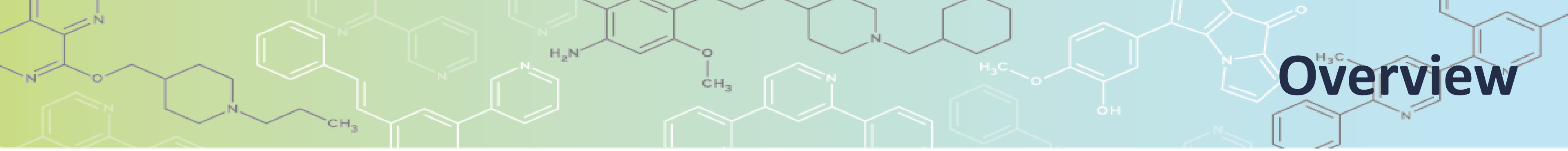
Emerging patterns mining and automated detection of contrasting chemical features

Alban Lepailleur

Centre d'Etudes et de Recherche sur le Médicament de Normandie (CERMN)
UNICAEN EA 4258 - FR CNRS 3038 INC3M - SF 4206 ICORE
Université de Caen Normandie, France

Chemoinformatics Strasbourg Summer School 2016



Overview

- Emerging Pattern (EP) mining
  - Generalities
  - Our contributions

- Case study I : Detection of structural alerts for the mutagenicity endpoint

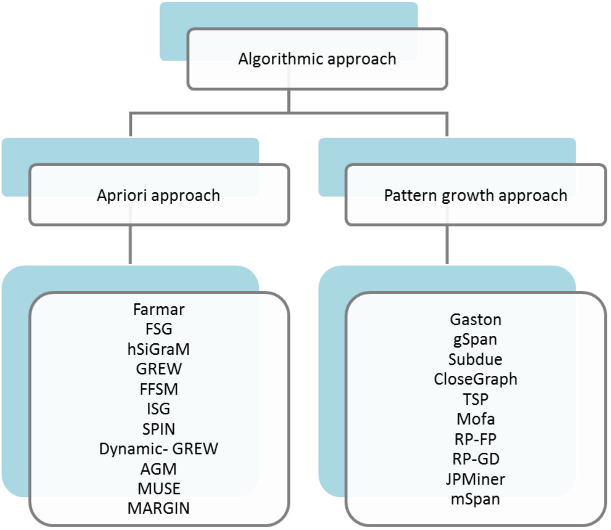
- Case study II : Polypharmacology of kinases

Data mining in chemoinformatics

- With the explosion of the availability of data, we need new methods to identify structure-activity relationships in large databases
  - LeadScope, ChEMBL, PubChem, …

- The calculation of the frequency of a descriptor is often at the core of the process

- Algorithms for the calculation of frequent descriptors often lead to the generation of myriads of such descriptors



Algorithmic approach

Apriori approach

Farmar
FSG
hSiGraM
GREW
FFSM
ISG
SPIN
Dynamic- GREW
AGM
MUSE
MARGIN

Pattern growth approach

Gaston
gSpan
Subdue
CloseGraph
TSP
Mofa
RP-FP
RP-GD
JPMiner
mSpan

Emerging patterns mining

- To limit the number of generated descriptors, methods have been proposed for finding representative and significant subsets
- Emerging pattern mining is a data mining technique introduced by Dong and Li [1,2] that captures differentiating features between 2 classes of data

Descriptors

	d1	d2	d3	d4	d5
mol1	X				X
mol2	X	X	X		X
mol3				X	
mol4	X	X			
mol5	X	X		X	
mol6	X	X			X
mol7					X
mol8			X		
mol9	X		X		X
mol10	X		X		

[1] Dong G. & Li J. Efficient mining of Emerging Patterns: Discovering trends and differences. *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Min., 5th* **1999**, 43-52.
[2] *Contrast Data Mining: Concepts, Algorithms, and Applications*; Dong G. & Bailey J., Eds.; CRC Press: Boca Raton, FL, **2013**.

Emerging patterns mining

- To limit the number of generated descriptors, methods have been proposed for finding representative and significant subsets
- Emerging pattern mining is a data mining technique introduced by Dong and Li [1,2] that captures differentiating features between 2 classes of data

Descriptors

	d1	d2	d3	d4	d5
mol1	X				X
mol2	**X**	**X**	X		X
mol3				X	
mol4	**X**	**X**			
mol5	**X**	**X**		X	
mol6	**X**	**X**			X
mol7					X
mol8			X		
mol9	X		X		X
mol10	X		X		

Emerging Pattern (EP)

{d1,d2}
is supported by
molecules [2,4,5]
and molecule [6]

Growth-rate
$\rho = 3$

[1] Dong G. & Li J. Efficient mining of Emerging Patterns: Discovering trends and differences. *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Min., 5th* **1999**, 43-52.
[2] *Contrast Data Mining: Concepts, Algorithms, and Applications*; Dong G. & Bailey J., Eds.; CRC Press: Boca Raton, FL, **2013**.

Emerging patterns mining

- To limit the number of generated descriptors, methods have been proposed for finding representative and significant subsets
- Emerging pattern mining is a data mining technique introduced by Dong and Li [1,2] that captures differentiating features between 2 classes of data

Descriptors

	d1	d2	d3	d4	d5
mol1	X				X
mol2	X	X	X		X
mol3				**X**	
mol4	X	X			
mol5	X	X		**X**	
mol6	X	X			X
mol7					X
mol8			X		
mol9	X		X		X
mol10	X		X		

Jumping Emerging Pattern (JEP)

{d4}
is supported by
molecules [3,5]

Growth-rate
$\rho = \infty$

[1] Dong G. & Li J. Efficient mining of Emerging Patterns: Discovering trends and differences. *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Min., 5th* **1999**, 43-52.
[2] *Contrast Data Mining: Concepts, Algorithms, and Applications*; Dong G. & Bailey J., Eds.; CRC Press: Boca Raton, FL, **2013**.

Applications of EP mining in chemoinformatics

- Auer and Bajorath were the firsts to apply EP mining in chemoinformatics [3,4]
  - Particularly, they introduced the notion of emerging chemical patterns (ECPs) for molecular classification

[3] Auer J. & Bajorath J. Emerging Chemical Patterns: A new methodology for molecular classification and compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502-2514.
[4] Namasivayam V.et al. Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2013**, *53*, 1272–1281.

- Sherhod and co-workers also applied EP mining for the identification of toxicophores for various toxicological endpoints [5,6]
  - Their method has been successfully used to implement new structural alerts for mutagenicity in the Derek Nexus expert system [7]

[5] Sherhod R.et al. Automating knowledge discovery for toxicity prediction using Jumping Emerging Pattern mining. *J. Chem. Inf. Model.* **2012**, *52*, 3074-3087.
[6] Sherhod R. et al. Emerging Pattern mining to aid toxicological knowledge discovery. *J. Chem. Inf. Model.* **2014**, *54*, 1864-1879.
[7] Coquin L. et al. New structural alerts for Ames mutagenicity discovered using Emerging Pattern mining techniques. *Toxicol. Res.* **2015**, *4*, 46- 56.

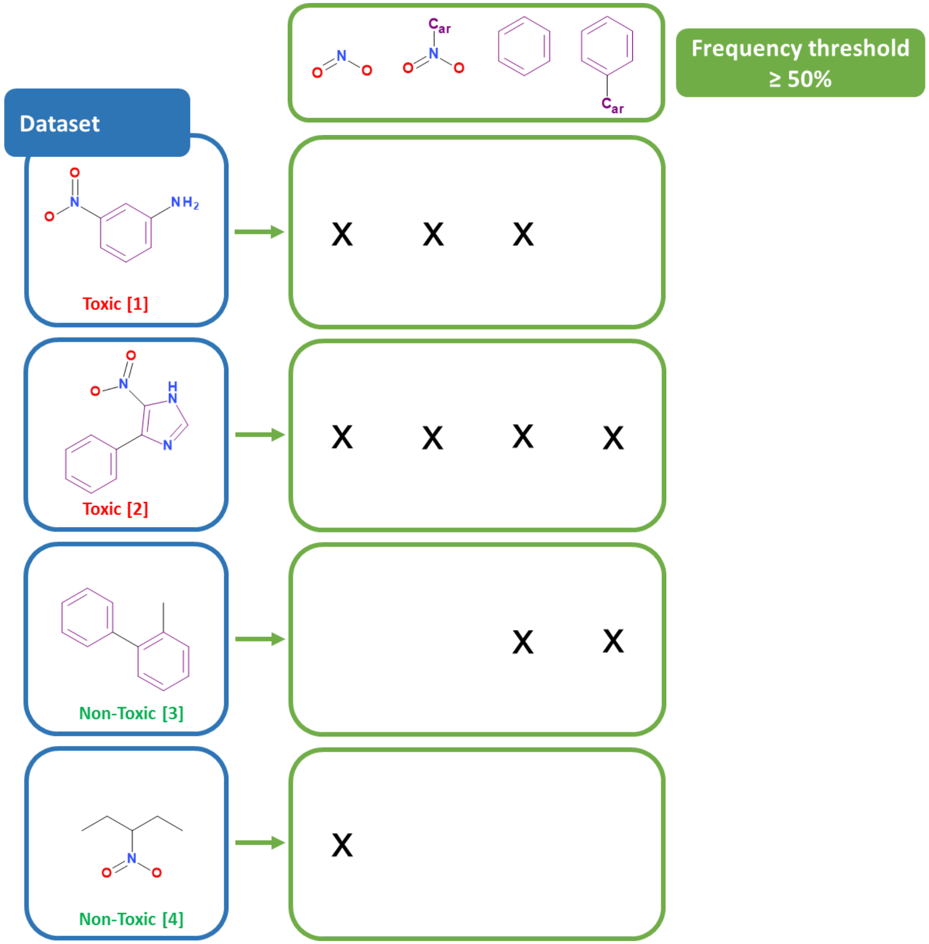
- Our contributions
  - We related the occurrences of jumping fragments to aquatic toxicity data [8]
  - We introduced the enumeration of combinations of chemical fragments [9,10]

[8] Lozano S. et al. Introduction of jumping fragments in combination with QSARs for the assessment of classification in ecotoxicology. *J. Chem. Inf. Model.* **2010**, *50*, 1330–1139.
[9] Poezevara G. et al. Extracting and summarizing the frequent emerging graph patterns from a dataset of graphs. *J. Intel. Inf. Syst.* **2011**, *37*, 333–353.
[10] Cuissart B. et al. Emerging Patterns as Structural Alerts for Computational Toxicology. In *Contrast Data Mining: Concepts, Algorithms and Applications*; Dong, G., Bailey, J., Eds.; Chapman and Hall/CRC, **2013**; pp 269–281.

Our method

- What makes our method [10] original:
  - We operate directly from the molecular graphs

[11] Metivier, J.P. et al. Discovering structural alerts for mutagenicity using Stable Emerging Molecular Patterns. *J. Chem. Inf Model.* **2015**, *55*, 925-940

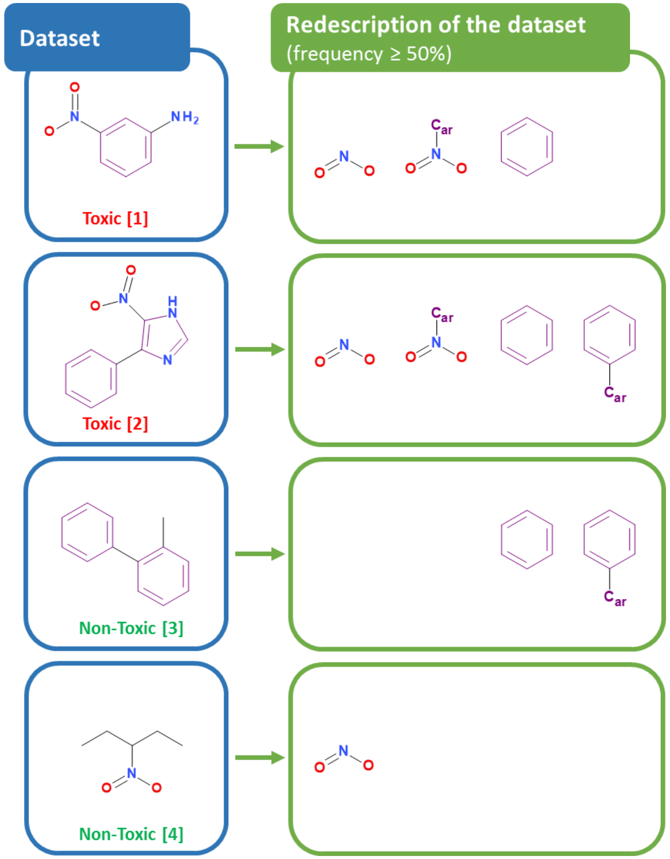


Toxic [1]	X	X	X	
Toxic [2]	X	X	X	X
Non-Toxic [3]			X	X
Non-Toxic [4]	X			

Our method

- What makes our method [11] original:
  - We operate directly from the molecular graphs

[11] Metivier, J.P. et al. Discovering structural alerts for mutagenicity using Stable Emerging Molecular Patterns. *J. Chem. Inf Model.* **2015**, *55*, 925-940



Dataset

Redescription of the dataset
(frequency ≥ 50%)

Toxic [1]

Toxic [2]

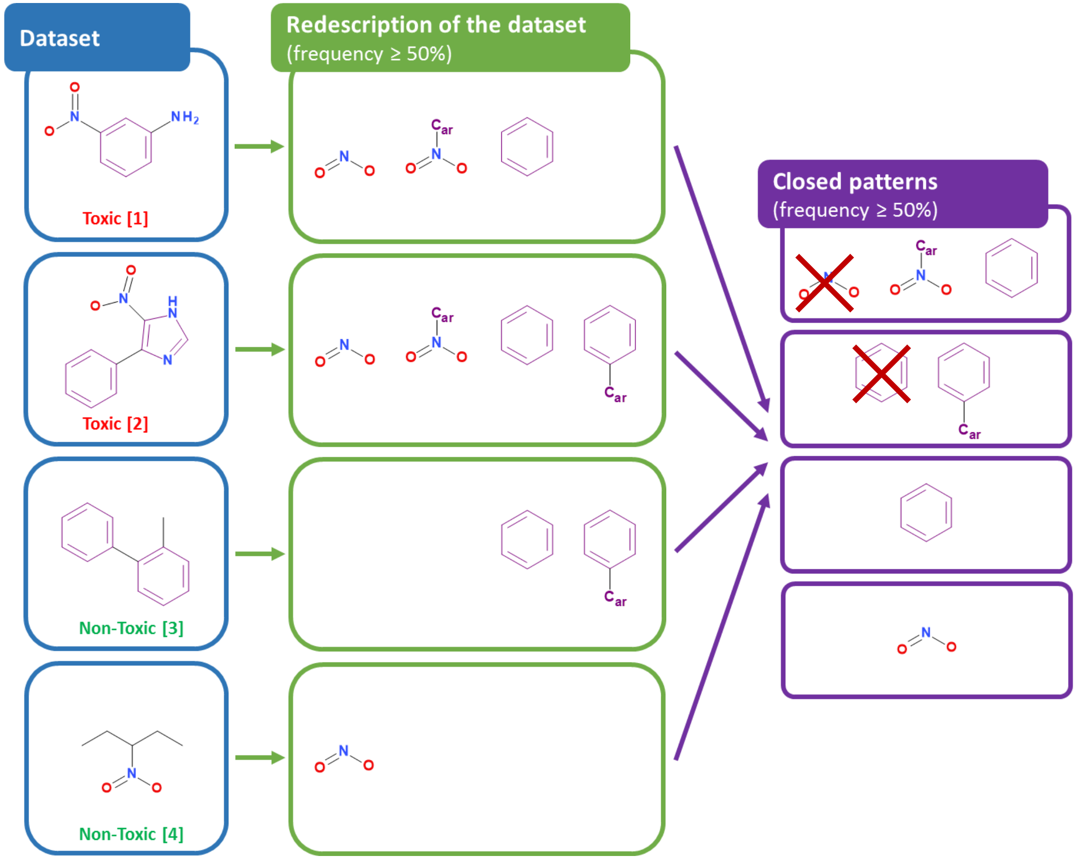
Non-Toxic [3]

Non-Toxic [4]

Our method

- What makes our method[11] original:
  - We operate directly from the molecular graphs
  - We enumerate the frequent closed patterns

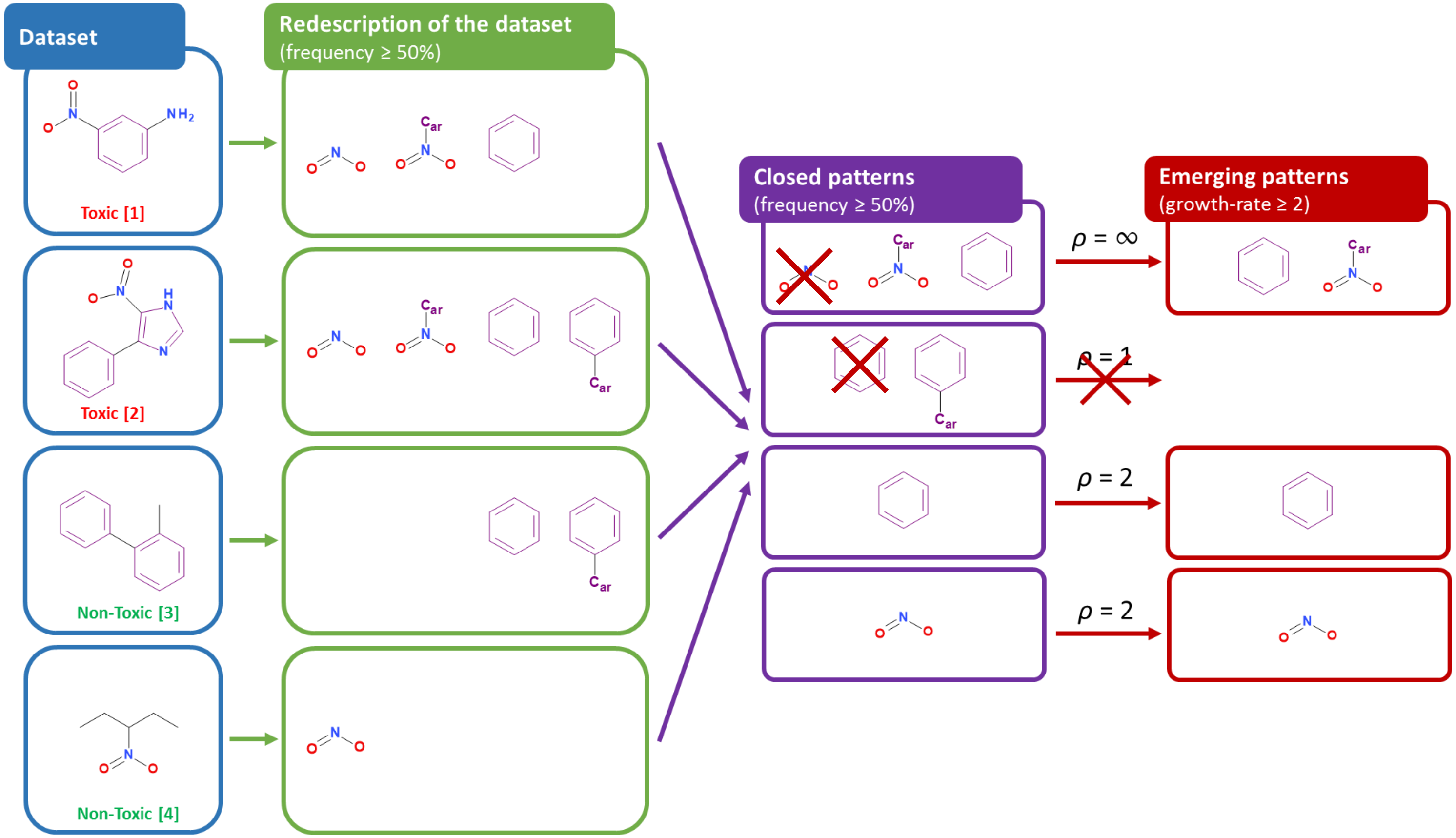
[11] Metivier, J.P. et al. Discovering structural alerts for mutagenicity using Stable Emerging Molecular Patterns. *J. Chem. Inf Model.* **2015**, *55*, 925-940



Our method

- What makes our method [11] original:
  - We operate directly from the molecular graphs
  - We enumerate the frequent closed patterns to extract the emerging patterns

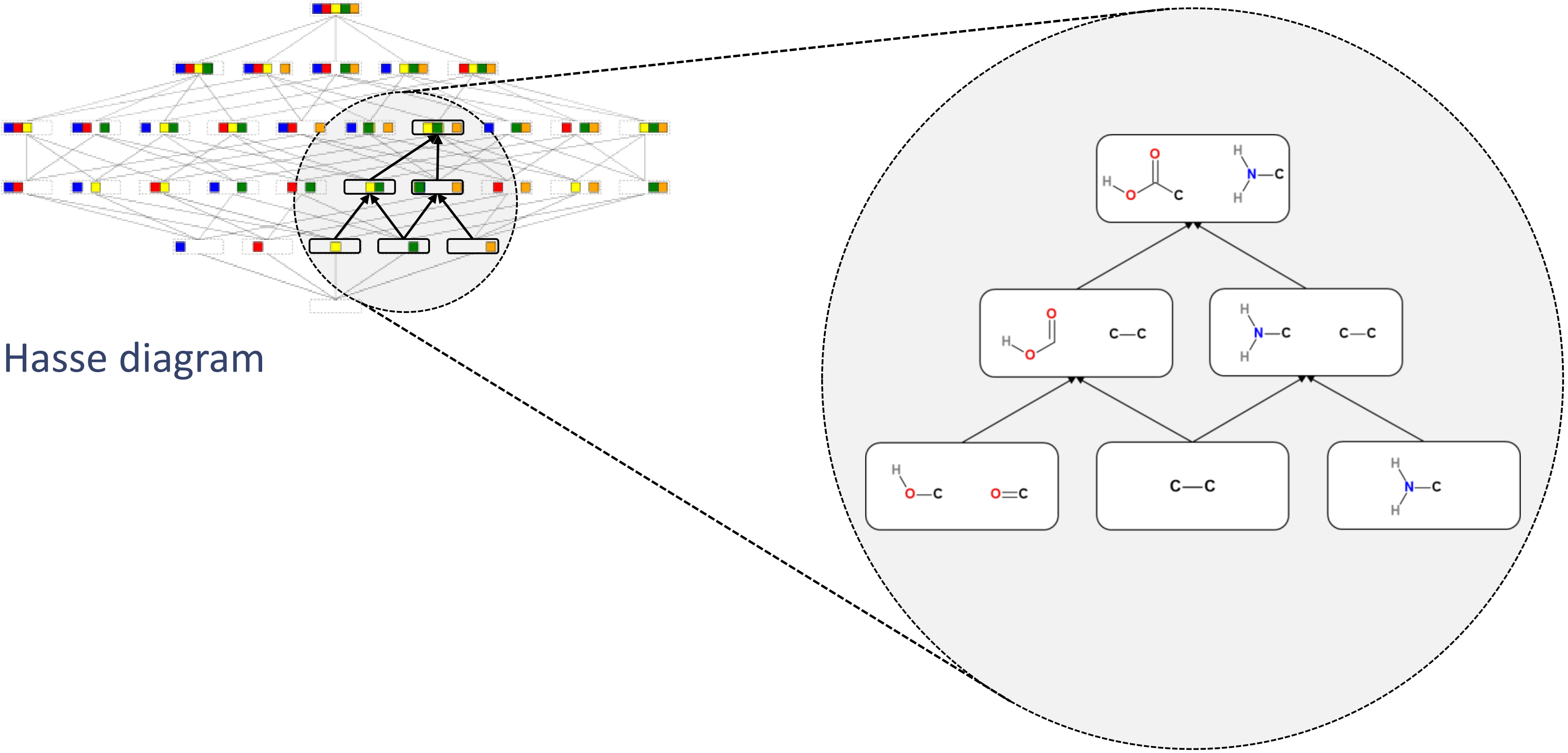
[11] Metivier, J.P. et al. Discovering structural alerts for mutagenicity using Stable Emerging Molecular Patterns. *J. Chem. Inf Model.* **2015**, *55*, 925-940



Our method

- What makes our method[11] original:
  - We operate directly from the molecular graphs
  - We enumerate the frequent closed patterns to extract the emerging patterns
  - We organize the patterns in a Hasse diagram

[11] Metivier, J.P. et al. Discovering structural alerts for mutagenicity using Stable Emerging Molecular Patterns. *J. Chem. Inf Model.* **2015**, *55*, 925-940



Hasse diagram

Case study (I): Structural alerts

- Search of structural alerts
  - One of the most interesting approach of predictive toxicology
  - Define the key features of a molecule that are required to initiate a toxicological pathway
  - Examples of domain experts rules
    - The Tennant and Ashby’s set for DNA reactivity
    - The Benigni and Bossa’s set for mutagenic and carcinogenic potential
    - ToxAlerts
  - The updating of a knowledge base is very time consuming since it requires strong investment of domain experts and a detailed analysis of the scientific literature

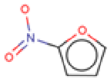
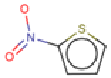
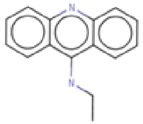
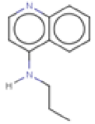
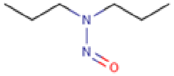
EP mining should reduce the time and efforts needed to identify new structural alerts

Case study (I): Structural alerts

❍ Search of structural alerts for mutagenicity

- The Hansen benchmark dataset (http://doc.ml.tu-berlin.de/toxbenchmark/)
  - □ 6512 compounds from the literature annotated with Ames mutagenicity data
    - ▪ 3503 Ames ⊕ and 3009 Ames ⊖
  - □ Use of a 0.36% frequency threshold (support of 20 molecules) → 15000 Eps

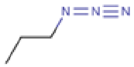
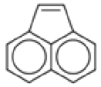
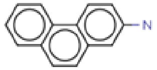
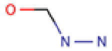
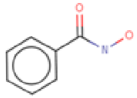
- 10 JEPs → only present in the mutagens

	JEPs	Support	ρ
MF_945		83	∞
MF_954		31	∞
MF_1666		27	∞
MF_991		32	∞
MF_252		27	∞

Nitro aromatic groups: MF_945, MF_954

Polycyclic aromatic amines: MF_1666, MF_991

Nitrosamine group: MF_252

	JEPs	Support	ρ
MF_4		28	∞
MF_1616		26	∞
MF_1211		25	∞
MF_87		25	∞
MF_414		37	∞

Azide group: MF_4

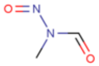
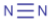
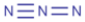
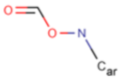
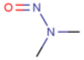
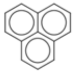
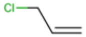
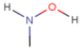
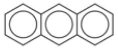
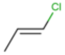
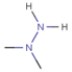
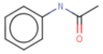
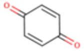
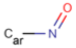
Polycyclic planar hydrocarbon system: MF_1616

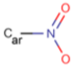
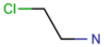
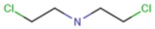
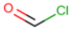
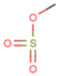
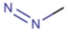
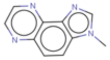
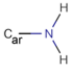
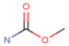
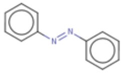
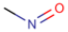
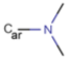
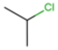
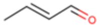
Generalized toxicophores: MF_1211, MF_87, MF_414

Case study (I): Structural alerts

❍ Search of structural alerts for mutagenicity

- Comparison with ToxAlerts (https://ochem.eu/) → 32 out of 50 toxicophores

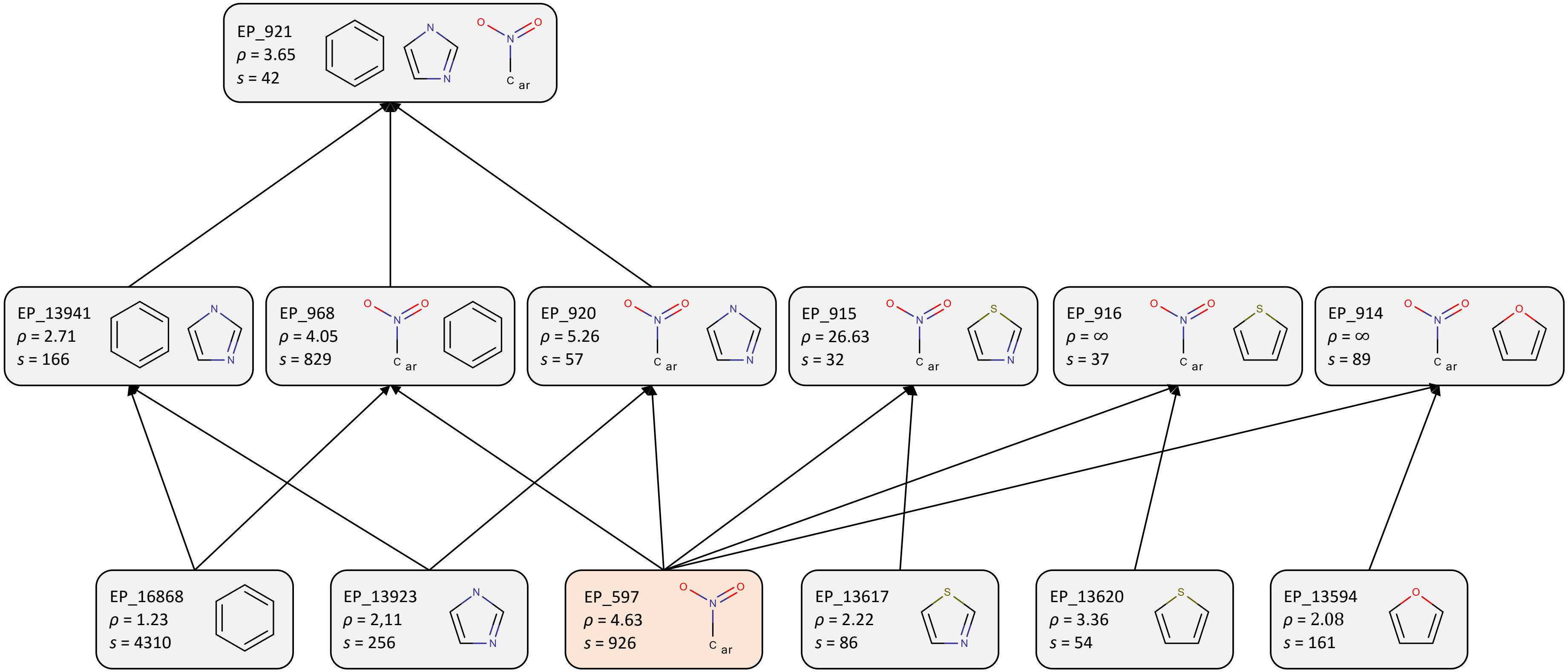
	EPs	Support	ρ	Structural Alert in ToxAlerts
MF_73		44	36.94	N-nitroso-N-alkylamides N-nitroso-N-alkylureas N-nitroso-N-alkylcarbamates
MF_0		67	27.92	Diazo
MF_1		55	22.76	Azide
MF_125		52	21.47	Aromatic and aliphatic aziridinyl derivatives
MF_156		47	12.60	Aromatic hydroxylamine ester
MF_72		27	9.09	Nitrosamine
MF_1651		178	6.79	Polycyclic aromatic hydrocarbons
MF_1841		26	6.59	Allylic halides
MF_824		48	6.01	Hydroxyl amine
MF_1667		226	5.84	Polycyclic aromatic hydrocarbons
MF_1831		36	5.33	Monohaloalkene
MF_75		162	5.19	Unsubstituted heteroatom-bonded heteroatom
MF_927		35	5.16	Aromatic N-acyl amine
MF_1298		27	4.94	Quinones
MF_920		964	4.70	Nitrosoarenes

	EPs	Support	ρ	Structural Alert in ToxAlerts
MF_134		926	4.63	Aromatic nitro groups
MF_444		93	4.47	Nitrogen mustard
MF_212		35	4.15	N mustard
MF_2188		29	4.12	Acyl halides
MF_2107		40	4.05	Alkyl ester of sulfonic and sulfuric acids
MF_34		74	3.68	Aliphatic azo
MF_1317		26	3.61	Heterocyclic polycyclic aromatic hydrocarbons
MF_1883		308	2.33	Aliphatic and aromatic epoxides
MF_916		656	2.17	Primary aromatic amine
MF_432		50	1.83	Alkyl carbamate
MF_41		151	1.54	Aromatic azo
MF_153		51	1.33	Aliphatic nitroso
MF_169		182	1.25	Tertiary aromatic amine
MF_2197		26	1.17	Aliphatic halogens
MF_1836		59	1.17	α,β-unsaturated carbonyl

Case study (I): Structural alerts

❍ Search of structural alerts for mutagenicity

- Extract of the Hasse diagram
  - □ Example: stimulation[12] of aromatic rings by the addition of a nitro group

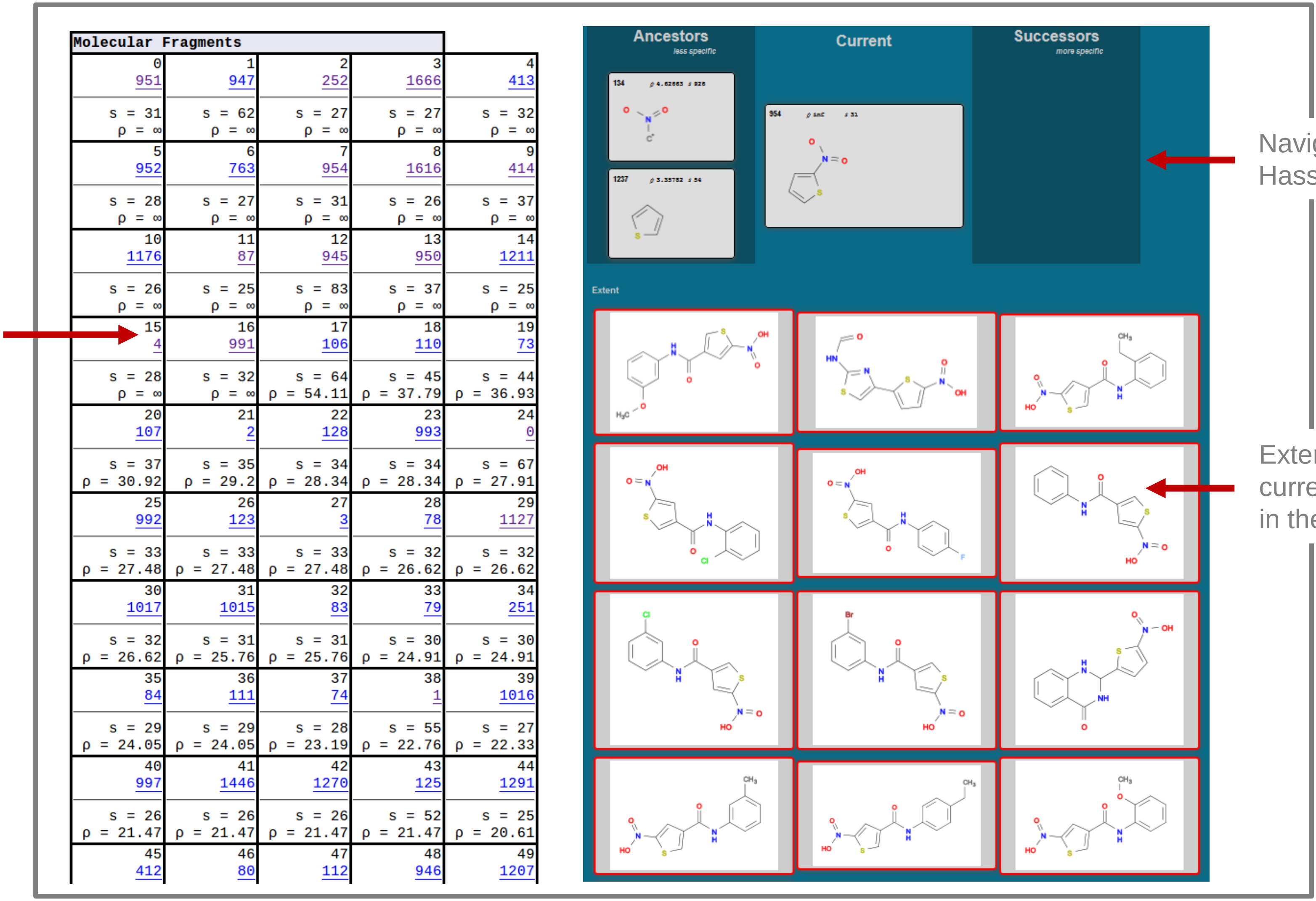


[12] Bissell-Siders R.et al. On the stimulation of patterns. *Lect. Notes Comput. Sci.* **2010**, *6208*, 56-69.

Case study (I): Structural alerts

❍ Our interactive visualization tool https://chemoinfo.greyc.fr/2014_Metivier/

List of patterns



Molecular Fragments

0
951
s = 31
ρ = ∞
	1
	947
	s = 62
	ρ = ∞
		2
		252
		s = 27
		ρ = ∞
			3
			1666
			s = 27
			ρ = ∞
				4
				413
				s = 32
				ρ = ∞

5
952
s = 28
ρ = ∞
	6
	763
	s = 27
	ρ = ∞
		7
		954
		s = 31
		ρ = ∞
			8
			1616
			s = 26
			ρ = ∞
				9
				414
				s = 37
				ρ = ∞

10
1176
s = 26
ρ = ∞
	11
	87
	s = 25
	ρ = ∞
		12
		945
		s = 83
		ρ = ∞
			13
			950
			s = 37
			ρ = ∞
				14
				1211
				s = 25
				ρ = ∞

15
4
s = 28
ρ = ∞
	16
	991
	s = 32
	ρ = ∞
		17
		106
		s = 64
		ρ = 54.11
			18
			110
			s = 45
			ρ = 37.79
				19
				73
				s = 44
				ρ = 36.93

20
107
s = 37
ρ = 30.92
	21
	2
	s = 35
	ρ = 29.2
		22
		128
		s = 34
		ρ = 28.34
			23
			993
			s = 34
			ρ = 28.34
				24
				0
				s = 67
				ρ = 27.91

25
992
s = 33
ρ = 27.48
	26
	123
	s = 33
	ρ = 27.48
		27
		3
		s = 33
		ρ = 27.48
			28
			78
			s = 32
			ρ = 26.62
				29
				1127
				s = 32
				ρ = 26.62

30
1017
s = 32
ρ = 26.62
	31
	1015
	s = 31
	ρ = 25.76
		32
		83
		s = 31
		ρ = 25.76
			33
			79
			s = 30
			ρ = 24.91
				34
				251
				s = 30
				ρ = 24.91

35
84
s = 29
ρ = 24.05
	36
	111
	s = 29
	ρ = 24.05
		37
		74
		s = 28
		ρ = 23.19
			38
			1
			s = 55
			ρ = 22.76
				39
				1016
				s = 27
				ρ = 22.33

40
997
s = 26
ρ = 21.47
	41
	1446
	s = 26
	ρ = 21.47
		42
		1270
		s = 26
		ρ = 21.47
			43
			125
			s = 52
			ρ = 21.47
				44
				1291
				s = 25
				ρ = 20.61

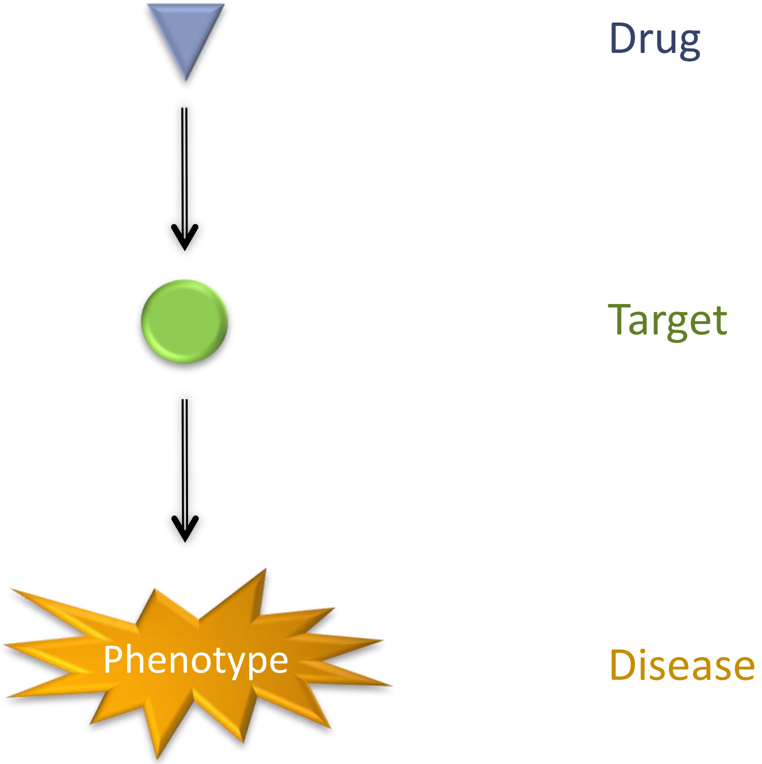
45 412	46 80	47 112	48 946	49 1207

Navigation in the Hasse diagram

Extent of the current pattern in the dataset

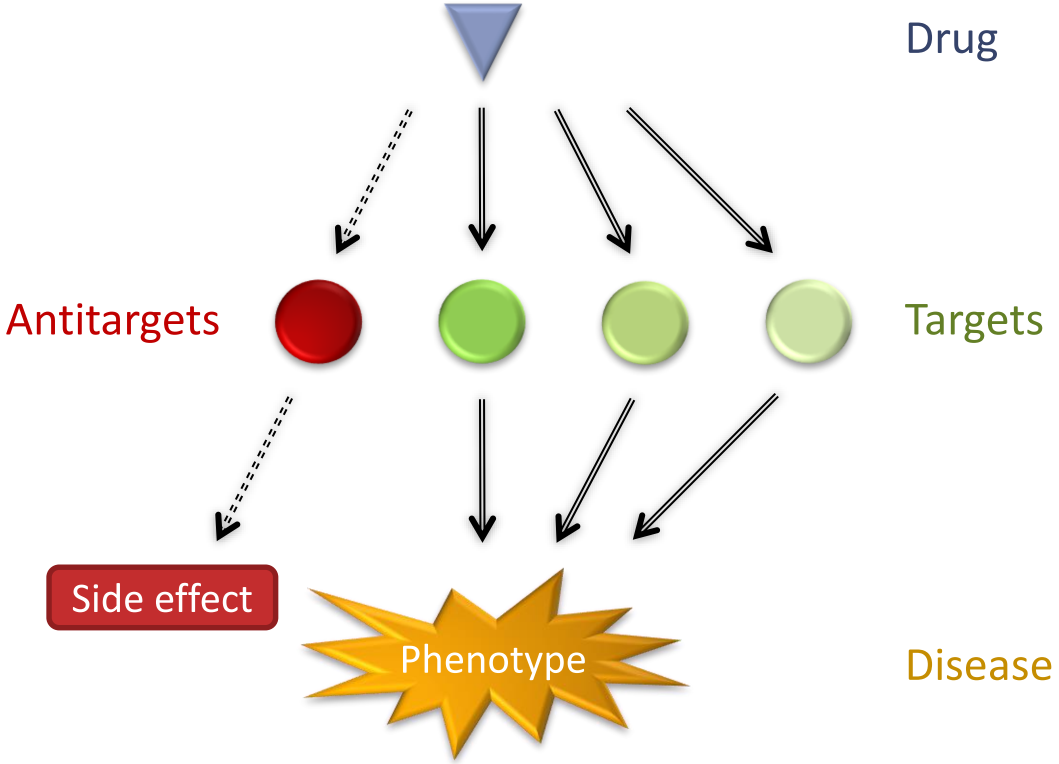
Case study (II): Polypharmacology

- The “one drug – one target – one disease” paradigm



Case study (II): Polypharmacology

- The "one drug – one target – one disease" paradigm
- Polypharmacological drug behavior
  - Many known drugs elicit their therapeutic effects by acting on multiple targets
  - But such drugs can also bind antitargets responsible for side effects

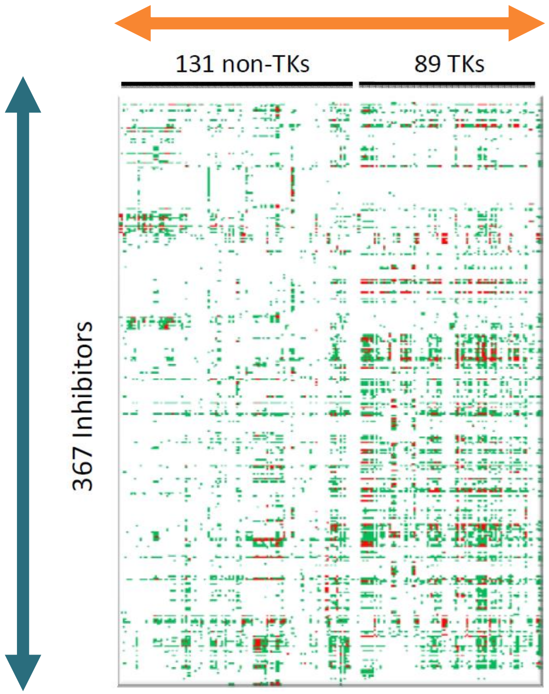


Case study (II): Polypharmacology

- Polypharmacology of kinases [13,14]
  - Most tumors can escape from the inhibition of any single kinase
  - The GSK Published Kinase Inhibitor Set (PKIS) as a source of knowledge

Annotation of the kinase inhibitors using affinity data for 220 kinases (active/inactive)

367 small molecules ATP-competitive kinase inhibitors



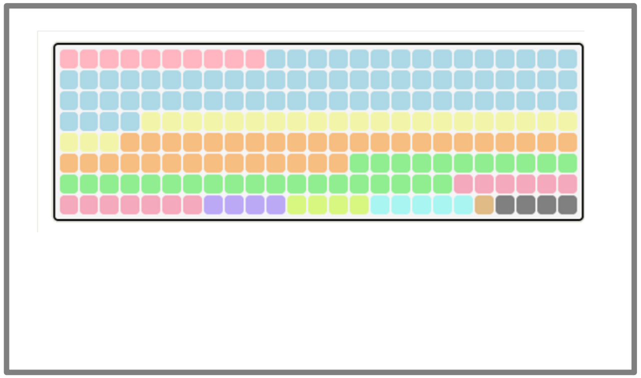
Data available from

ChEMBL

[13] Knight, Z.A. et al. Targeting the Cancer Kinome through Polypharmacology. *Nat. Rev. Cancer* **2010**, *10*, 130–137.
[14] Wu, P. et al. Small-Molecule Kinase Inhibitors: An Analysis of FDA-Approved Drugs. *Drug Discovery Today* **2016**, *21* (1), 5–10.

Case study (II): Polypharmacology

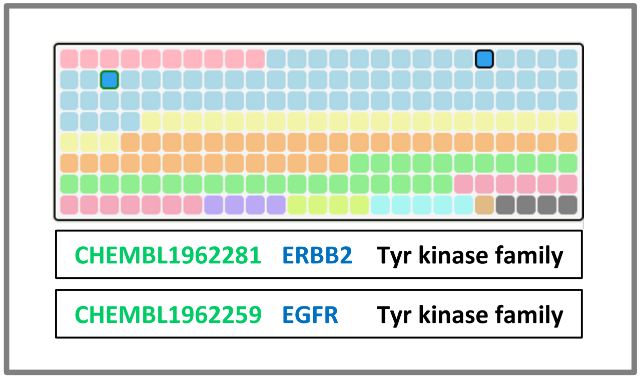
- Kinase Miner
  - Interactive tool dedicated to polypharmacology of kinases



Case study (II): Polypharmacology

❍ Kinase Miner

- Interactive tool dedicated to polypharmacology of kinases
- Example: dual inhibition of ERBB2 and EGFR



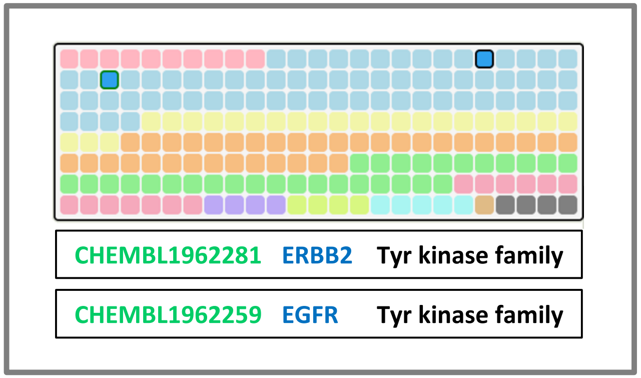
CHEMBL1962281 ERBB2 Tyr kinase family

CHEMBL1962259 EGFR Tyr kinase family

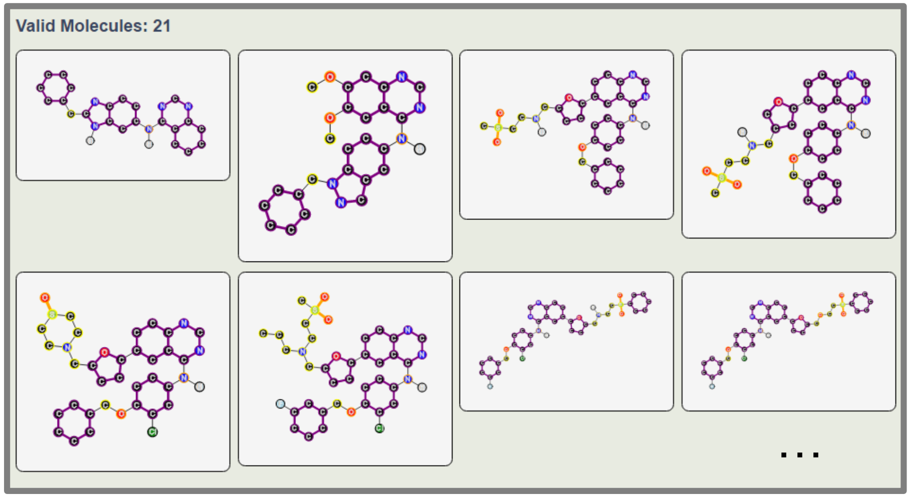
Case study (II): Polypharmacology

❍ Kinase Miner

- Interactive tool dedicated to polypharmacology of kinases
- Example: dual inhibition of ERBB2 and EGFR



CHEMBL1962281	ERBB2	Tyr kinase family
CHEMBL1962259	EGFR	Tyr kinase family



Valid Molecules: 21

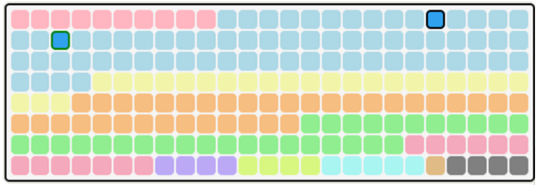
. . .

Molecules in agreement with the dual ERBB2 and EGFR inhibition

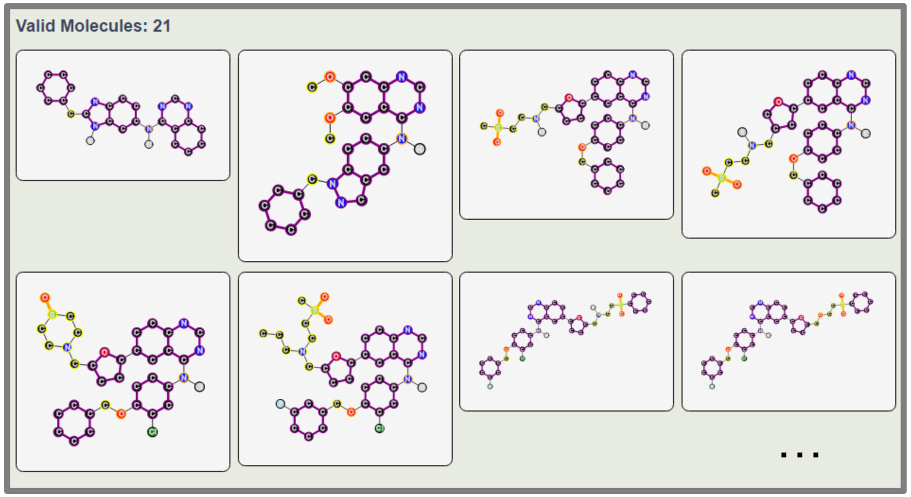
Case study (II): Polypharmacology

❍ Kinase Miner

- Interactive tool dedicated to polypharmacology of kinases
- Example: dual inhibition of ERBB2 and EGFR

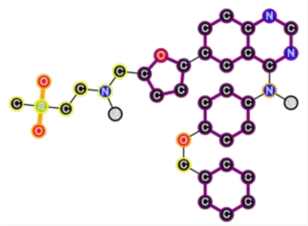


CHEMBL1962281	ERBB2	Tyr kinase family
CHEMBL1962259	EGFR	Tyr kinase family

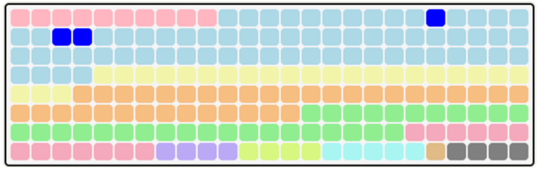


Valid Molecules: 21

. . .



CHEMBL529066

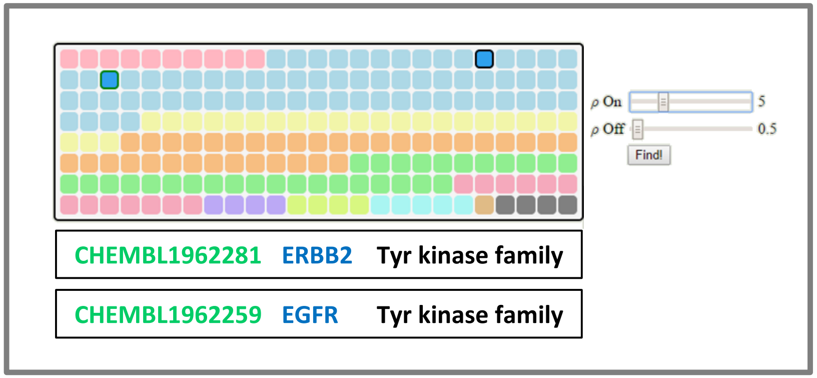


ERBB2	73.03 % inh @ 1µM
ERBB4	96.64 % inh @ 1µM
EGFR	94.13 % inh @ 1µM

Case study (II): Polypharmacology

❍ Kinase Miner

- Interactive tool dedicated to polypharmacology of kinases
- Example: dual inhibition of ERBB2 and EGFR



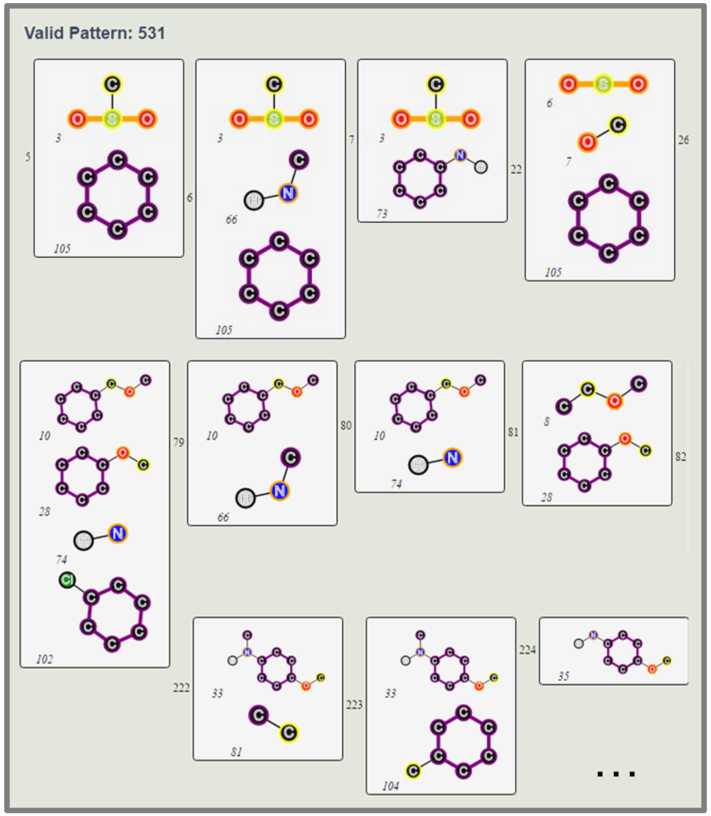
ρ On 5

ρ Off 0.5

Find!

CHEMBL1962281 ERBB2 Tyr kinase family

CHEMBL1962259 EGFR Tyr kinase family



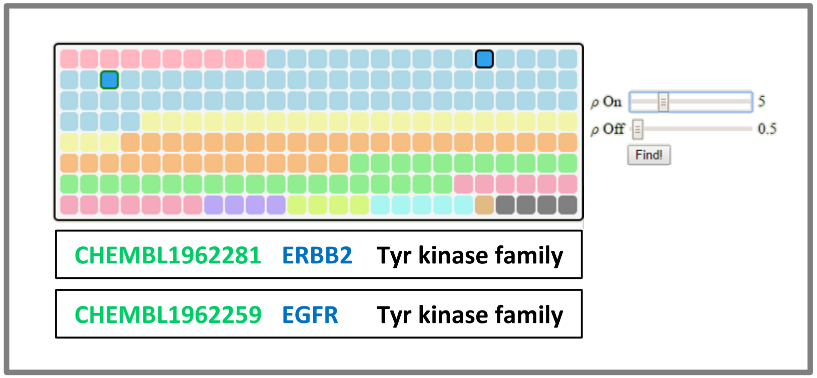
Valid Pattern: 531

. . .

Case study (II): Polypharmacology

❍ Kinase Miner

- Interactive tool dedicated to polypharmacology of kinases
- Example: dual inhibition of ERBB2 and EGFR

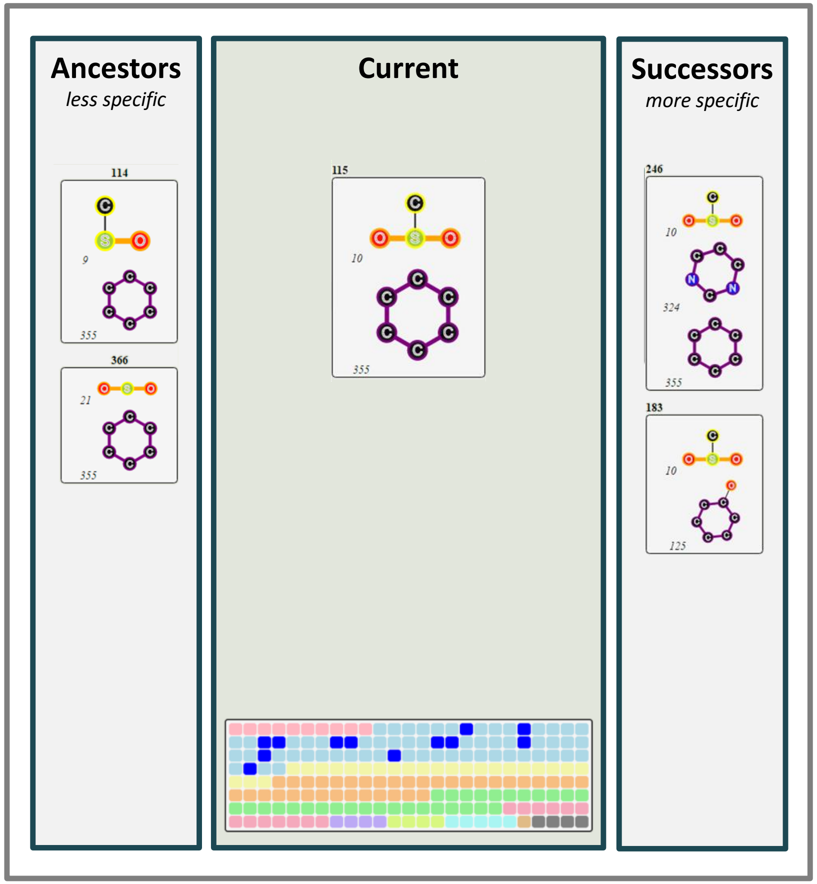


ρ On 5

ρ Off 0.5

Find!

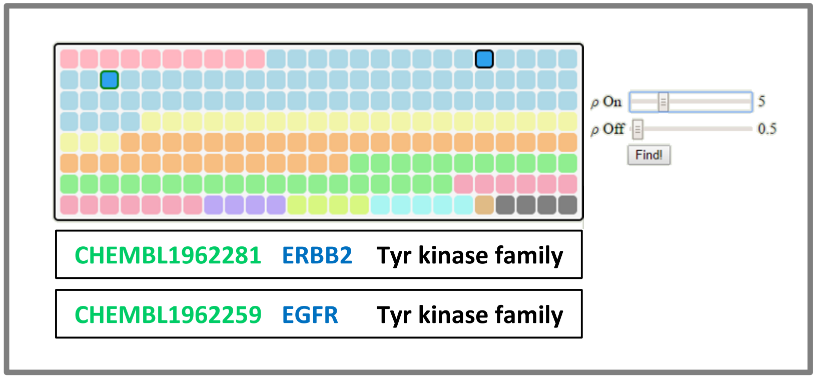
CHEMBL1962281	ERBB2	Tyr kinase family
CHEMBL1962259	EGFR	Tyr kinase family



Case study (II): Polypharmacology

❍ Kinase Miner

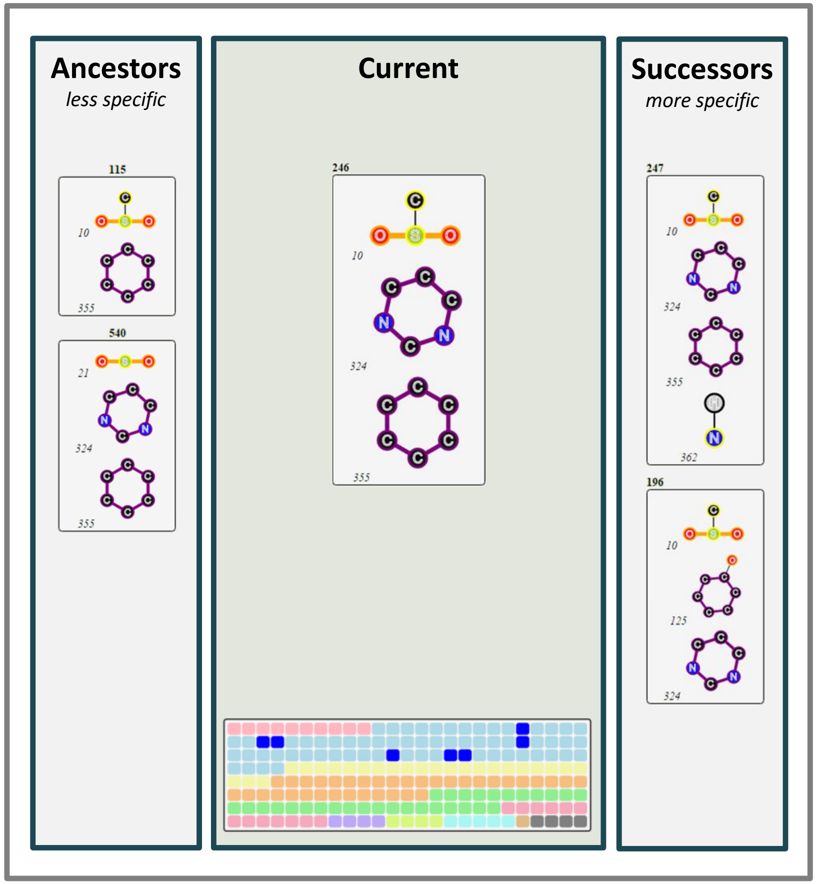
- Interactive tool dedicated to polypharmacology of kinases
- Example: dual inhibition of ERBB2 and EGFR



ρ On 5

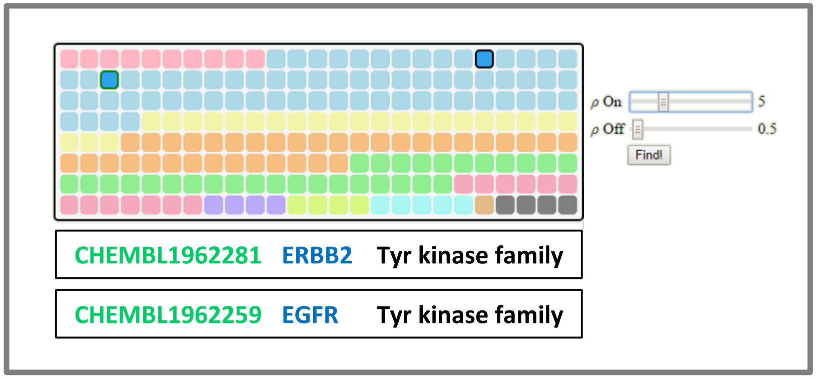
ρ Off 0.5

CHEMBL1962281	ERBB2	Tyr kinase family
CHEMBL1962259	EGFR	Tyr kinase family

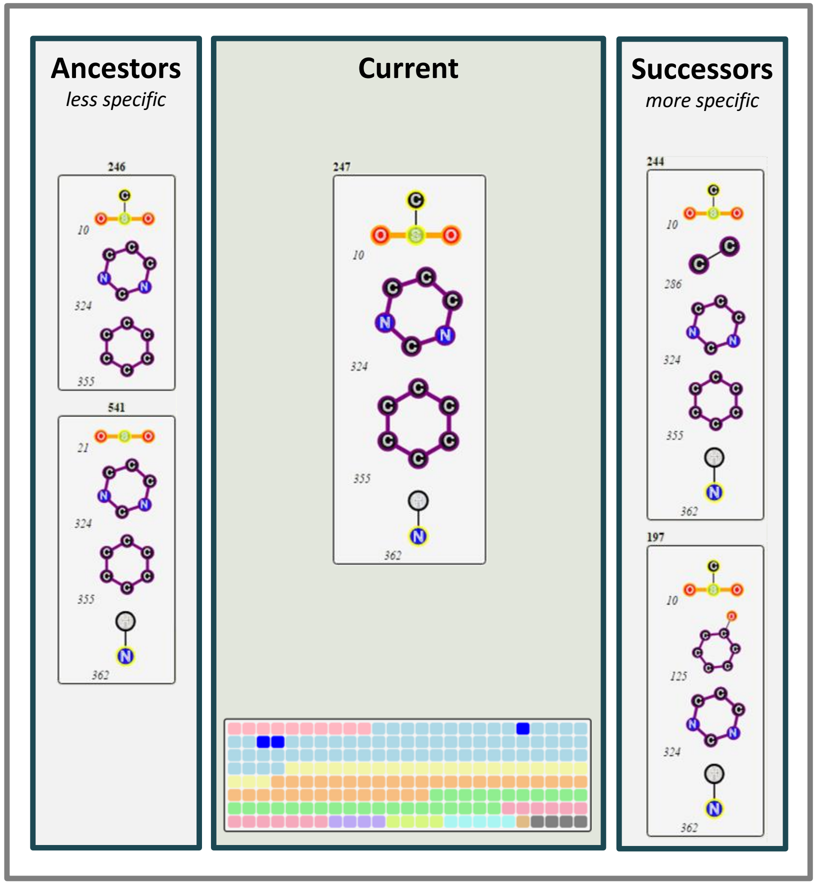


Case study (II): Polypharmacology

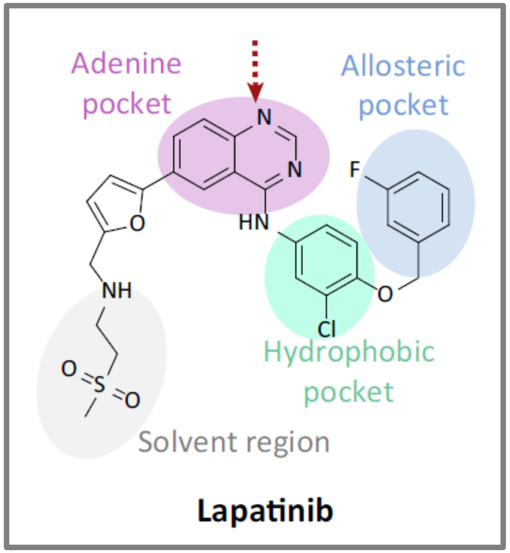
- Kinase Miner
  - Interactive tool dedicated to polypharmacology of kinases
  - Example: dual inhibition of ERBB2 and EGFR



CHEMBL1962281	ERBB2	Tyr kinase family
CHEMBL1962259	EGFR	Tyr kinase family

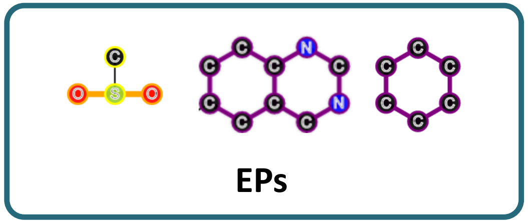


Last development



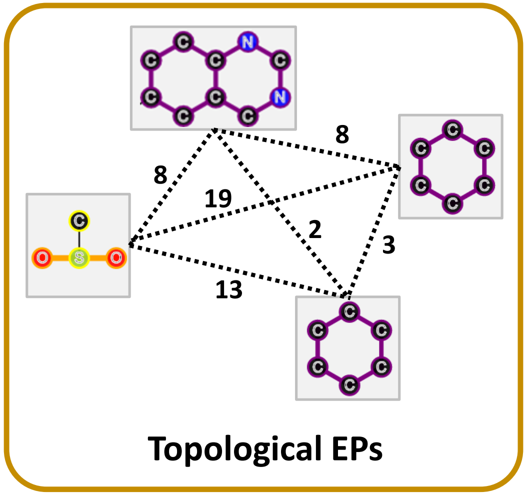
Lapatinib

1 Initial method



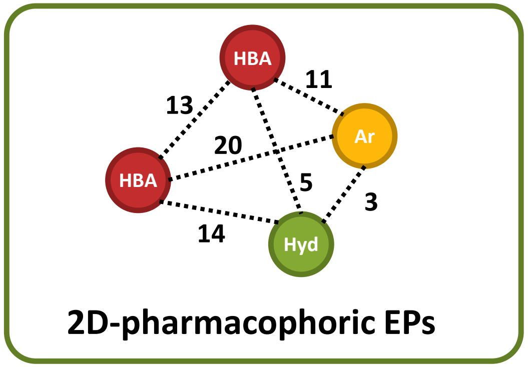
EPs

2 Addition of the distances



Topological EPs

3 Pharmacophoric definitions



2D-pharmacophoric EPs

Conclusion and perspectives

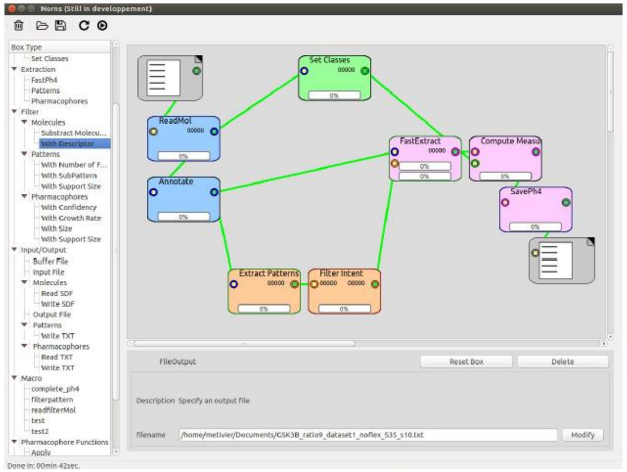
- The aim of the EP mining described here is to support the knowledge discovery in large and multidimensional data sets

- Validation
  - Identification of toxicophores
  - Understanding of the polypharmacological profile of kinase inhibitors

- Development of a workflow tool

Poster P17

Métivier et al.
Automated Generation of 2D-Pharmacophores from Large Datasets

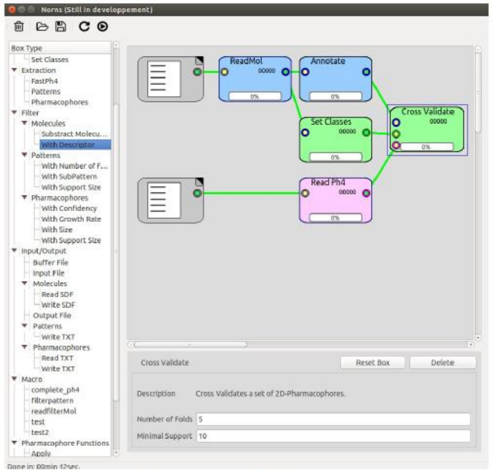


FileOutput

Description Specify an output file

filename /home/metivier/Documents/GSK3B_ratio9_dataset1_noflex_S35_s10.txt

2D-Pharmacophore Extraction



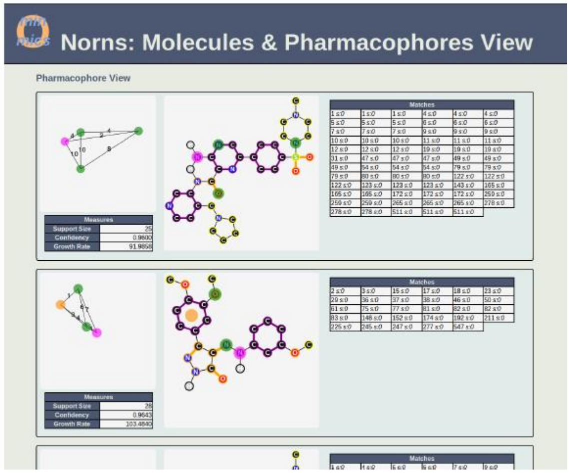
Cross Validate

Description Cross Validates a set of 2D-Pharmacophores.

Number of Folds 5

Minimal Support 10

Cross-validation



Measures	
Support Size	25
Confidency	0.9600
Growth Rate	91.9856

Measures	
Support Size	28
Confidency	0.9643
Growth Rate	103.4840

Results as webpages

Acknowledgments

GREYC

Bertrand Cuissart
Bruno Crémilleux

CERMN

Ronan Bureau
Jean-Philippe Métivier

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Aleksey Buzmakov
Amedeo Napoli

QUIID

Guillaume Poezevara
www.quiid.tech - contact@quiid.tech

Higher School of Economics

Sergei Kuznetsov

adn'TOX
EXPERT EN GENOTOXICOLOGIE

Jérémie Le Goff

Financial supports

MinOmics program

Basse Normandie
CONSEIL RÉGIONAL

RÉGION
NORMANDIE

NORMANDIE
INCUBATION

Agence Nationale de la Recherche
ANR