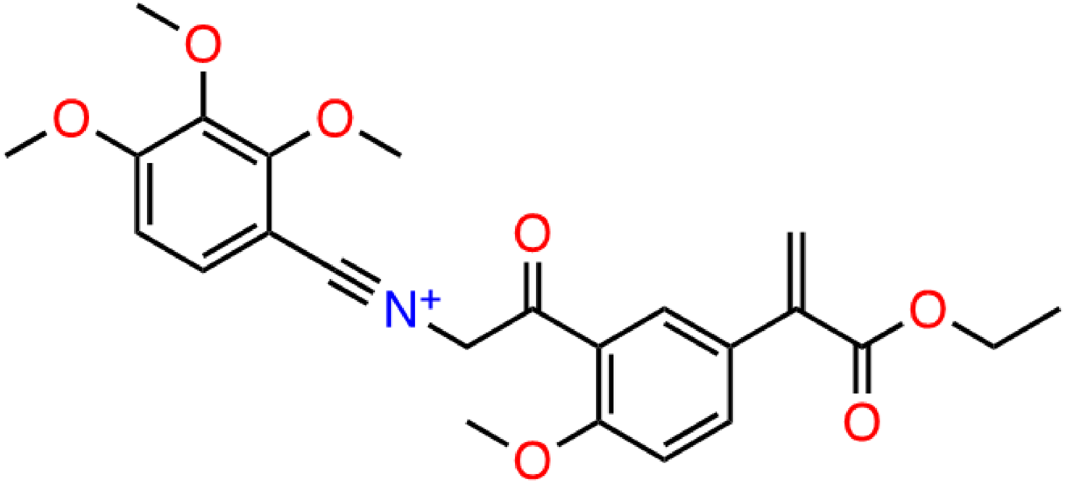
G-10

- *De novo* design may return close analogues of seed compounds, it is not bound to always aim for revolutionary novelty!
- Advantage: at least these compounds are sure to be within the AD!
- Liability: intellectually disappointing… and not automatically easy to make!

A BIT TOO "ORIGINAL"....



G-17



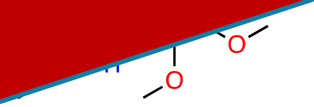
G-18

- Yet, peroxides are not unheard of in drug design (artemisinin)

REALITY CHECK

- A debatable QSAR odel, and a docking ow akly corre with real affinity are *not* sufficient arg to convinc and in synthesizing “original” molecules!

G-14



6.70

Unfortunately, none of the purchased compounds was seen to bind (at the colchicine site, or elsewhere) in a soaking experiment on Tubulin crystals

(Andrea Prota, Paul-Scherrer-Institut, CH)

- However, purchasing their closest analogues within the Enamine Real Space was acceptable!

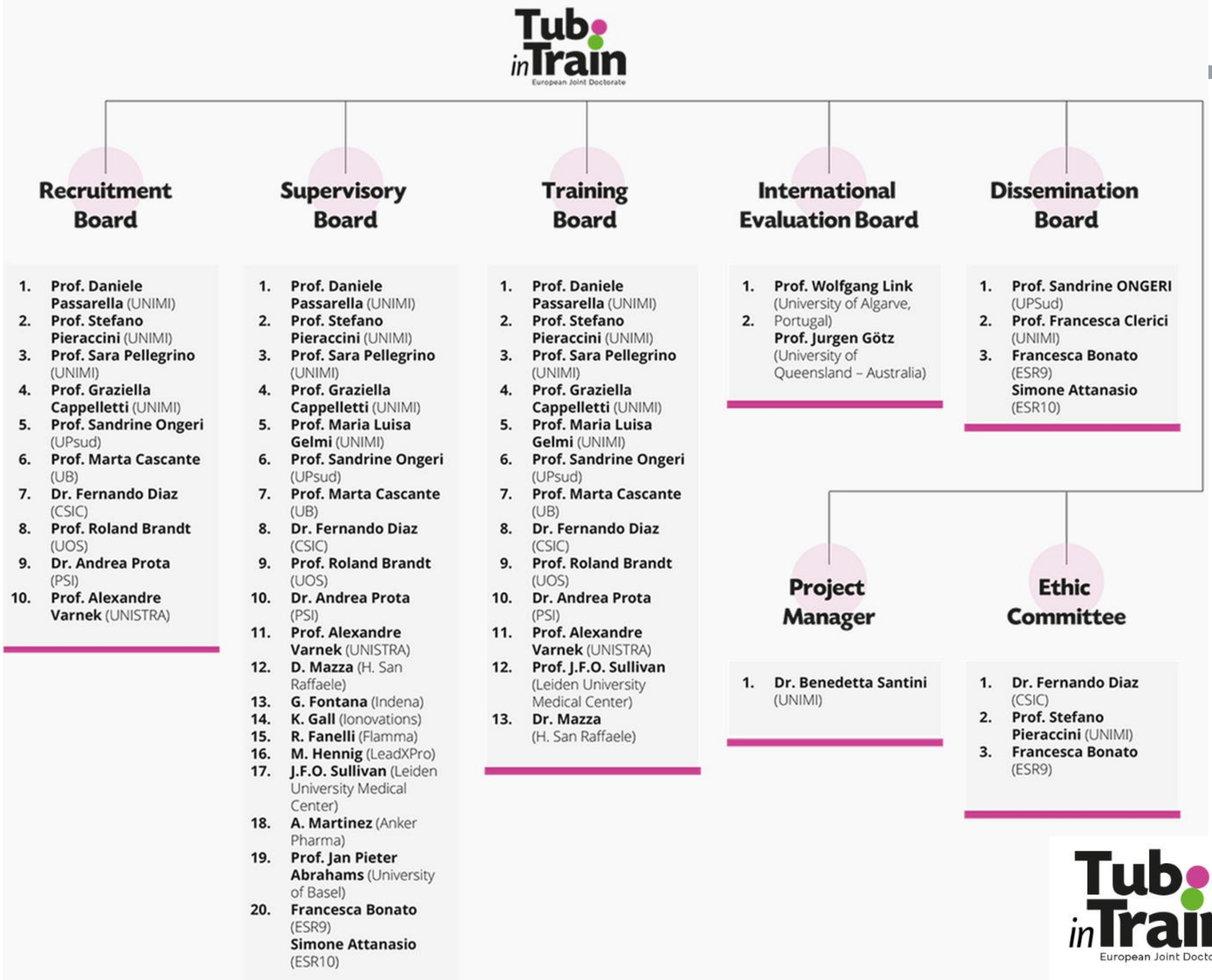
OUTLINE

- Generative Models as a Chemical Space Navigator: Inverse QSAR
- Tubulin…
- QSAR Model and Virtual Screening
- “Inverting” Virtual Screening Hits
- Conclusions…

IN REAL LIFE, NOTHING IS EVER IDEAL…

- Yes, the Encoder/Decoder paradigm is a powerful solution to chemical space navigation and Inverse QSAR!
- However, we all dived into Deep Learning, and left over old unsolved issues – which came to seek revenge! No robust affinity predictions – no trust – no synthesis!
  - Even FEASIBLE molecules require time and money to make – and a good reason to invest in them!
- Originality = Out of Applicability Domain – an insolvable conundrum?
- AI has no monopoly on Originality – Scaffold Hop-supporting descriptors do it as well!
- Medicinal Chemistry is rather conservative – and for good reasons! The AI may occasionally generate “yet another privileged scaffold” with different “ornaments” – but do we really need AI for that?

- **WHAT IS “ORIGINALITY”, ANYWAY ?**



Recruitment Board

1. **Prof. Daniele Passarella** (UNIMI)
2. **Prof. Stefano Pieraccini** (UNIMI)
3. **Prof. Sara Pellegrino** (UNIMI)
4. **Prof. Graziella Cappelletti** (UNIMI)
5. **Prof. Sandrine Ongeri** (UPsud)
6. **Prof. Marta Cascante** (UB)
7. **Dr. Fernando Diaz** (CSIC)
8. **Prof. Roland Brandt** (UOS)
9. **Dr. Andrea Prota** (PSI)
10. **Prof. Alexandre Varnek** (UNISTRA)

Supervisory Board

1. **Prof. Daniele Passarella** (UNIMI)
2. **Prof. Stefano Pieraccini** (UNIMI)
3. **Prof. Sara Pellegrino** (UNIMI)
4. **Prof. Graziella Cappelletti** (UNIMI)
5. **Prof. Maria Luisa Gelmi** (UNIMI)
6. **Prof. Sandrine Ongeri** (UPsud)
7. **Prof. Marta Cascante** (UB)
8. **Dr. Fernando Diaz** (CSIC)
9. **Prof. Roland Brandt** (UOS)
10. **Dr. Andrea Prota** (PSI)
11. **Prof. Alexandre Varnek** (UNISTRA)
12. **D. Mazza** (H. San Raffaele)
13. **G. Fontana** (Indena)
14. **K. Gall** (Ionovations)
15. **R. Fanelli** (Flamma)
16. **M. Hennig** (LeadXPro)
17. **J.F.O. Sullivan** (Leiden University Medical Center)
18. **A. Martinez** (Anker Pharma)
19. **Prof. Jan Pieter Abrahams** (University of Basel)
20. **Francesca Bonato** (ESR9) **Simone Attanasio** (ESR10)

Training Board

1. **Prof. Daniele Passarella** (UNIMI)
2. **Prof. Stefano Pieraccini** (UNIMI)
3. **Prof. Sara Pellegrino** (UNIMI)
4. **Prof. Graziella Cappelletti** (UNIMI)
5. **Prof. Maria Luisa Gelmi** (UNIMI)
6. **Prof. Sandrine Ongeri** (UPsud)
7. **Prof. Marta Cascante** (UB)
8. **Dr. Fernando Diaz** (CSIC)
9. **Prof. Roland Brandt** (UOS)
10. **Dr. Andrea Prota** (PSI)
11. **Prof. Alexandre Varnek** (UNISTRA)
12. **Prof. J.F.O. Sullivan** (Leiden University Medical Center)
13. **Dr. Mazza** (H. San Raffaele)

International Evaluation Board

1. **Prof. Wolfgang Link** (University of Algarve, Portugal)
2. **Prof. Jurgen Götz** (University of Queensland – Australia)

Dissemination Board

1. **Prof. Sandrine ONGERI** (UPSud)
2. **Prof. Francesca Clerici** (UNIMI)
3. **Francesca Bonato** (ESR9) **Simone Attanasio** (ESR10)

Project Manager

1. **Dr. Benedetta Santini** (UNIMI)

Ethic Committee

1. **Dr. Fernando Diaz** (CSIC)
2. **Prof. Stefano Pieraccini** (UNIMI)
3. **Francesca Bonato** (ESR9)

- This work was funded and supported by the TubInTrain European Doctoral Consortium
- https://www.tubintrain.eu/