Machine-Learning for chemogenomics: Application to target identification for Cystic Fibrosis

Or : How to identify the targets of drugs identified in phenotypic screens ?

Pr Isabelle Sermet (Institut Necker Enfants Malades, INEM)

Pr Véronique Stoven (CBIO, Mines Paris PSL)

Laurence Calzone, Loredana Martignetti (U1331, Institut Curie)

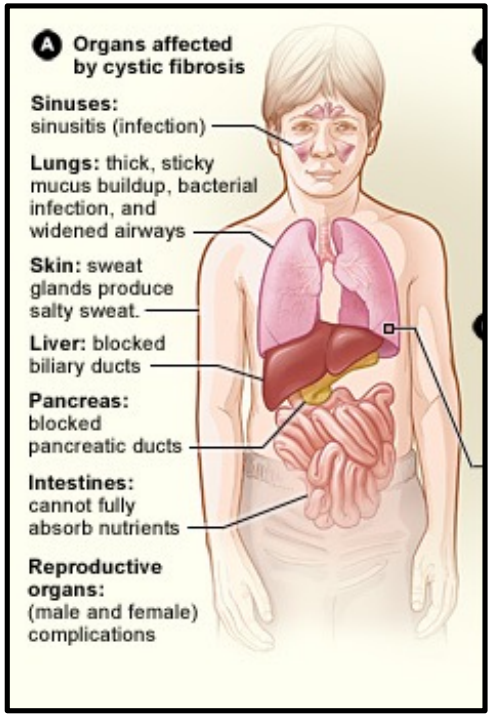
MINES PARIS | PSL

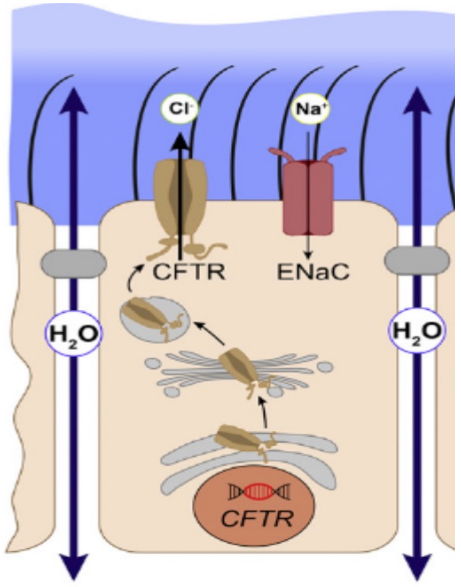
institut Curie

Inserm
La science pour la santé
From science to health

Cystic Fibrosis (CF)

- Most frequent genetic disease in caucasian population (1/3500 birth)
- **Mutation in *cftr* gene** (**C**ystic **F**ibrosis **T**ransmembrane **R**egulator) coding for **CFTR protein**: **Cl^- channel** at the apical surface of epithelial cells
- **> 2000 mutations,** various **impacts** on CFTR protein (classes)
- **70% cases: deletion of F508**, folding defect in CFTR, elimination via proteasome, absence at membrane (Class II)
- **Pulmonary symptoms**: chronic infections, inflammation, tissue damage, altered pulmonary function





CFTR in healthy airways

Until recently (10 years): symptomatic treatments (antibiotics, physiotherapy, nutritional support etc…)

CFTR modulators : proteic therapy

CFTR Activators and **Correctors** : improve processing, maturation and channel activity of mutated CFTR

Limitations :

- Patients bearing **at least one F508del mutation**
- Some patients are **non responders**
- **15% patients** bear "un-rescuable" mutations → **not eligible**.
- **Mechanism of action not fully understood**: identified in phenotypic screens (increase of chloride conductance).

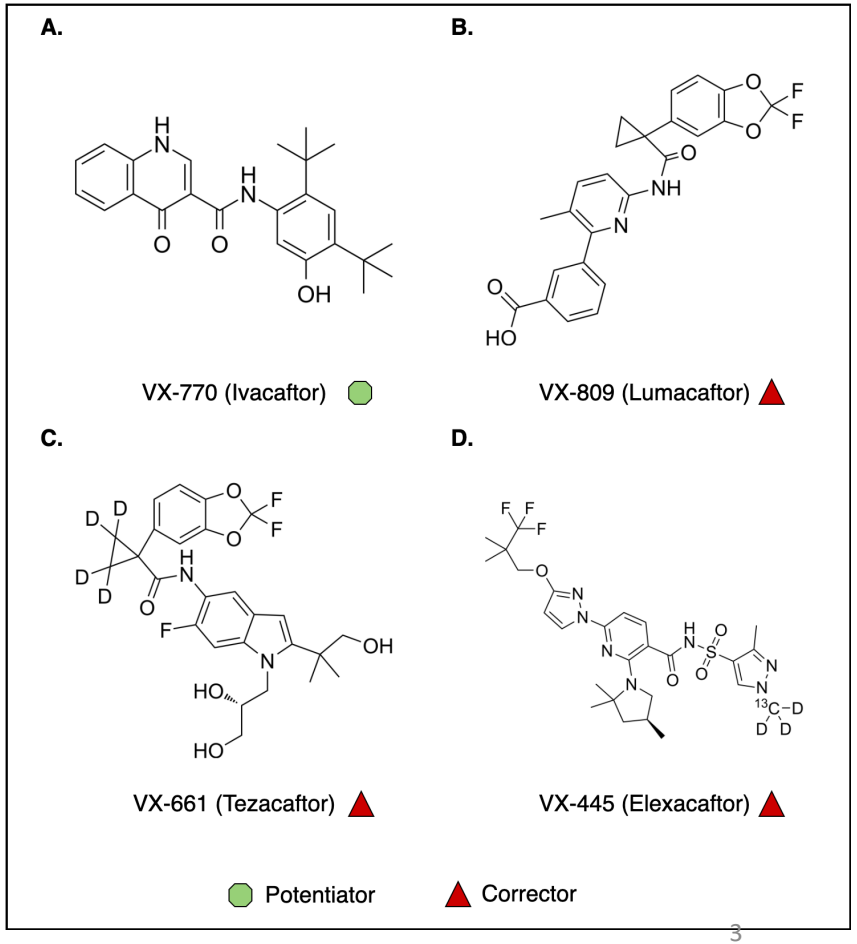
Issues:

Identify ETI protein targets other than CFTR

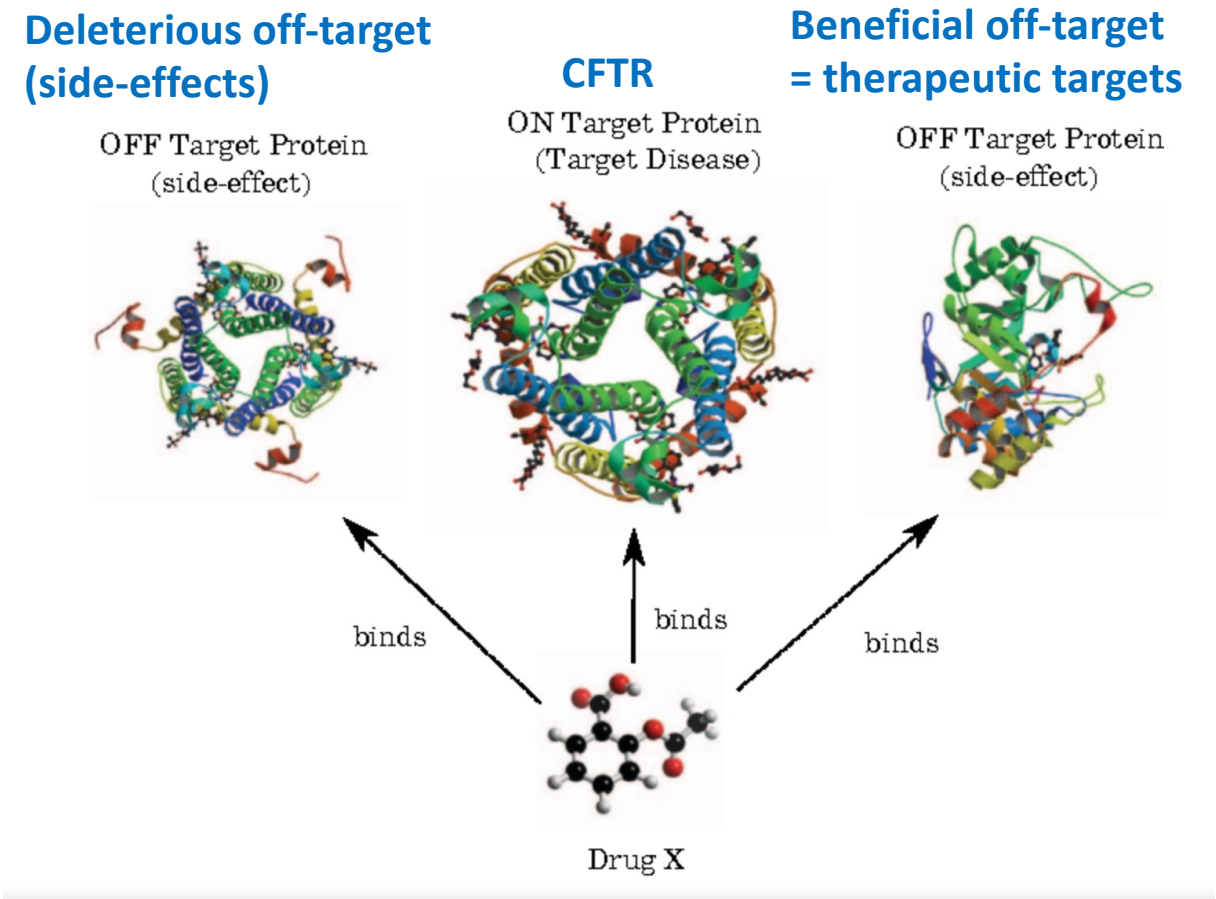
(clinical response not correlated with restoration of Cl^- conductance)

Propose new therapeutic targets and their associated "mutation agnostic" drugs (ETI ot optimized for them).

Trikafta™ (**ETI**, 4 years)= 2 correctors (Elexacaftor, Tezacaftor) + 1 activator (Ivacaftor)



Identify new therapeutic targets among ETI binding proteins



Difficult problem : 25.000 human genes, 3000 known druggable proteins

→ **Predict ETI protein interaction profile**

Boils down to a chemogenomic problem:

Predict drug-target interactions (DTI) at large scale

Of general interest in drug discovery

Phenotypic drugs candidates in cancer (survival assays)

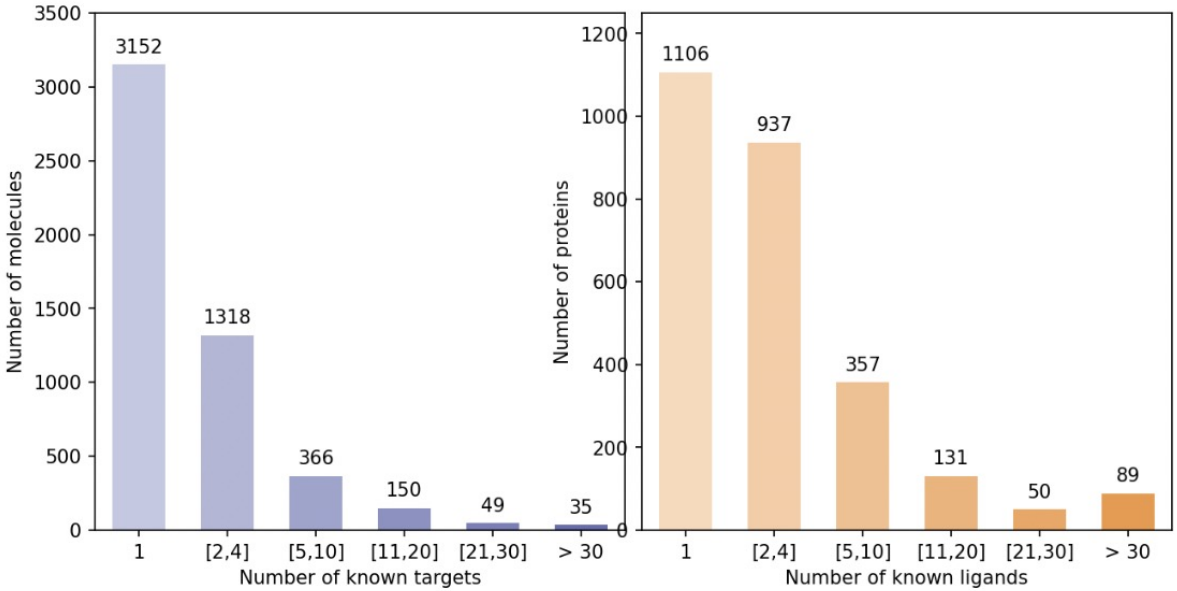
Anticipate side-effects, drug repurposing etc…

Using a **ML algorithm** and a **training set of DTIs**

Quality of training set (known interactions): critical for DTI prediction with ML

Limitations in classical DTI training sets:

- bias for well-studied families of proteins (kinases, GPCRs, NRs etc…). Idem for mol
- no negative interactions. **Randomly chosen among unknown interactions**



DrugBank : 8.600 interactions out of 12.210 involve proteins with > 10 known ligands

Protein nb of Ligands	nb of Interactions
1	1106
2 to 4	2527
5 to 10	2404
11 to 20	1920
21 to 30	1238
>30	5442

Nb of known ligands **FPR**

Prot in Category	RN-Datasets
0	2.2 ± 0.4
1	3.7 ± 0.5
2 to 4	5.1 ± 0.9
5 to 10	9.9 ± 0.9
11 to 20	13.8 ± 1.7
21 to 30	23.0 ± 4.9
> 30	18.6 ± 2.8

Proteins with many ligands viewed by ML algorithm as probable targets:
leads to False Positive predictions

Correct bias by “balanced” positive and negative DTIs per protein and molecule

Nb of known ligands

Prot in Category	FPR (Threshold = 0.5) RN-Datasets	BN-Datasets
0	2.2 ± 0.4	3.1 ± 0.5
1	3.7 ± 0.5	3.1 ± 0.8
2 to 4	5.1 ± 0.9	6.4 ± 0.8
5 to 10	9.9 ± 0.9	8.3 ± 0.6
11 to 20	13.8 ± 1.7	10.6 ± 0.5
21 to 30	23.0 ± 4.9	12.0 ± 3.0
> 30	18.6 ± 2.8	9.0 ± 0.4

→ Performance of ML algorithm according to False Positive Rate (FDR) on DrugBank

randomly choose negatives such that **equal nb of positives and negatives for each protein and molecule**

Target prediction improvement:
Improves score and rank of true target for 3 marketed drugs

	RN-Datasets		BN-Datasets	
Drug	Target Score	Target Rank	Target Score	Target Rank
DB11363	0.8	31	0.8	3
DB11842	0.76	31	0.85	18
DB11732	0.67	107	0.83	17

Najm et al, IJMS 2021

Build a large high-quality DTI training dataset : LCIdb

Binary interaction databases

DrugBank: DTI between FDA approved or in development molecules and their targets

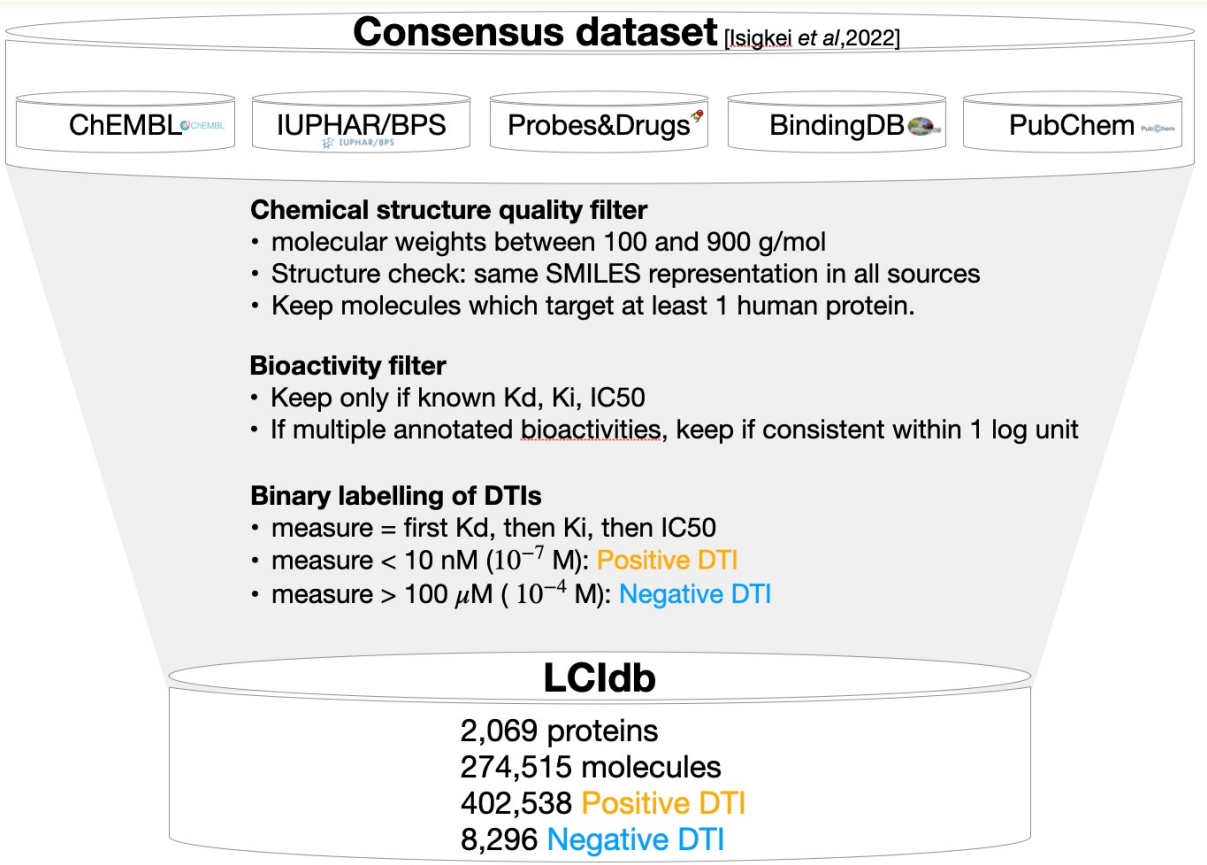
Drugbank v1.5.1 [Wishart *et al*, 2018]

2.513 proteins
4.813 molecules
13.716 + interactions

BIOSNAP similar size

BindingDB smaller

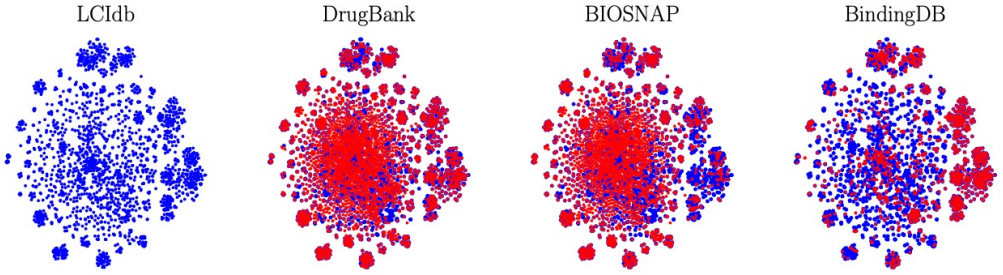
Bioactivity databases



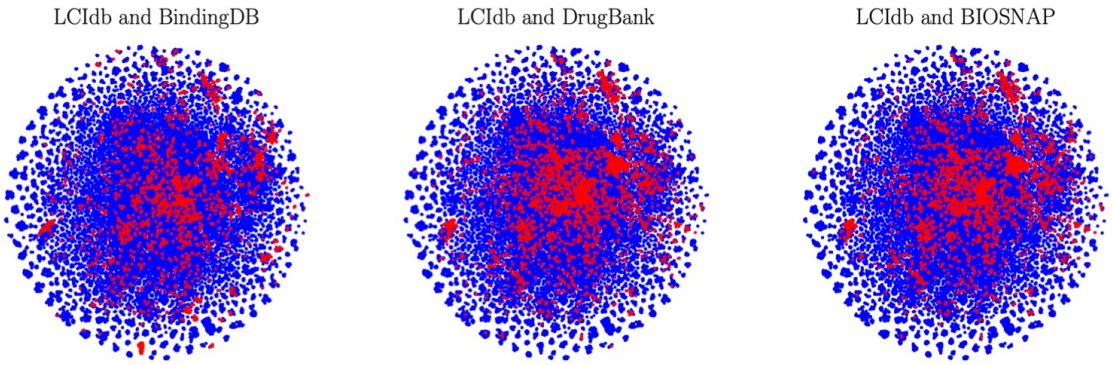
Complete with random “balanced” negatives per protein and molecule

Characterisation of LCIdb

2D representation of protein space coverage
(t-SNE algorithm on protein features)



2D representation of chemical space coverage
(t-SNE algorithm on molecule features)



Datasets	Molecules	Proteins	Positive DTIs
BIOSNAP	4,510	2,181	13,836
Unseen_drugs			13,836
Unseen_targets			13,836
BindingDB	7,161	1,254	9,166
DrugBank	4,813	2,507	13,715
DrugBank (Ext)	4,257	1,216	10,838
LCIdb	271,180	2,060	396,798
Unseen_drugs	271,180	2,060	396,798
Unseen_targets	271,180	2,060	396,798
Orphan	191,901	2,060	208,041

LCIdb

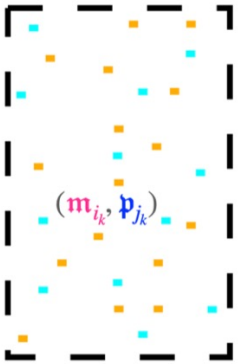
30 times larger than BIOSNAP and DrugBank
43 times larger than BindingDB

Similar nb of proteins (“druggable” human proteins)
Much larger nb of molecules

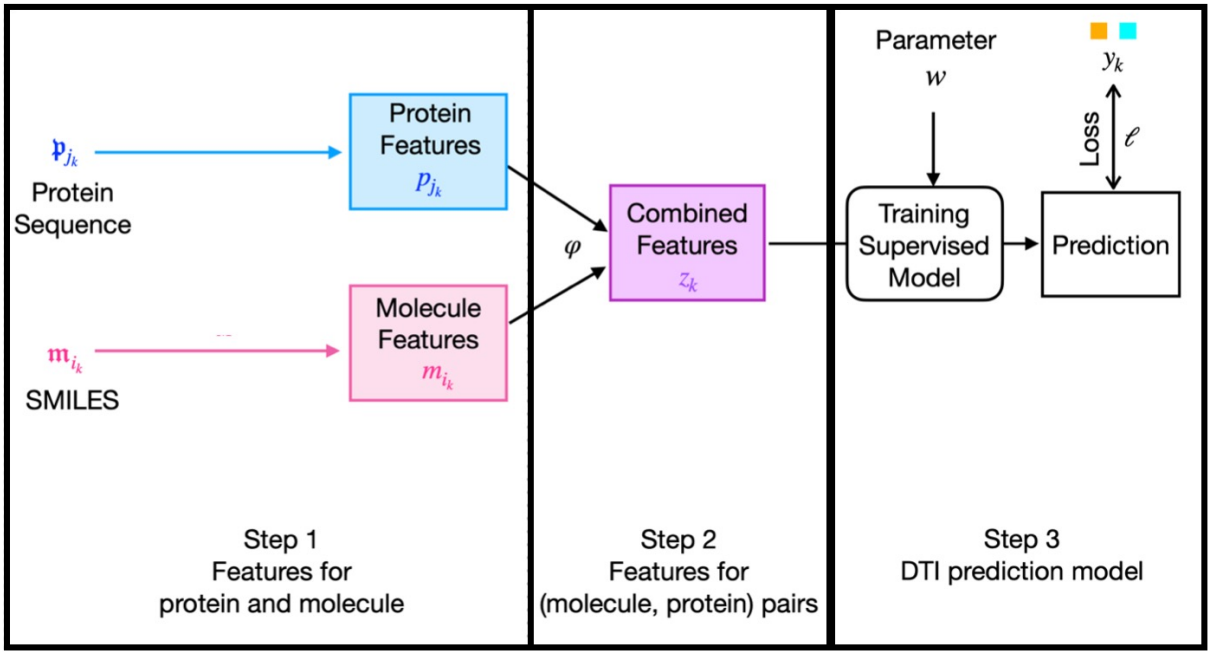
Guichaoua et al, JCIM 2024

General pipeline for DTI prediction in ML chemogenomic framework

A classification problem: distinguish **interacting** and **not interacting** (*m, p*) pairs



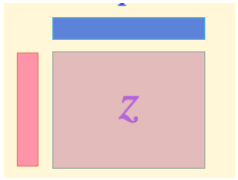
Training dataset : incomplete binary (+1 or 1) interaction matrix



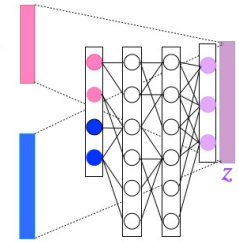
Combine features:

Concatenation:

Tensor product:



Neural Network:



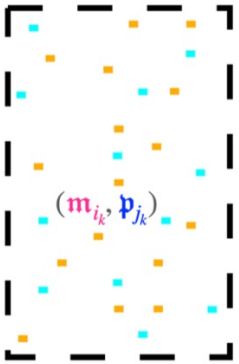
Computational burden for large training sets (LCIdb)

KOMET : efficient ML algorithm

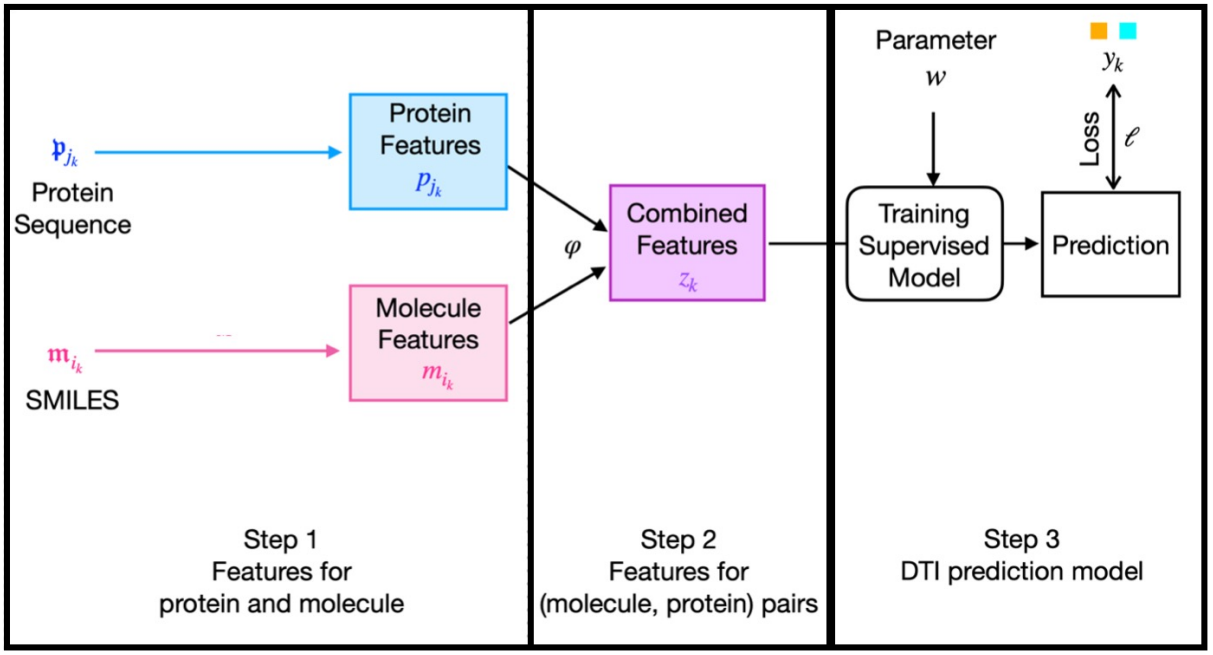
- Step 1: molecule and protein **features derived from Kernel methods**
- Steps 2-3: **tensor product** with "classical" **SVM** algorithm in the space of (m, p) pairs

General pipeline for DTI prediction in ML chemogenomic framework

A classification problem: distinguish **interacting** and **not interacting** (*m, p*) pairs



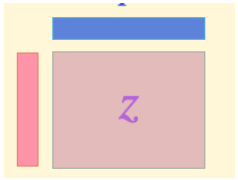
Training dataset : incomplete binary (+1 or 1) interaction matrix



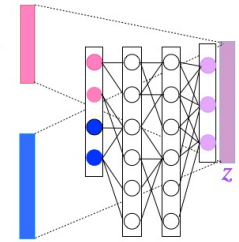
Combine features:

Concatenation:

Tensor product:



Neural Network:



Computational burden for large training sets (LCIdb)

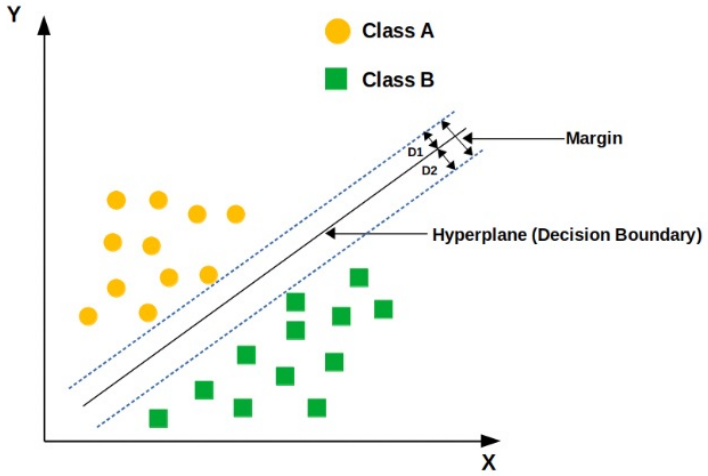
KOMET : efficient ML algorithm

- Step 1: molecule and protein **features derived from Kernel methods**
- Steps 2-3: **tensor product** with “classical" **SVM** algorithm in the space of (m, p) pairs

Train a SVM = Learn hyperplane separating positive/negative data points

Training set: $(\boldsymbol{x_i}, y_i)$ with $y_i = \mp 1$ (labels)

Linearly separable case

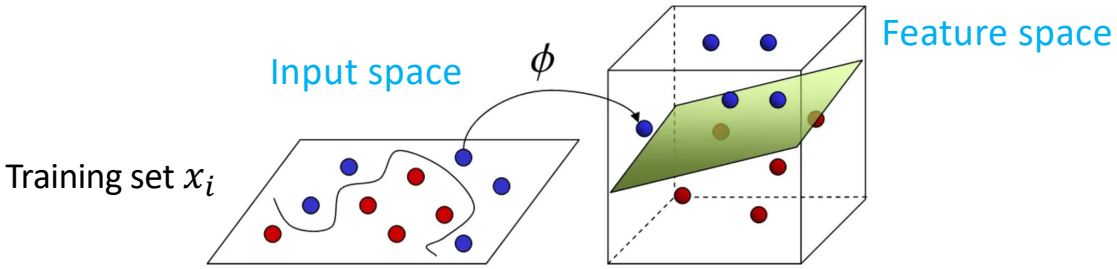


compute hyperplane with largest margin :
only involves $< \boldsymbol{x_i}.\boldsymbol{x_j} >$

⟶ Good generalization properties

Prediction based on position of points

Linearly non-separable case: may be linearly separable after transformation ϕ in a feature space (of higher dimension)



Kernel trick provides a solution to this problem:

- **define $\boldsymbol{k}$** : a similarity measure (kernel) in the input space
- **corresponds to a mapping function $\boldsymbol{\phi}$** in a feature space such that:

$< \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) > = k(\boldsymbol{x_i}, \boldsymbol{x_j})$ ⟶ kernel representation of data

- compute hyperplane only based on $k(\boldsymbol{x_i}, \boldsymbol{x_j})$ (**$\boldsymbol{\phi}$ not computed explicitly**)

Predictions based on $k(\boldsymbol{x}, \boldsymbol{x_i})$

The “classical” approach for SVM with kernel methods for chemogenomics

Training set : (m, p) **pairs** (with labels +1 or -1)

- **choose encodings** for molecules and proteins $\phi_M(m)$ and $\phi_P(p)$ (i.e. descriptors)

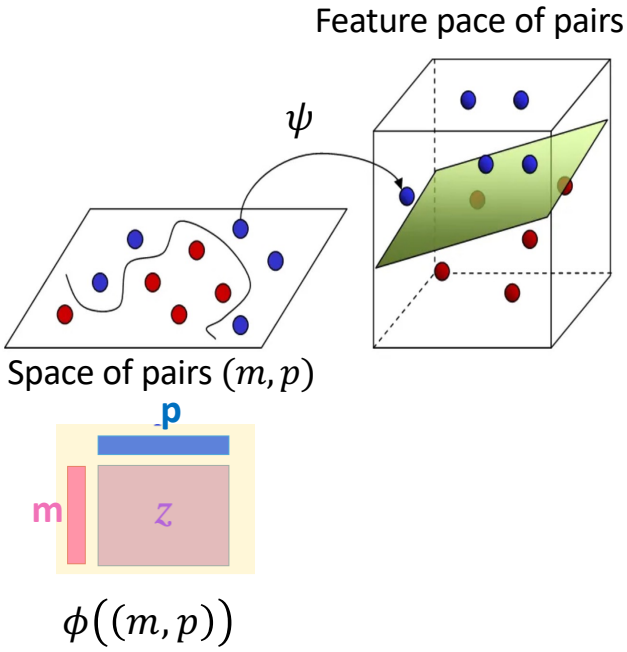
- **define similarity measures (kernels)** on molecules and proteins : k_M and k_P

- **choose tensor product to encode** $(\boldsymbol{m}, \boldsymbol{p})$ **pairs**: $\phi\big((m,p)\big) = \phi_M(m) \otimes \phi_P(p)$

- **hyperplane in the feature space of pairs computed based only on** $\boldsymbol{k_M}$ **and** $\boldsymbol{k_P}$
(based on $k_M(m_i, m_j).k_P(p_i, p_j)$, $\phi\big((m,p)\big)$ not computed)

Motivation to use Tensor product with SVM kernel methods :
- **Systematic definition and good “mixing properties”**
- **Efficient computation**

Predictions only based on $k_M(m, m_i).k_P(p, p_i)$



Limitation: computational burden for training on large DTIs datasets (LCIdb)

Principle of the KOMET algorithm: conceived for large training sets

KOMET solves a "classical SVM optimization problem" in the feature space of (m, p) pairs based on:

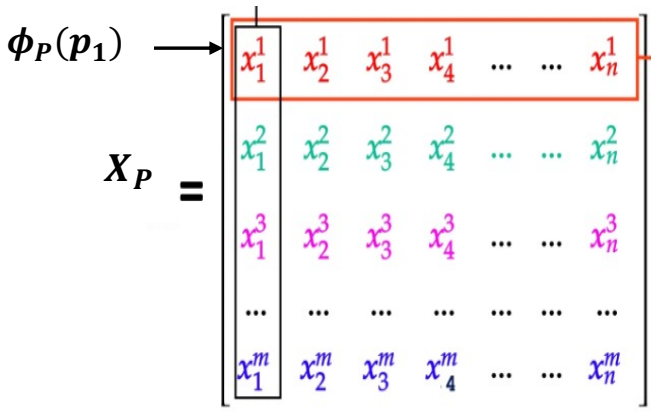
- **"kernel-derived" encoding of molecules and proteins ϕ_M and ϕ_P**
- **(m, p) pairs encoded as tensor product**: $\phi\big((m,p)\big) = \phi_M(m) \otimes \phi_P(p)$
- **"classical" SVM optimisation in (m, p) space**

Compute protein mapping function ϕ_P :

- **choose a similarity measure k_P defining matrix K_P** as: $(K_p)_{ij} = k_P(p_i, p_j)$
- **define X_P such as: $K_P = X_P . X_P^T$**
- **compute matrix X_P** : defines kernel-derived encoding ϕ_P

X_P is computed by SVD decomposition of K_P

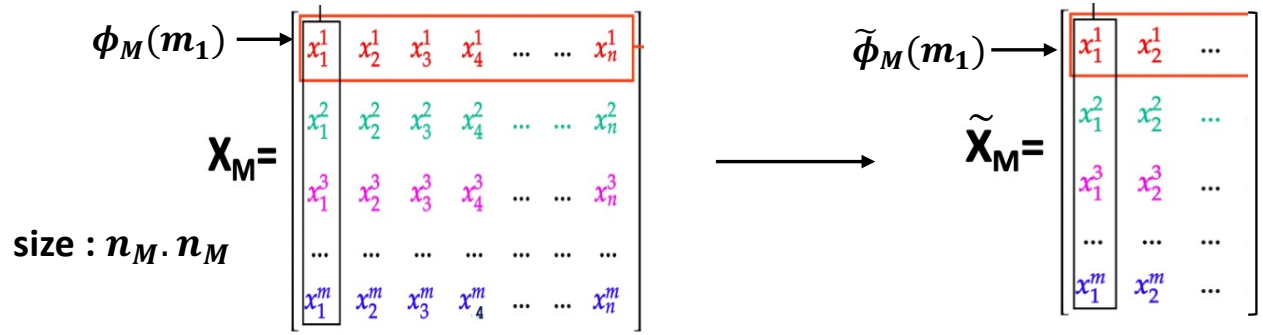
Computation of K_P and X_P "feasible" : size $n_P . n_P$ (n_P= 2060 in LCIdb)



Matrix representation of the protein data

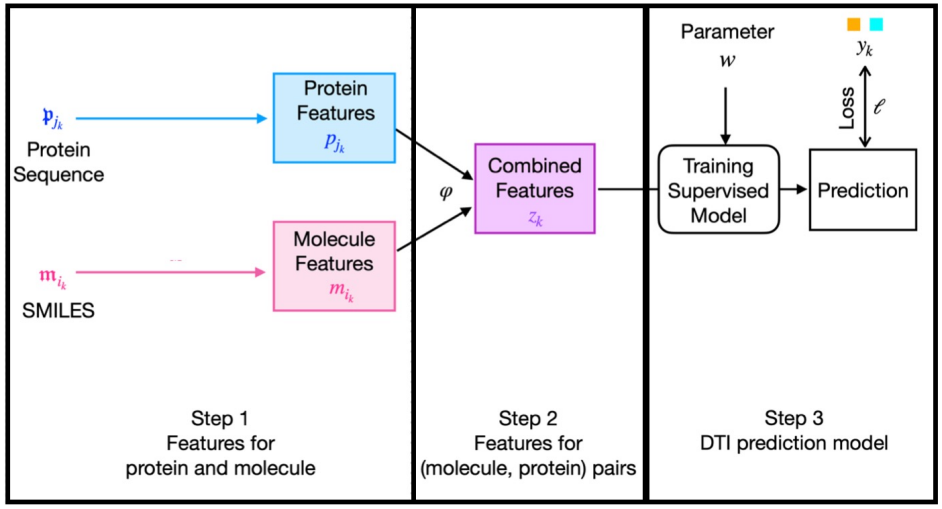
Compute molecule mapping function ϕ_M : not feasible because X_M and K_M too large (size $n_M . n_M$= (271.180)^2 in LCIbd)

Komet efficiently computes $\widetilde{X_M}$ of much smaller size, such that : $K_M \approx \widetilde{X_M} . \widetilde{X_M}^T$



size : $n_M . d_M$ with $d_M \ll n_M$

Uses sampling landmark molecules in LCIdb and dimension reduction during SVD



Step 1: compute $\phi_P(p)$ and $\tilde{\phi}_M(m)$
Steps 2 and 3: "classical" SVM optimisation computation of hyperplane in space $\phi\big((m,p)\big) = \tilde{\phi}_M(m) \otimes \phi_P(p)$
KOMET efficiently solves the optimisation

molecules: SMILE format, **ECFP4 fingerprints** = 1024-bit binary vector, K_M = **Tanimoto kernel**
m_M= 3000 and d_M=1000

proteins: aa sequence, K_P **LA-kernel** (local alignment).

KOMET prediction performances and computational efficiency (LCIdb)

Table 4: Comparison of AUPR scores on large-sized datasets, in 5-fold cross-validation.

Dataset	Komet	ConPLex	MolTrans	RF with concatenated features
LCIdb	**0.9925±0.0004**	0.9783±0.0008	0.9721±0.0011	0.9865±0.0002
Unseen_drugs	**0.9944±0.0003**	0.9831±0.0009	0.9710±0.0004	0.9829±0.0006
Unseen_targets	**0.8952±0.0186**	0.8780±0.0223	0.5987±0.0131	0.6886±0.0232
Orphan	**0.8671±0.0075**	0.8175±.0130	0.5455±0.0004	0.5961±0.0070

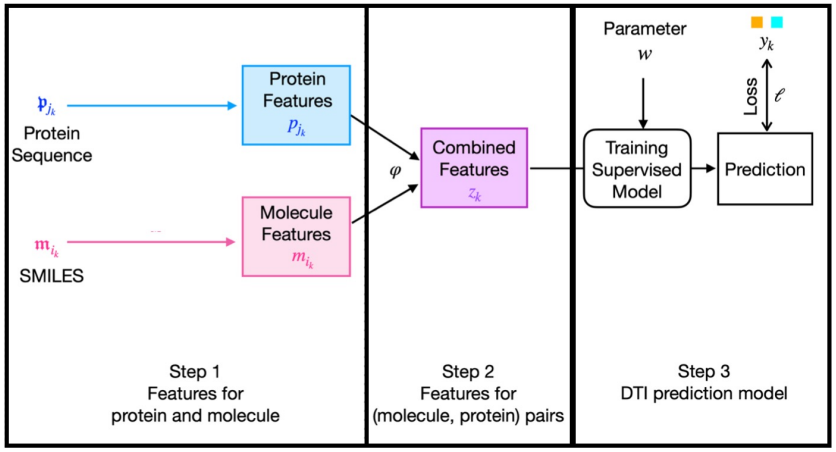
Table 5: Comparison of training time for the considered algorithms.

	Komet	ConPLex	MolTrans	RF with concatenated features
LCIdb	**15s**	907.3s	69838s	4391s
Unseen_drugs	**15s**	1734s	68400s	4213s
Unseen_targets	**15s**	888s	64800s	4100s
Orphan	**8s**	1329s	25200s	1297s

2 CPUs and 1 NVIDIA A40 GPU
with 48 GB of memory

Deep-learning algorithms

Guichaoua et al, JCIM 2024



combine with NN

KOMET also outperforms deep-learning approaches on medium size datasets (BIOSNAP, BindingDB, DrugBank)

Unpublished results

Target prediction for CFTR modulators with ML chemogenomics

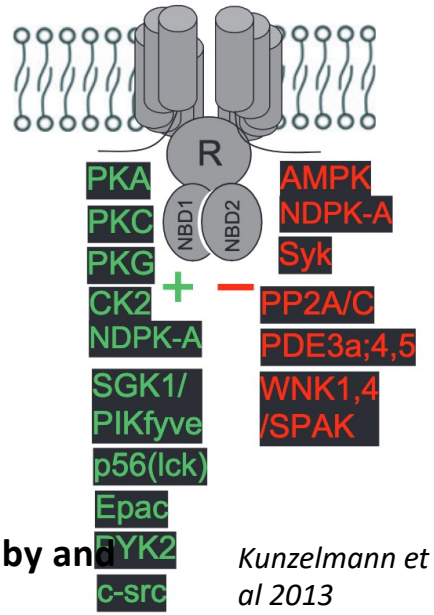
For **Elexacaftor,** Tezacaftor, and Ivacaftor (ETI= Trikafta™)

No high confidence predictions for Ivacaftor and Tezacaftor

For **Elexacaftor**: various **kinases** with high DTI probability scores (> 0.8)
50 kinases experimentally tested.
Modest but significant anti-kinase activity (µM range) for: **SYK, GSK3B** (*in vitro* tests)
Confirmed by *in vivo* functional tests

SYK and GSK3B appear highly relevant in the context of CF
- CFTR phosphorylation by **SYK** reduces its stability at the membrane (recycling)
- **GSK3B** inhibitor Kenpaullone improves maturation and processing of F508del-CFTR
(Trzcinska 2012)

Moderate inhibition of SYK and GSK3B by Elexacaftor may contribute to rescue of mutated CFTR by and increase Chloride conductance



Kunzelmann et al 2013

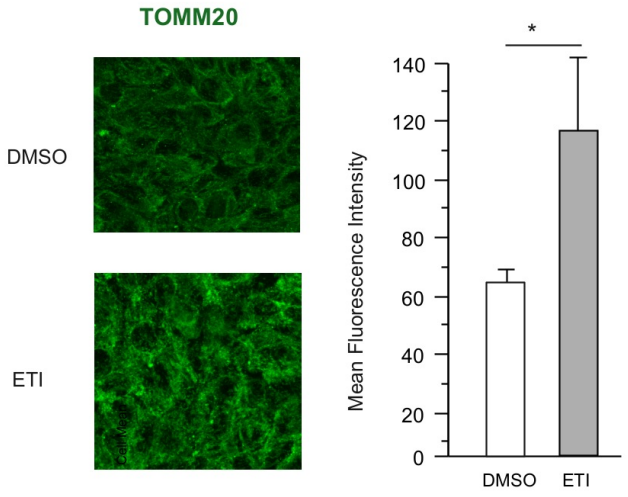
Further relevance of GSK3B new therapeutic target in CF

Unpublished results

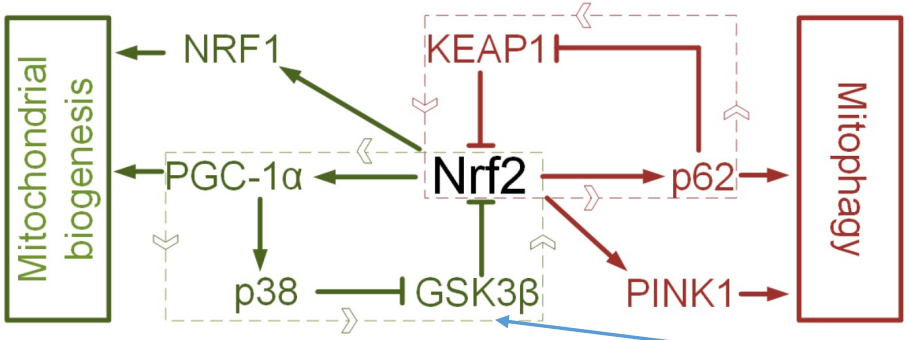
ETI combination improves mitochondrial fitness in CF airway epithelial cells:

- Increases mitochondrial mass, restores Oxydative Phosphorylation activity

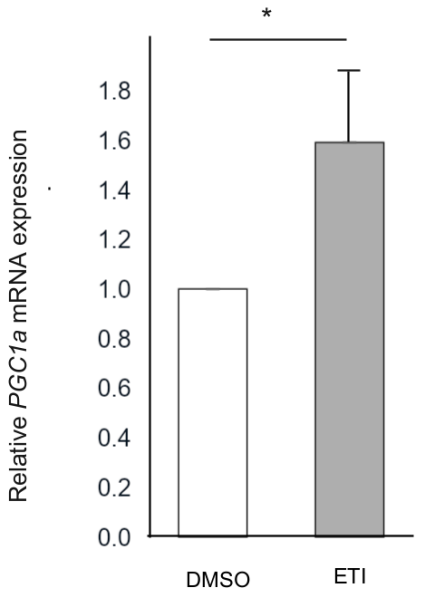
- Increases level of PGC1-α (major mitochondrial biogenesis regulator)
- Increases NRF2 levels (TF involved in response to oxidative stress)



From Guerev *et al,* Brain sciences 2020



Inhibition of GSK3B by Elexacaftor may explain metabolic benefit of ETI



→ **Specific inhibitors of GSK3B could improve CF mitochondrial fitness**

Further relevance of SYK as therapeutic target in CF

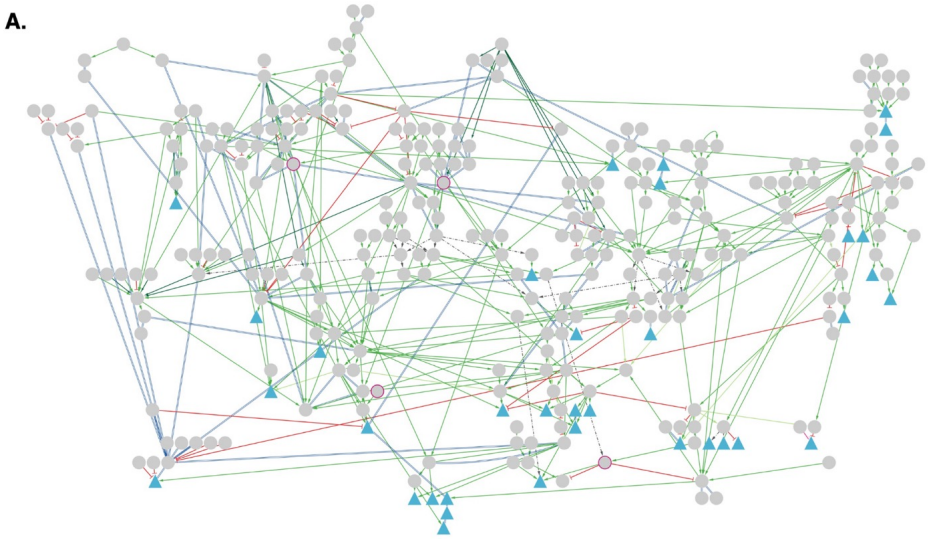
Systems Biology approach of CF *(Najm et al 2024)*:

Rationale: **Absence of CFTR propagates further molecular deregulation via its interaction partners leading to various cellular phenotypes that explain diversity of CF symptoms**

Goal: Build a signalling network that recapitulates CF signalling deregulations
- better understanding of the disease
- identify new therapeutic target whose modulation may modulate CF phenotypes

Network built based on transcriptomic data of CF airway Epithelial cells vs WT counterparts

317 nodes, 529 interactions



Najm et al, BMC Genomics 2024

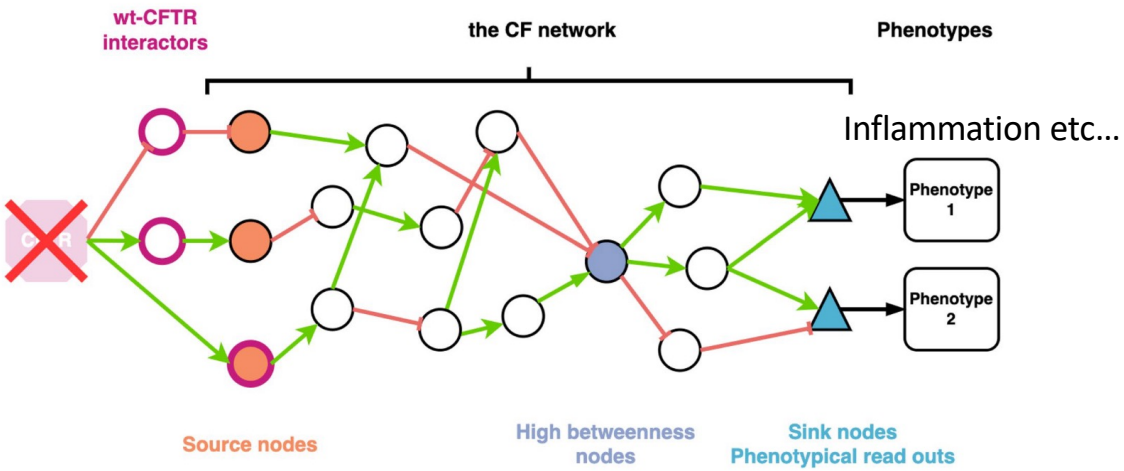
Identify therapeutic targets based on topological analysis of the network

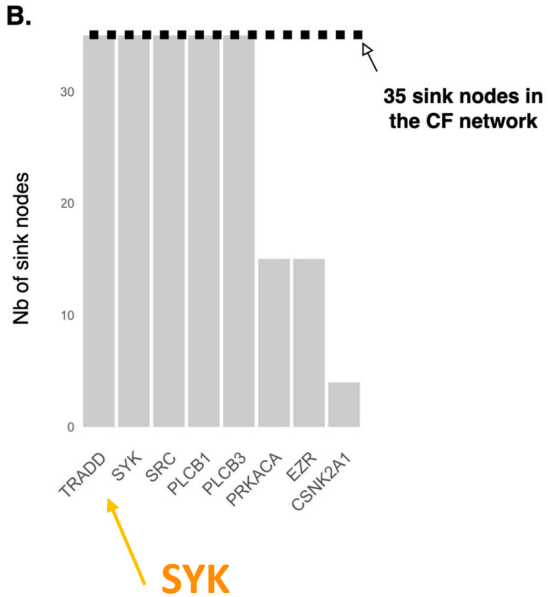
”remarkable” nodes

Source nodes (initiate propagation of deregulations)

Hubs (receive and propagate deregulations)

Output nodes (modulate CF phenotypes)





Among source nodes, SYK is connected to the 35 sink nodes.

Moderate inhibition of SYK by Elexacaftor may contribute to reduction of inflammation by ETI

Specific inhibition of SYK may modulate all CF phenotypes in the network

Najm et al, BMC Genomics 2024

Ongoing work

Further experimental validations of SYK, GSK3B on CF airway respiratory cells using readouts related to CF phenotypes (other than chloride conductance)

Effect of available specific inhibitors of SYK and GSK3B on :
- inflammation markers
- Trans Epithelial Electrical Resistance (TEER)
- wound repair (scratch tests)

Longer term:
SYK : an interesting target for numerous inflammatory diseases. Fostamatinib FDA approved, others in development

No specific inhibitor of GSK3B approved yet, but several are in development

→ **”mutation agnostic” therapies** to modulate deleterious CF symptoms

Acknowledgements

CBIO: Mathieu Najm, Gwenn Guichaoua, Philippe Pinel, Victor Laigle, Chloé-Agathe Azencott





MINES PARIS | PSL

INEM, U1151: Isabelle Sermet-Gaudelus, Mairead Aubert


Inserm
La science pour la santé
From science to health

Institut Curie, U1331: Loradana Martignetti, Laurence Calzone


institut Curie

Financial Support: La Fondation Dassault Systèmes, Vaincre la Mucoviscidose, Fondation pour la Recherche Médicale, La Fondation Maladies Rares

Datasets	Molecules	Proteins	Positive DTIs	Negative DTIs
BIOSNAP	4,510	2,181	13,836	(13,647 random)
Unseen_drugs			13,836	(13,647 random)
Unseen_targets			13,836	(13,647 random)
BindingDB	7,161	1,254	9,166	23,435
DrugBank	4,813	2,507	13,715	(13,715 balanced)
DrugBank (Ext)	4,257	1,216	10,838	(10,838 balanced)
LCIdb	271,180	2,060	396,798	7,965 (+ 388,833 balanced)
Unseen_drugs	271,180	2,060	396,798	7,965 (+ 388,833 balanced)
Unseen_targets	271,180	2,060	396,798	7,965 (+ 388,833 balanced)
Orphan	191,901	2,060	208,041	7,965 (+ 200,076 balanced)

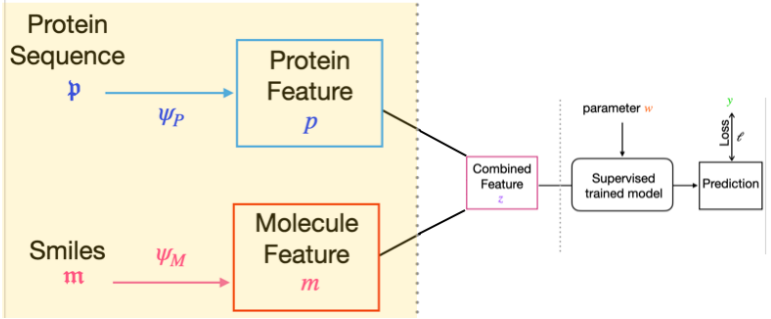
LCIdb

30 times larger than BIOSNAP and DrugBank

43 times larger than BindingDB

Guichaoua et al, JCIM 2024

Impact of molecule and protein features choice



Comparison of fixed and pre-trained learned features

Performance in the more difficult scenario: LCIdb_Orphan Dataset

AUPR		Protein feature	
		Feature map of LA kernel	**ESM2** [Lin *et al, 2022*]
Molecule Feature	**Feature map of Tanimoto kernel on ECFP4**	**0.897**	0.864
	ECFP4	0.893	0.866
	Dgl-lifesci [Li et al, 2021] **(GNN supervised contextpred)**	0.887	0.858

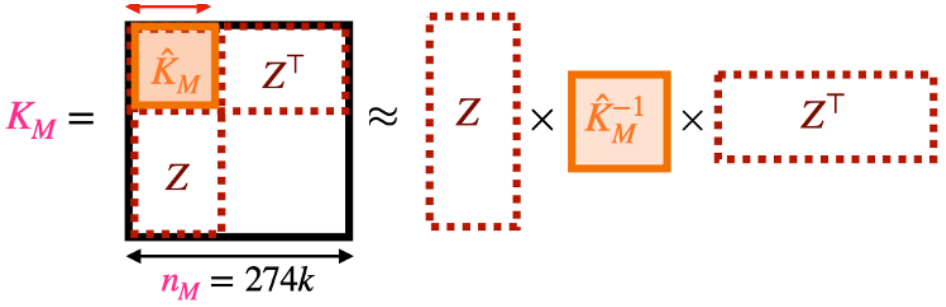
Fixed features better than DL features for this specific problem

Drug-like molecules and human druggable proteins

Molecule kernel $k_M(\mathfrak{m}, \mathfrak{m}') = \langle \psi_M(\mathfrak{m}), \psi_M(\mathfrak{m}') \rangle$ $K_M = X_M X_M^\top$

Nyström approximation

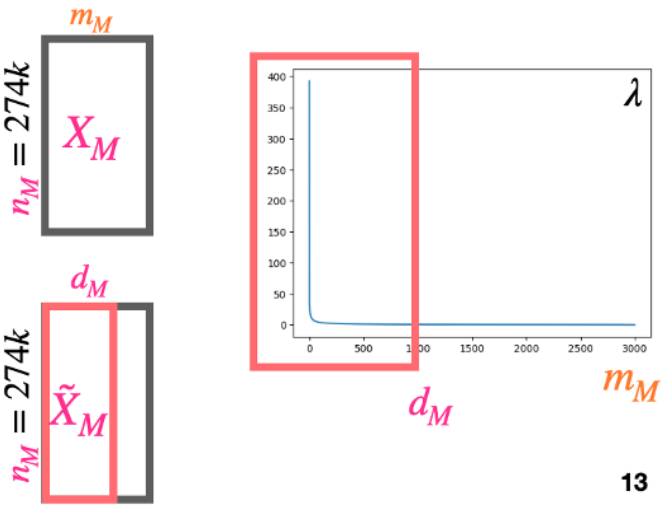
m_M random landmarks molecules



Singular value decomposition (SVD)

$\hat{K}_M = U \text{diag}(\lambda) U^\top$

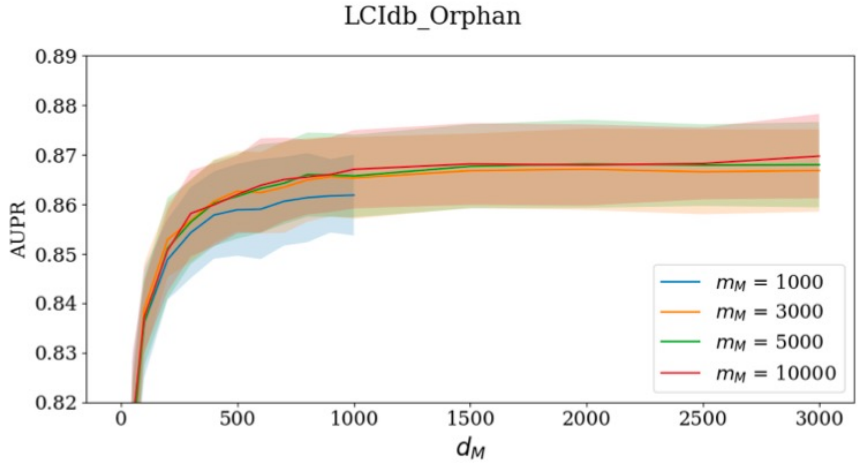
$K_M \approx X_M X_M^\top$ where $X_M = ZU\text{diag}(1/\sqrt{\lambda})$

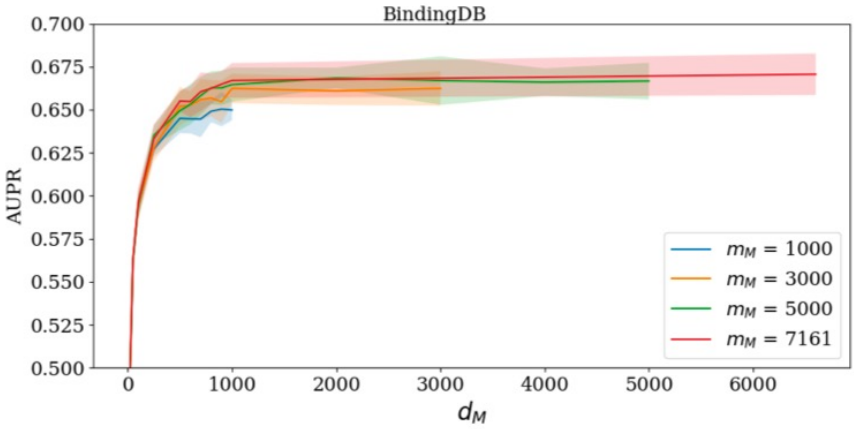


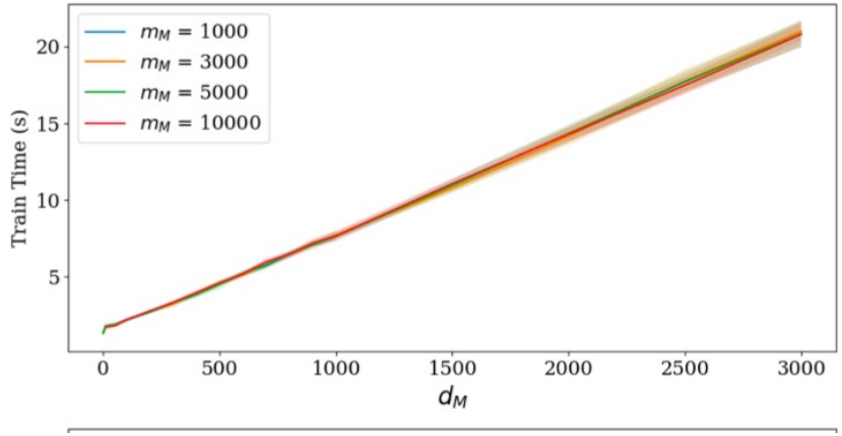
Reduction dimension

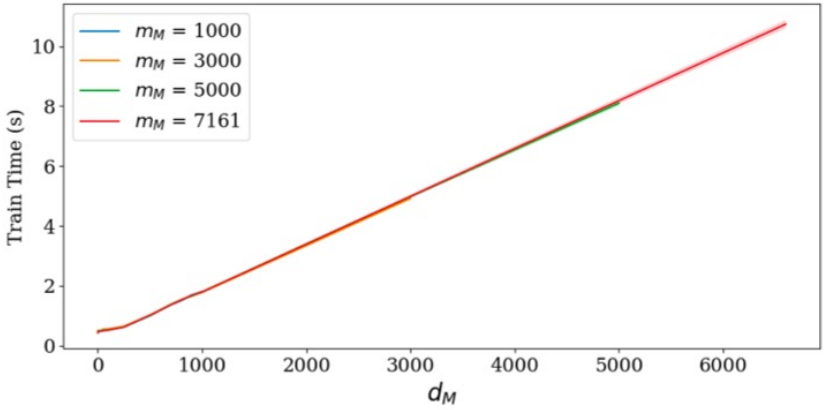
$X_M \approx \tilde{X}_M$ where $\tilde{X}_M \in \mathbb{R}^{n_M \times d_M}$

13


LCIdb_Orphan
AUPR
d_M
m_M = 1000
m_M = 3000
m_M = 5000
m_M = 10000


BindingDB
AUPR
d_M
m_M = 1000
m_M = 3000
m_M = 5000
m_M = 7161


Train Time (s)
d_M
m_M = 1000
m_M = 3000
m_M = 5000
m_M = 10000


Train Time (s)
d_M
m_M = 1000
m_M = 3000
m_M = 5000
m_M = 7161

KOMET Prediction performances:

On large size datasets (LCIdb)

Table 4: Comparison of AUPR scores on large-sized datasets, in 5-fold cross-validation.

Dataset	Komet	ConPLex	MolTrans	RF with concatenated features
LCIdb	**0.9925±0.0004**	0.9783±0.0008	0.9721±0.0011	0.9865±0.0002
Unseen_drugs	**0.9944±0.0003**	0.9831±0.0009	0.9710±0.0004	0.9829±0.0006
Unseen_targets	**0.8952±0.0186**	0.8780±0.0223	0.5987±0.0131	0.6886±0.0232
Orphan	**0.8671±0.0075**	0.8175±.0130	0.5455±0.0004	0.5961±0.0070

Table 5: Comparison of training time for the considered algorithms.

	Komet	ConPLex	MolTrans	RF with concatenated features
LCIdb	**15s**	907.3s	69838s	4391s
Unseen_drugs	**15s**	1734s	68400s	4213s
Unseen_targets	**15s**	888s	64800s	4100s
Orphan	**8s**	1329s	25200s	1297s

On medium size datasets

Dataset	Komet	ConPLex	MolTrans	RF with concatenated features
BIOSNAP	**0.9429±0.0008**	0.9246±0.0037	0.8989±0.0048	0.9121±0.0032
Unseen_drugs	**0.8979±0.0051**	0.8763±0.0072	0.8547±0.0045	0.8620±0.0100
Unseen_targets	**0.8754±0.0099**	0.8641±0.0100	0.7058±0.0273	0.8127±0.0116
BindingDB	0.6598±0.0074	**0.6765±0.0178**	0.6196±0.0150	0.6454±0.0075
DrugBank	**0.9400±0.0030**	0.8961±0.0070	0.8068±0.0100	0.8018±0.0086

Deep-learning algorithms