Statistical problems from Syngenta

Ranking of chemicals

Response: Kill score (0-100) for each chemical from experiments.

Explanatory:

- Plant species
- Application timing
- Application rate

Problem: Rank the chemicals.

What is meant by chem A is better than chem B?

Current approaches:

- Non-linear regression, rank by ED50. Each chem modelled separately.
- Elo/Fifa ranking. Not order invariant (depends on screening order).

Data from experiments

Fix screen and species

	Chem A	Chem B	Chem C	Chem D
100g/ha	0	20	20	20
500g/ha	60	80	40	60
1000g/ha	80	100	80	80

Another screen and species

	Chem A	Chem B	Chem C	Chem D
100g/ha	10	10	40	??
500g/ha	50	60	50	??
1000g/ha	85	90	100	??

From chem space to chem ranking

Response: Kill score (0-100)

Explanatory:

- Chem composition
- Everything in previous slide.

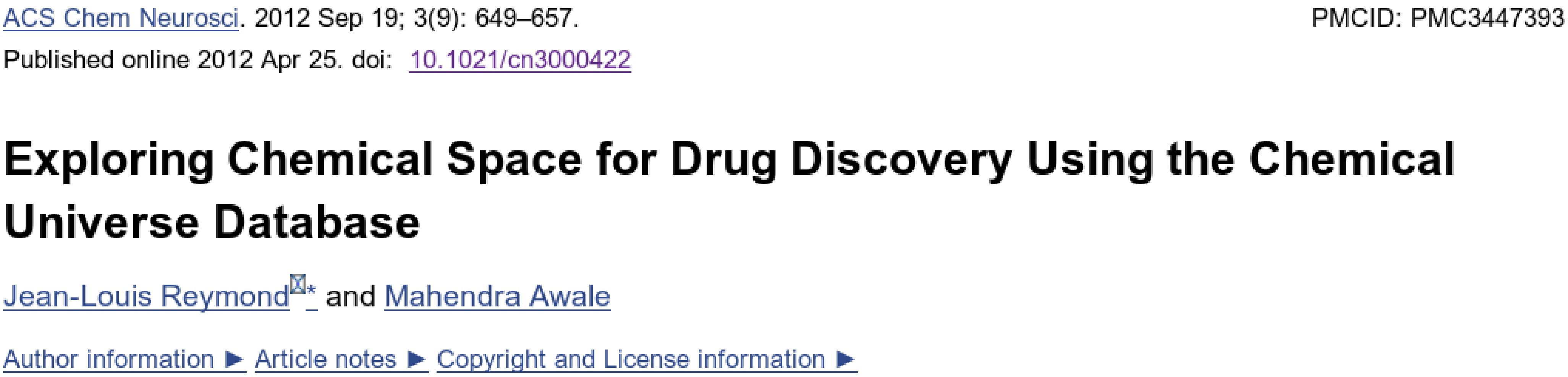
Question: Predict kill score (y) from composition ($\underline{x}$)

$$y = f(\underline{x}) + \varepsilon$$

Non-smooth fcn

A small change in the composition may have a big impact on kill score.

Chem Space



ACS Chem Neurosci. 2012 Sep 19; 3(9): 649–657. PMCID: PMC3447393

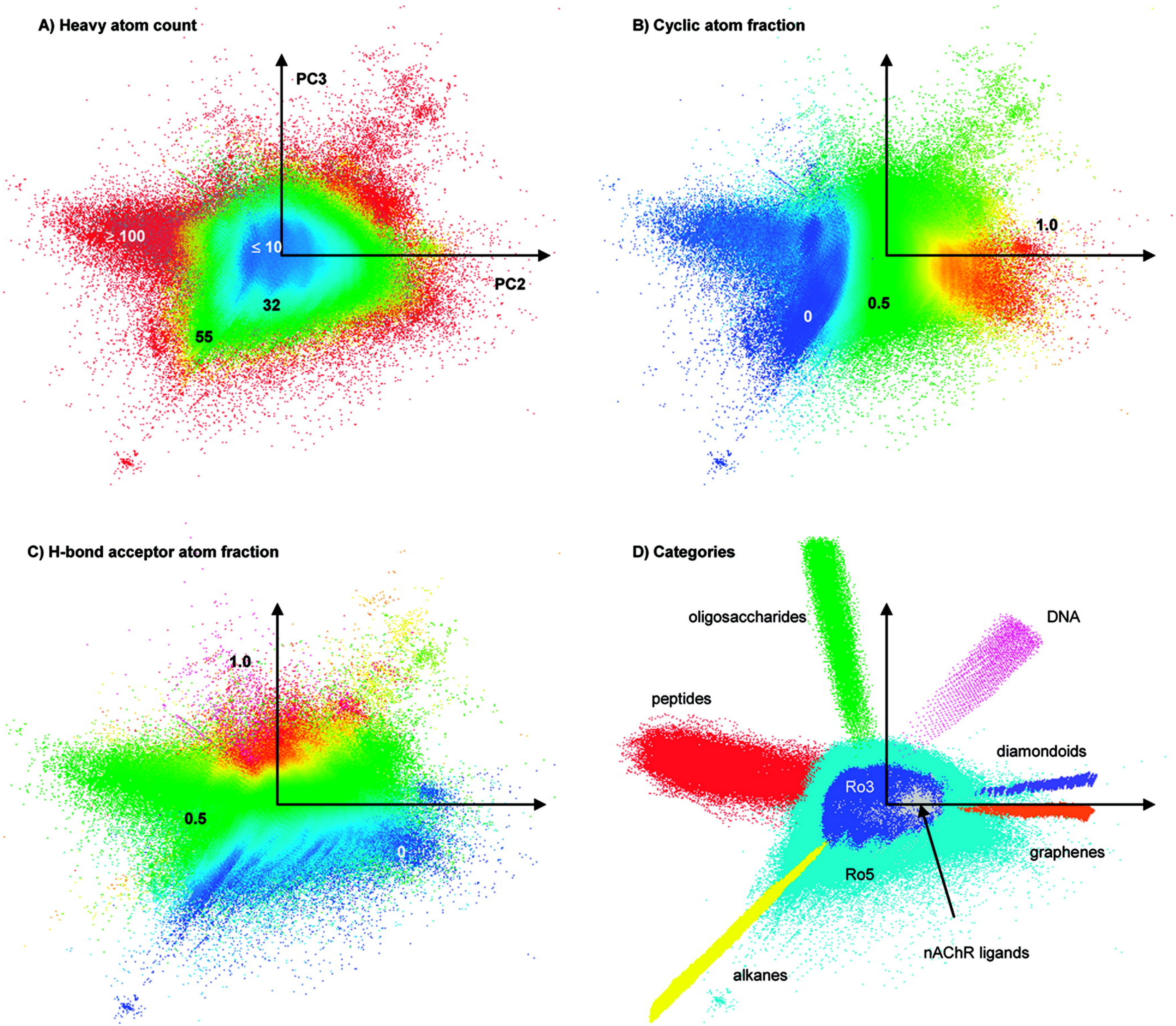
Published online 2012 Apr 25. doi: 10.1021/cn3000422

Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database

Jean-Louis Reymond☒* and Mahendra Awale

Author information ► Article notes ► Copyright and License information ►

- ~ 1B sturctures (2.6GB) www.gdb.unibe.ch
- Molecular quantum numbers (MQNs): 42 integer-valued descriptors
    - Atom counts
    - Polarity counts
    - Bond counts
    - Topology counts


A) Heavy atom count
PC3
PC2
≥ 100
≤ 10
32
55
B) Cyclic atom fraction
1.0
0.5
0
C) H-bond acceptor atom fraction
1.0
0.5
0
D) Categories
oligosaccharides
DNA
peptides
diamondoids
Ro3
graphenes
Ro5
nAChR ligands
alkanes

Formulation toxicity

Response: Compound toxicity (0-100 or categories).

Explanatory: Chem composition.

Problem: Estimate individual toxicities. Interactions?

$$y = f(x,\theta) + \varepsilon$$

This sounds like a standard regression problem.