Bayesian joint modeling of chemical structure and dose response curves

Kelly Moran

Duke University, Department of Statistical Science

kelly.r.moran@duke.edu

CSGF Outgoing Fellow Talk (wow!)

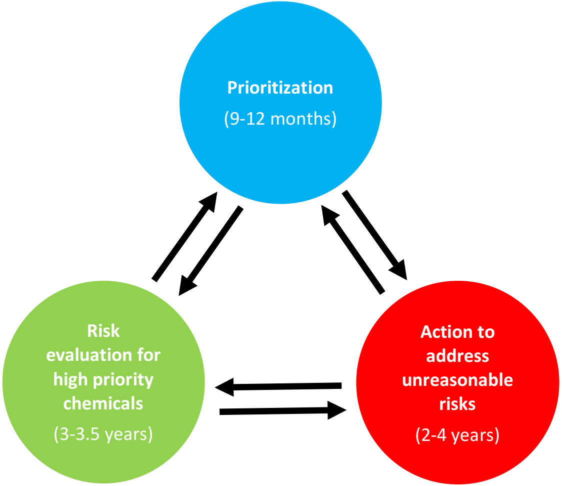
Motivation

Image from https://deadhomersociety.com/2016/06/05/quote-of-the-day-2632/.

Motivation

*From http://gender-chemicals.org/blog.

The Toxic Substances Control Act (TSCA)



Prioritization

(9-12 months)

Risk evaluation for high priority chemicals

(3-3.5 years)

Action to address unreasonable risks

(2-4 years)

*From `https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/how-epa-assesses-chemical-safety`.

Chemical testing

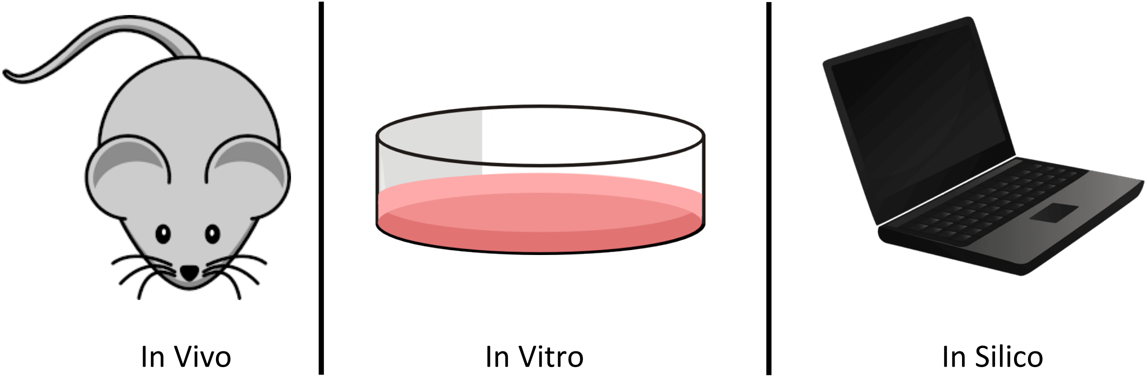
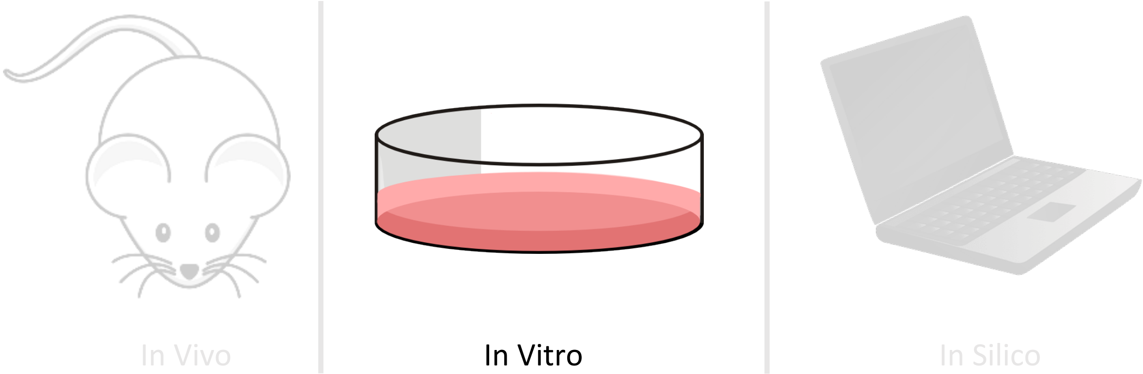


Figure: Means of chemical testing, from slow and expensive to fast and cheap.

Chemical testing



In Vivo

In Vitro

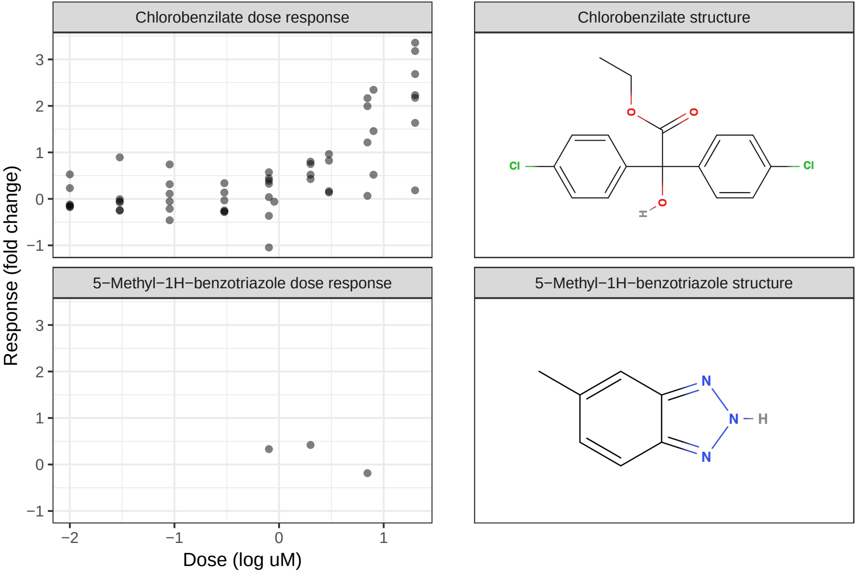
In Silico

Figure: Fast(er) and cheap(er).

ToxCast: EPA's high-throughput screening program

Figure: ToxCast has data on over 9,000 chemicals with over 1,000 assay endpoints. (Left) High-throughput assay plate is filled. (Right) High-throughput screening robot.

Data in ToxCast



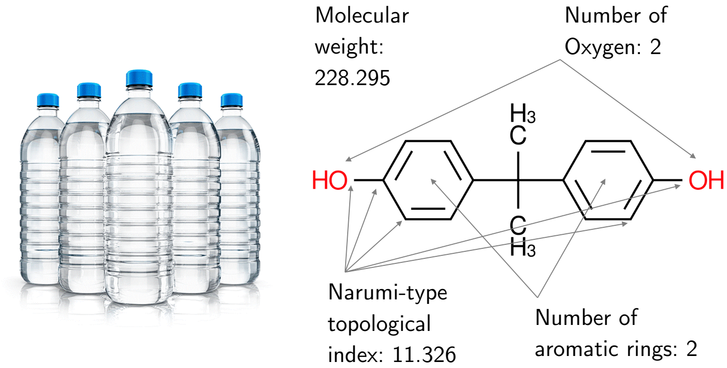
Chemical testing

Figure: Fast(est) and cheap(est).

What's the use?

- Learning about toxicologically relevant chemical distance in silico helps in:
  - Designing new studies.
  - Increasing efficiency of studies.
  - Supplementing the results from lab-based studies.
  - Bridging the gap between the # of chemicals of interest and the # with known toxicological profiles.

Chemical structure (BPA)



Molecular weight: 228.295

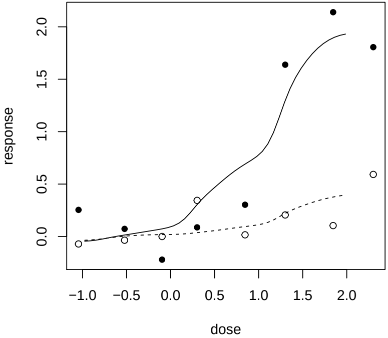
Number of Oxygen: 2

Narumi-type topological index: 11.326

Number of aromatic rings: 2

Figure: Software such as Mold2 extract chemical features using SMILES. The SMILES for Bisphenol A (BPA) is CC(C)(C1=CC=C(C=C1)O) C2=CC=C(C=C2)O.

“Near” in structure but “far” in activity



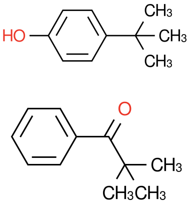
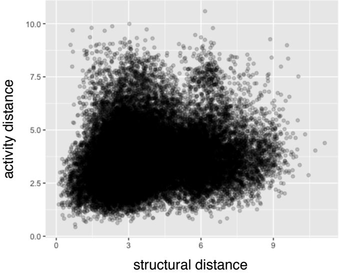
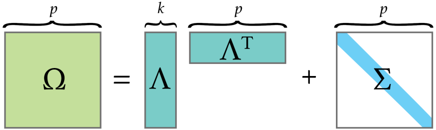


Figure: 4-tert-Butylphenol (left solid/solid, right top) and tert-Butyl phenyl ketone (left open/dashed, right bottom).

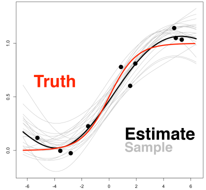


Key model components

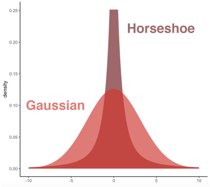
- Factor modeling



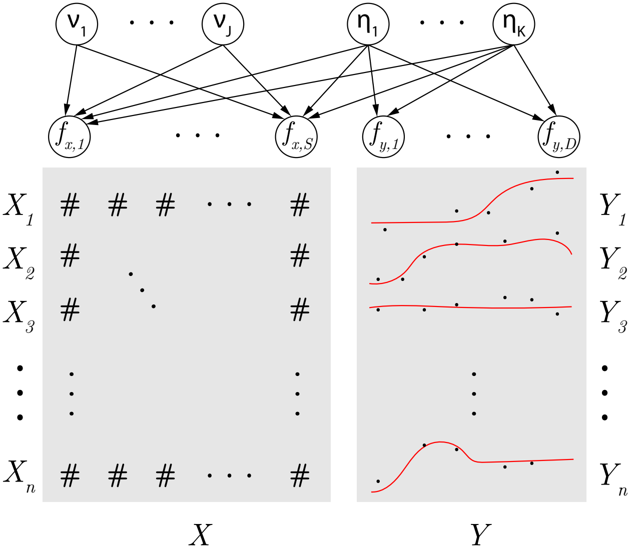
- Gaussian processes



- Sparsity-inducing priors



Model visual summary



Chemical “distance” and prediction

- In mathematical notation

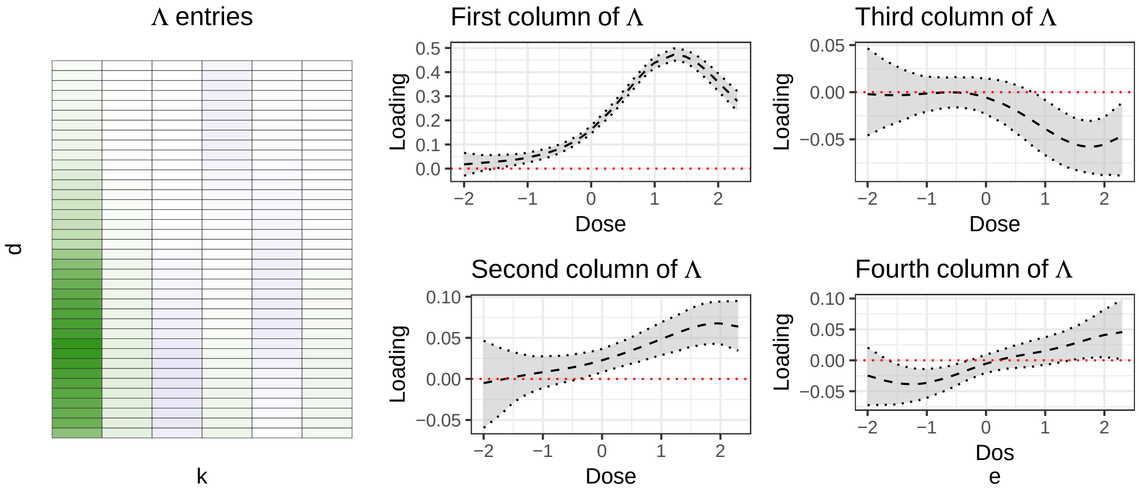
$$\underset{D\times 1}{Y_i} = \underset{D\times K}{\Lambda}\;\underset{K\times 1}{\eta_i} + \underset{D\times 1}{\varepsilon_i}\,, \quad \underset{S\times 1}{X_i} = \underset{S\times K}{\Theta}\;\underset{K\times 1}{\eta_i} + \underset{S\times J}{\Xi}\;\underset{J\times 1}{\nu_i} + \underset{S\times 1}{\mathrm{e}_i}\,.$$

- Toxicity “distance” between chemicals i and j can be represented in the shared factor space (i.e., how far apart the vectors η_i and η_j are)

- Two chemicals that are very close in this space will have similar dose-response curves, and similar toxicity-relevant features

- They may not have similar toxicity-irrelevant features

ToxCast results (Λ)



Λ entries

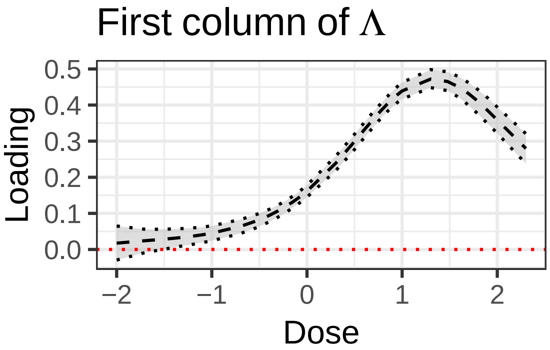
First column of Λ

Third column of Λ

Second column of Λ

Fourth column of Λ

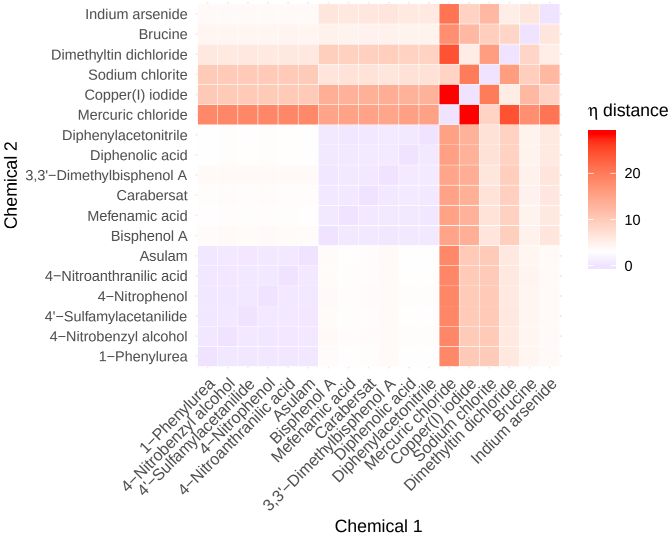
Significant features associated with first column of Λ



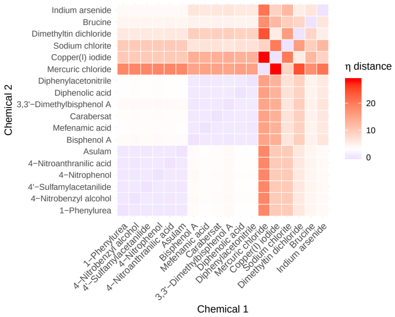
- The number of group X-C on aromatic ring
- Molecular regresson coefficients surface LogP index
- Sum eigenvalue weighted by van der Waals distance matrix
- Sum of topological distance between the vertices O and Cl
- Number of Chlorine

In the training set, the chemicals having the largest expected value for η_1 are Mercuric chloride, Benzyltriphenylphosphonium chloride, Sodium chlorite, 1,1-Bis(3-cyclohexyl-4-hydroxyphenyl)cyclohexane, and Basic Blue 7. All but 1,1-Bis(3-cyclohexyl-4-hydroxyphenyl)cyclohexane, which is a known irritant, are known toxins.

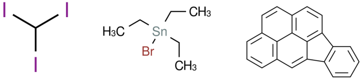
Chemical distance



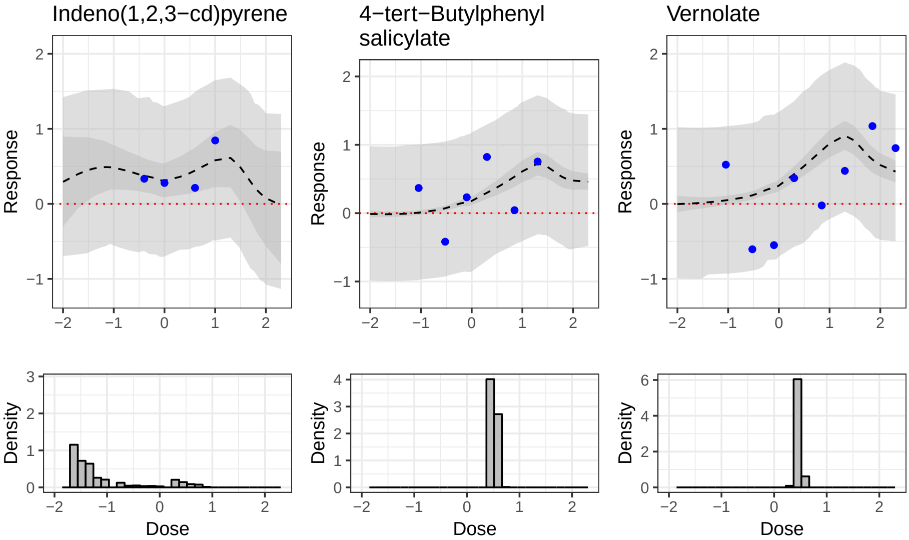
“Fill in” or “venture out”



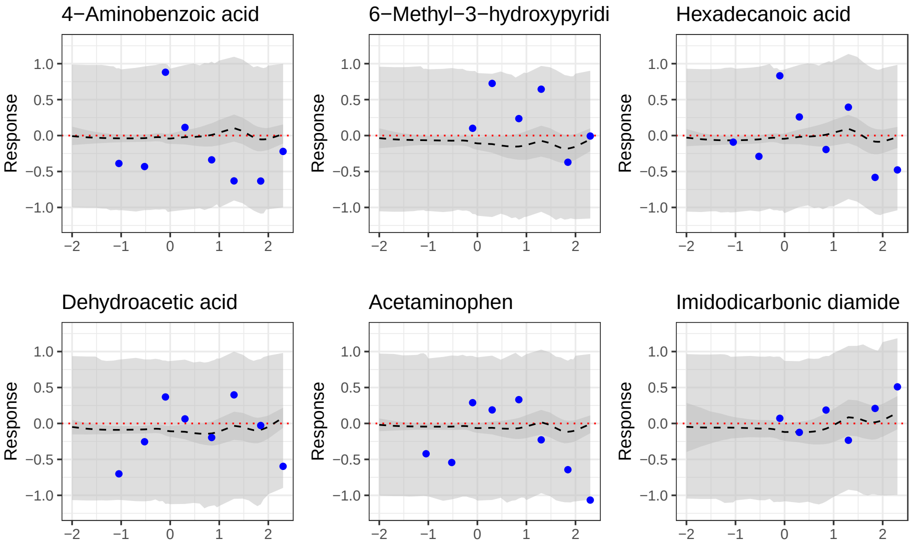
The three “farthest” chemicals in the hold-out set. From left to right: Iodoform, Triethyltin bromide, and Indeno(1,2,3-cd)pyrene.



Predictions for hold-out activating chemicals



Predictions for hold-out non-activating chemicals

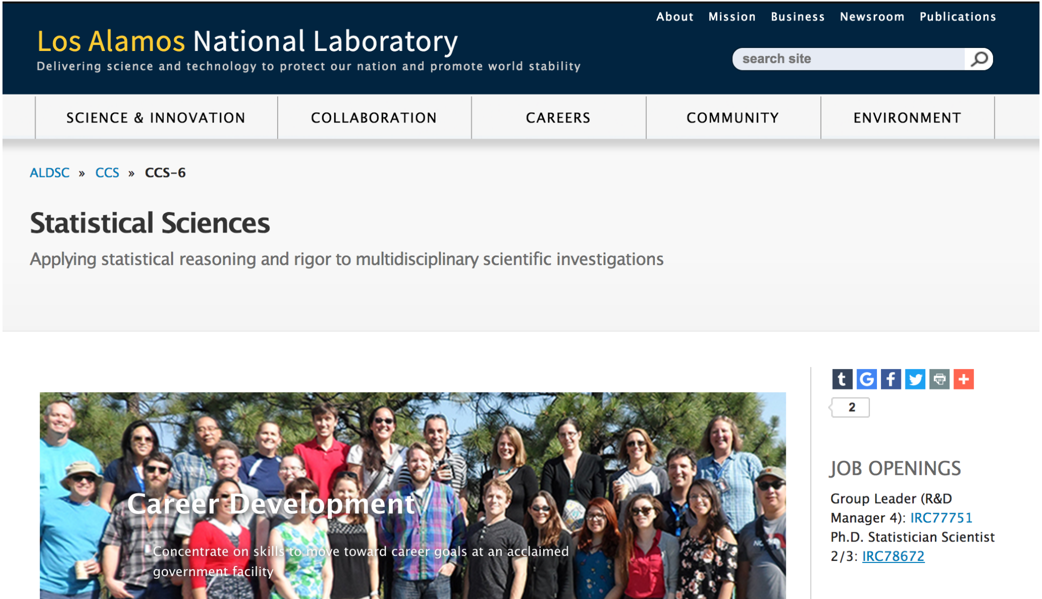


What next (for this model)?

Future work includes:

- Using distance to inform mixture models
- Direct model specification of active/inactive
- Nonlinear dimension reduction
- Linking (combinations of) assays to human health outcomes
- Integrating information from multiple assays and multiple feature sets

What next (for me)?



Los Alamos National Laboratory

Delivering science and technology to protect our nation and promote world stability

ALDSC » CCS » CCS-6

Statistical Sciences

Applying statistical reasoning and rigor to multidisciplinary scientific investigations

Career Development

Concentrate on skills to move toward career goals at an acclaimed government facility

JOB OPENINGS

Group Leader (R&D Manager 4): IRC77751

Ph.D. Statistician Scientist 2/3: IRC78672

Computing and CSGF

CSGF gave me the freedom to explore beyond just my advisor's projects:

- First experience with Gaussian processes was a LANL practicum my first summer in grad school
- Through work on this toxicology project, I met a wonderful collaborator who was interested in GPs
- Developed a new fast GP algorithm
- Will be able to pull that in to expand this method to bigger applications

Thanks and CSGF

DOE
CSGF

Thanks at Duke are due to my advisor, Amy Herring, and the rest of my committee. I'm also grateful to Matt Wheeler.

At LANL I am particularly grateful to Earl Lawrence, Dave Osthus, and Kary Myers. The motivation to finish my PhD came from wanting to work with such amazing people!