Karolinska Institutet

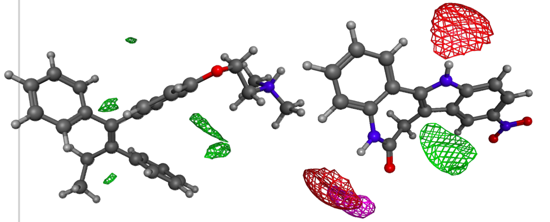
Is our QSAR modeling of growth inhibition any good?

Egon Willighagen

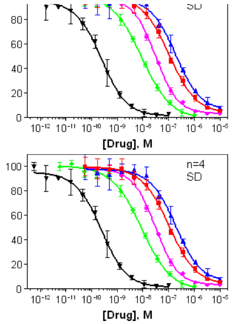
2011-05-30

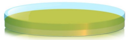
Goal: predict growth inhibition from molecular structures

Molecular descriptors



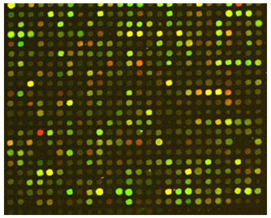
Drug responses (GI50)





Regression

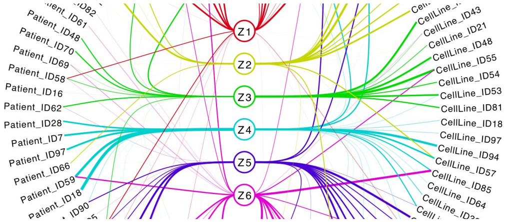
Gene expression



Genesets

Regression

Probabilistic component model



Interpretation

- Downstream biological profiling
- Toxicity prediction
- Biomarker discovery

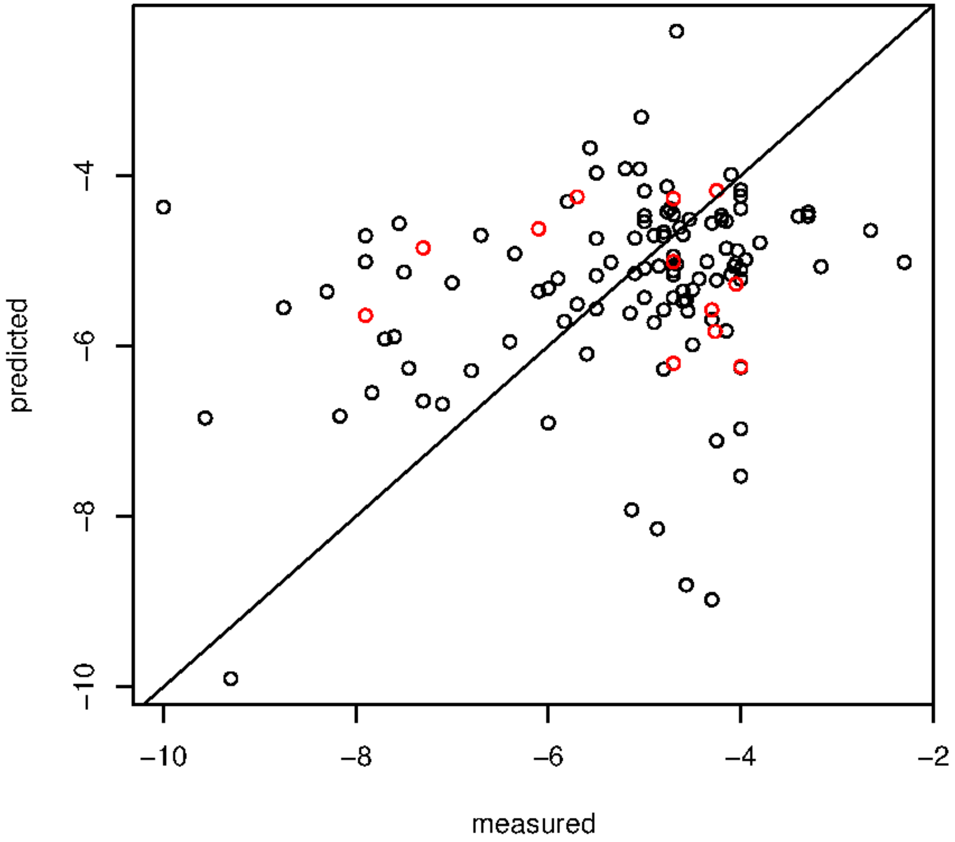
Data and Goal

- **Compounds**
  - 230 drugs → molecular descriptors
    - logP, number of acidic groups, etc

- **Activities**
  - log GI_{50} values (between -7 and -2), where GI_{50} is in molar
    - GI_{50} is the dose where the growth is inhibited 50%
  - Three cell lines: HL60, PC3, MCF7

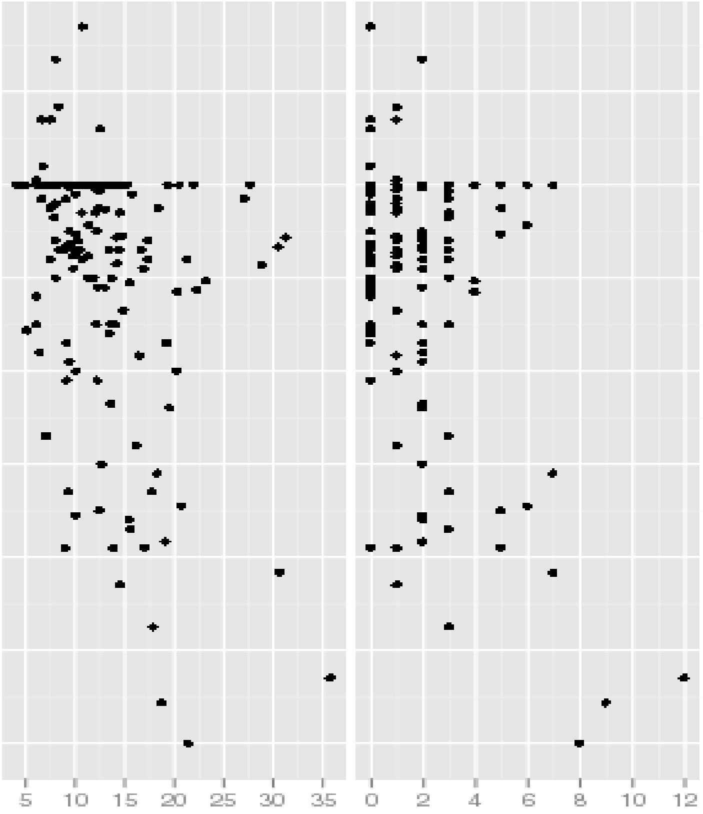
- Understand why some molecules have higher GI_{50} values

GI_{50} cannot be predicted from molecular structures



(Non-)linear regression methods cannot predict the GI_{50} values from the molecular structures using ~290 QSAR descriptors, like logP, number of hydrogen donors, etc.

The “best” descriptors...



Skeletal variation

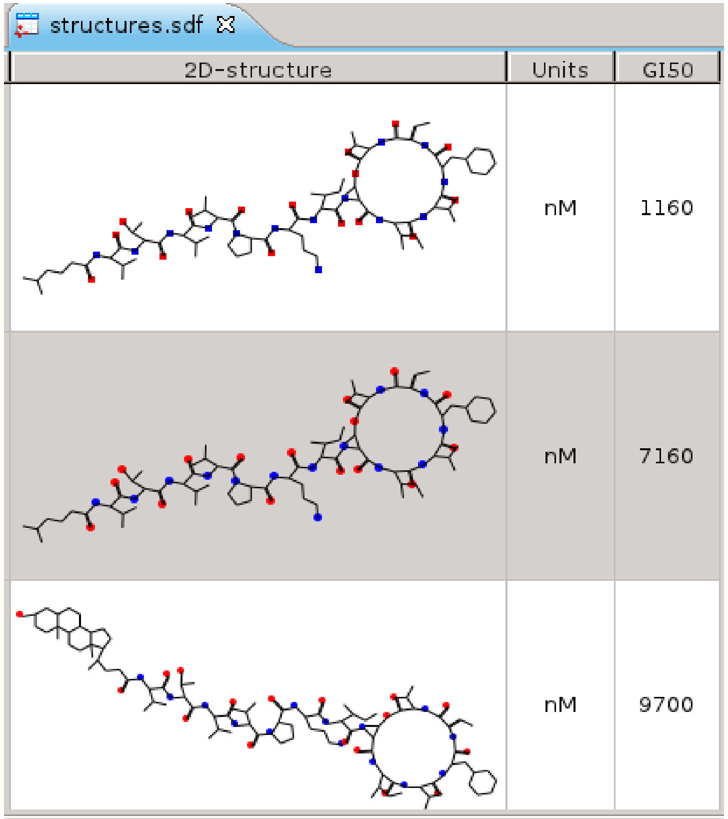
Number of double bonded carbons

Even the best QSAR descriptors show little correlation (*0.38* and *0.4*). and they do not complement.

What about published GI_{50} QSAR models?

ChEMBL database

- → > 50000 GI_{50} values from literature (GI_{50}, log GI_{50}, …)
- → Largest studies have < 120 structures
- → 7 largest study of one paper, with highly congeneric *kahalalide F* compounds (see **screenshot**)
- → Next study has only 78 structures



2D-structure	Units	GI50
	nM	1160
	nM	7160
	nM	9700

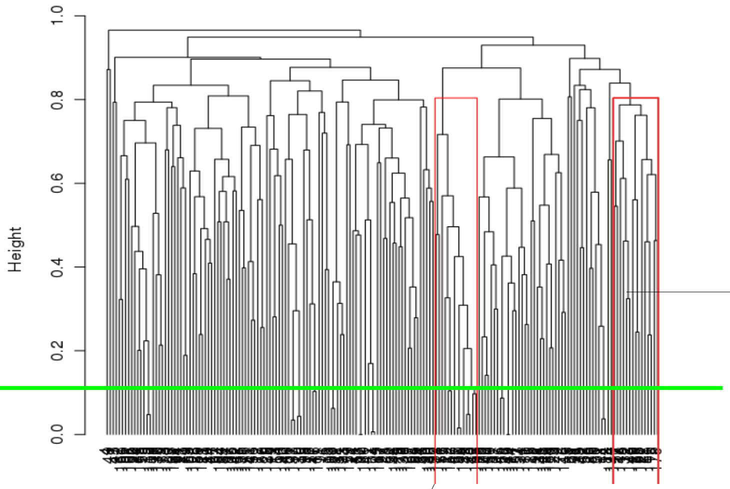
(No specs on how GI values are *counted*.)

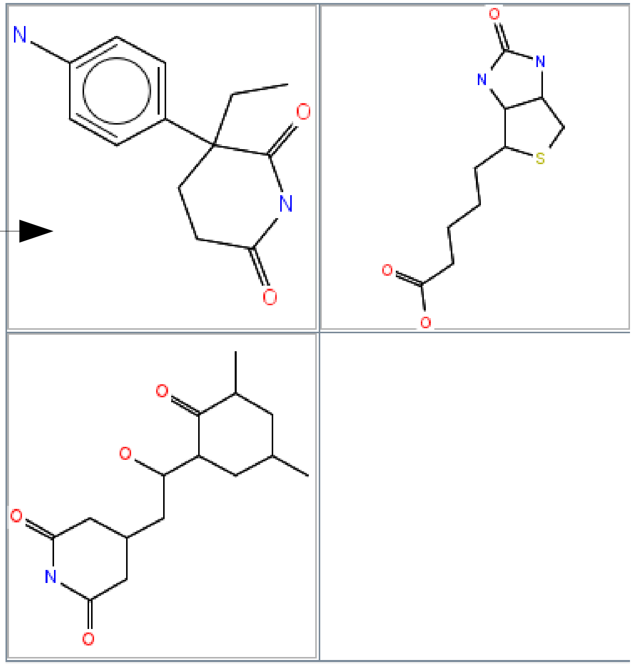
Jiménez, J.C., et al. J. Med. Chem. 2008. 51(16):4920-4931.

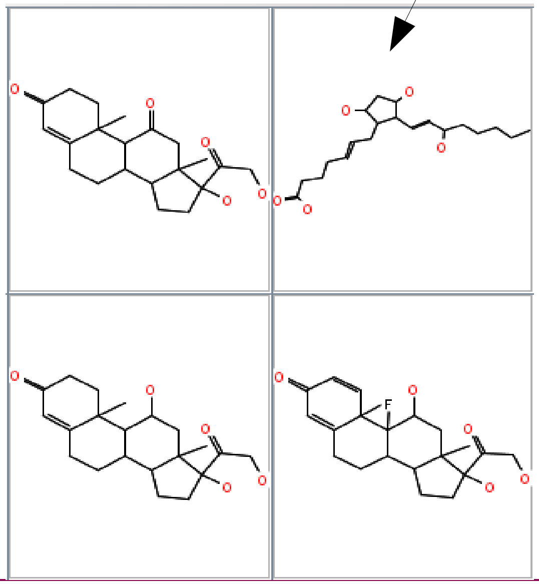
Hierarchical clustering of compounds

Dissimilar compounds

Similar compounds







The ***high diversity*** of molecules makes it harder for the QSAR modeling to find patterns.

Conclusions

- Our data is more complex than QSAR studies for GI50 in literature
  - → Clusters too diverse, too many modes of action(?)
  - → Statistical method cannot find any significant patterns
  - → Correlation found is hard to interpret (at best)
- Bad for our paper? No.
  - → For data sets with high diversity we propose gene expression as alternative