BAYESIALAB

7TH ANNUAL BayesiaLab Conference DURHAM 2019

NORTH CAROLINA BIOTECHNOLOGY CENTER

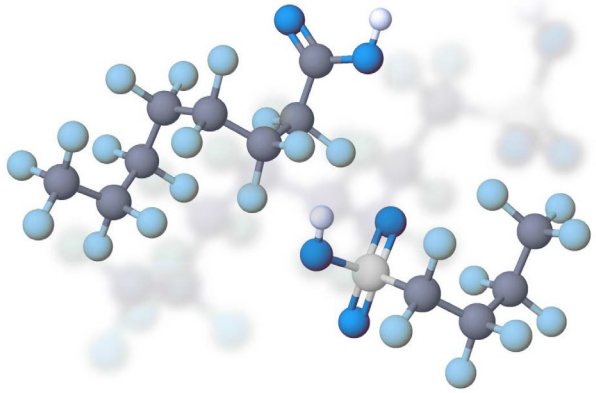
Risk Analysis of PFAS Contamination in Private Water Wells: a Bayesian Network Model

Javad Roostaei, PhD
Jacqueline MacDonald Gibson, PhD

10/10/2019

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Outline



- **Introduction**
  - PFAS, forever chemicals
  - PFAS in NC private wells
  - Objectives
- **Methods**
  - Bayesian network model development
  - Model evaluation and cross-validation
- **Results**
  - Insights on risk factors
  - Bayesian network model accuracy
- **Conclusion**
  - Intervention to decrease risk in private wells

Introduction

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Per-and polyfluoroalkyl substances (PFAS)

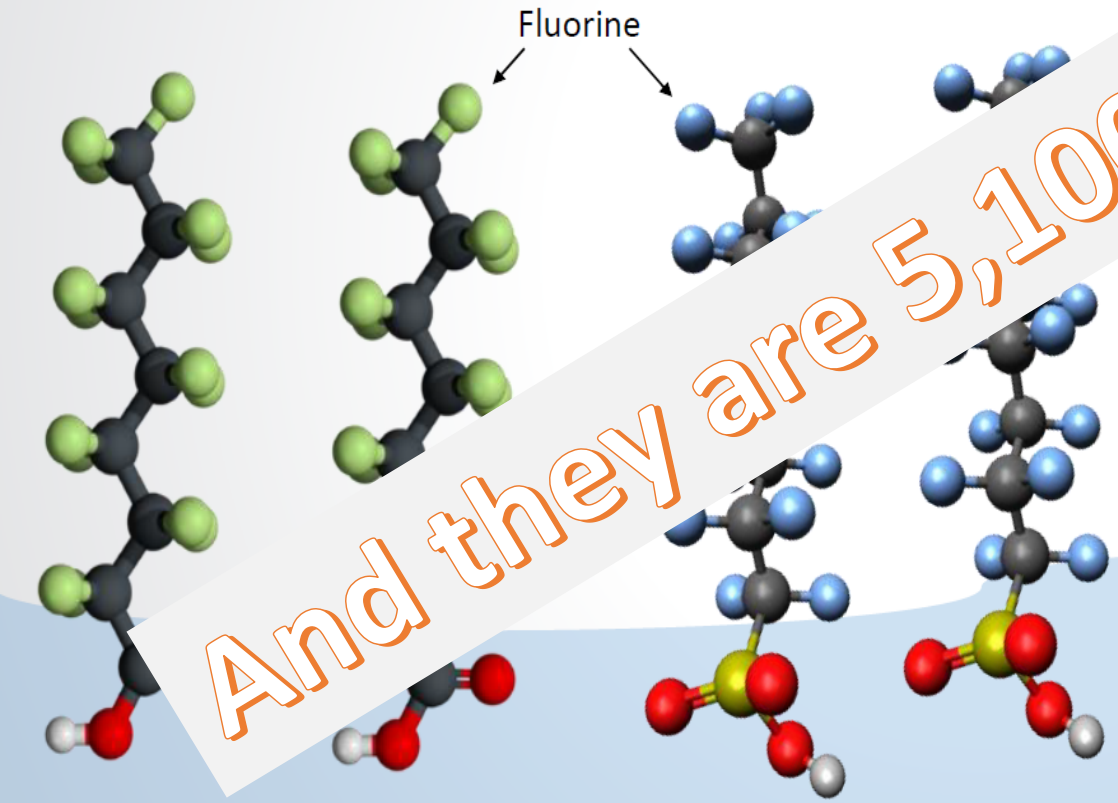
PFAS are found in many consumer products

http://www.sixclasses.org/

Forever chemicals!!!

EPA

Per- and Polyfluoroalkyl Substa… (PFAS)



Perfluorooctanoic acid (PFOA)

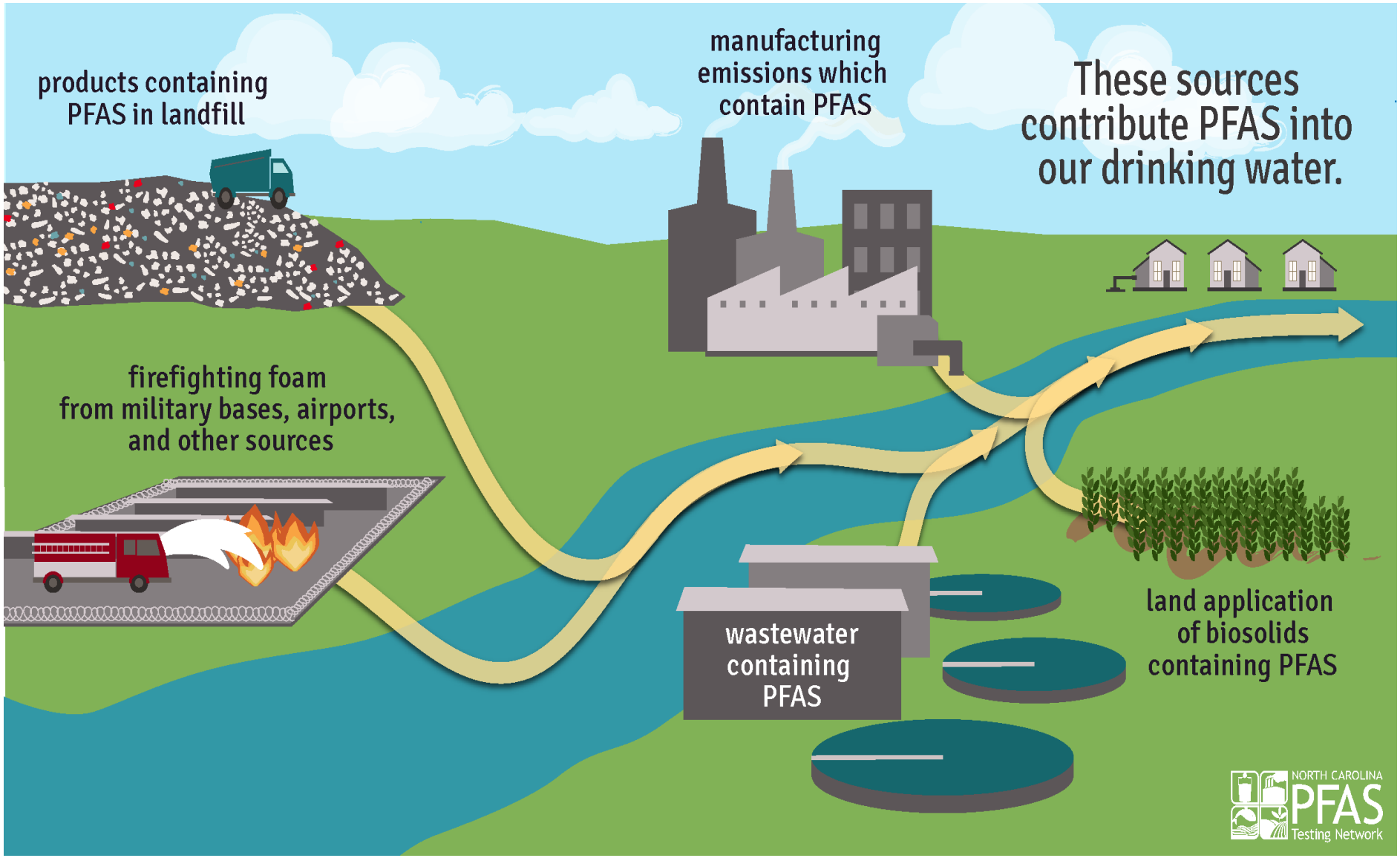
Perfluorooctanesulfonic acid (PFOS)

A clas…

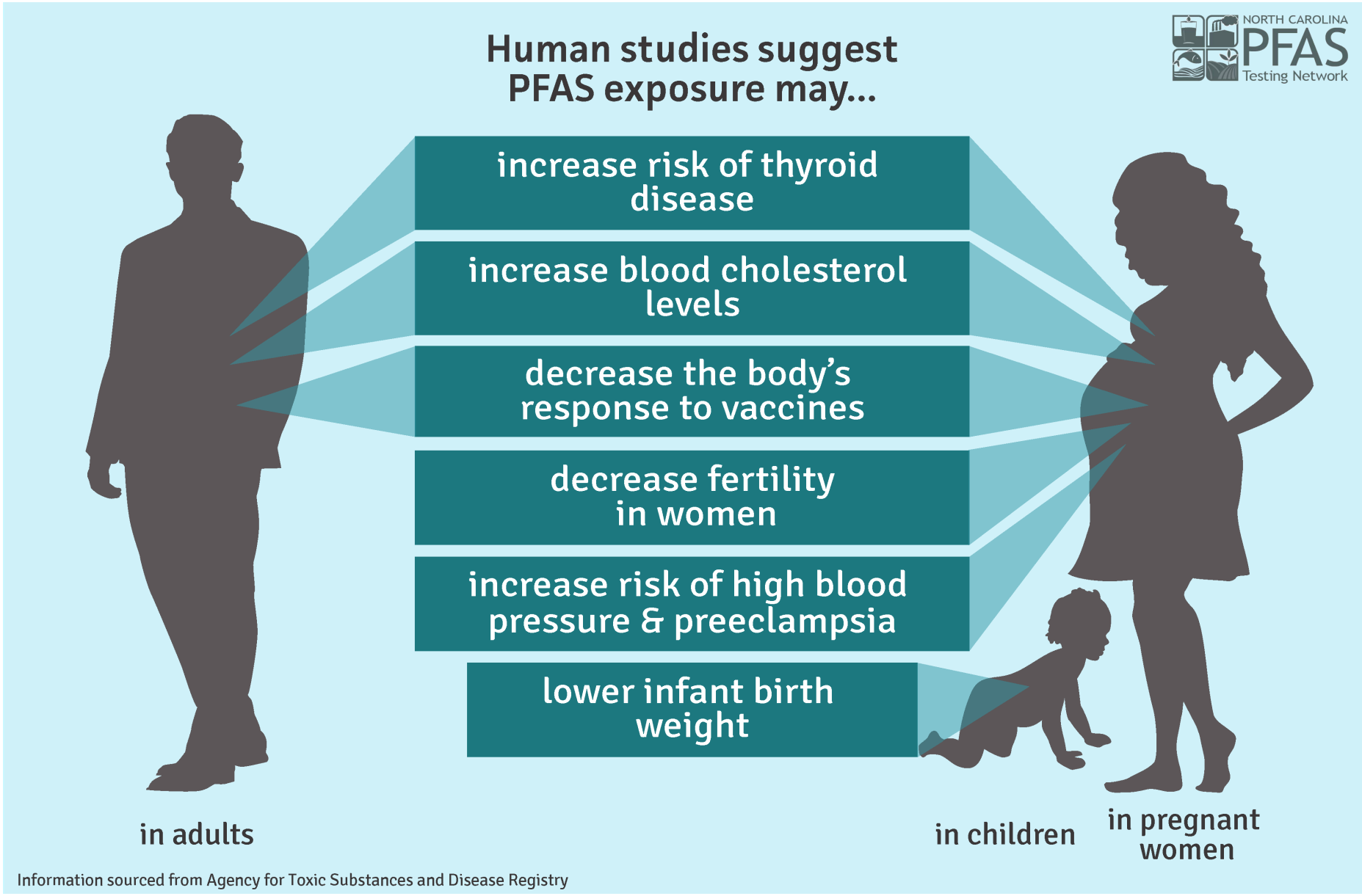
- …arbon (C) atoms …unded by fluorine (F) atoms
  - Water-repellent (hydrophobic body)
  - Stable C-F bond
- Some PFAS include oxygen, hydrogen, sulfur and/or nitrogen atoms, creating a polar end.

And they are 5,100 of them!!!

PFAS can reach to our drinking water!!

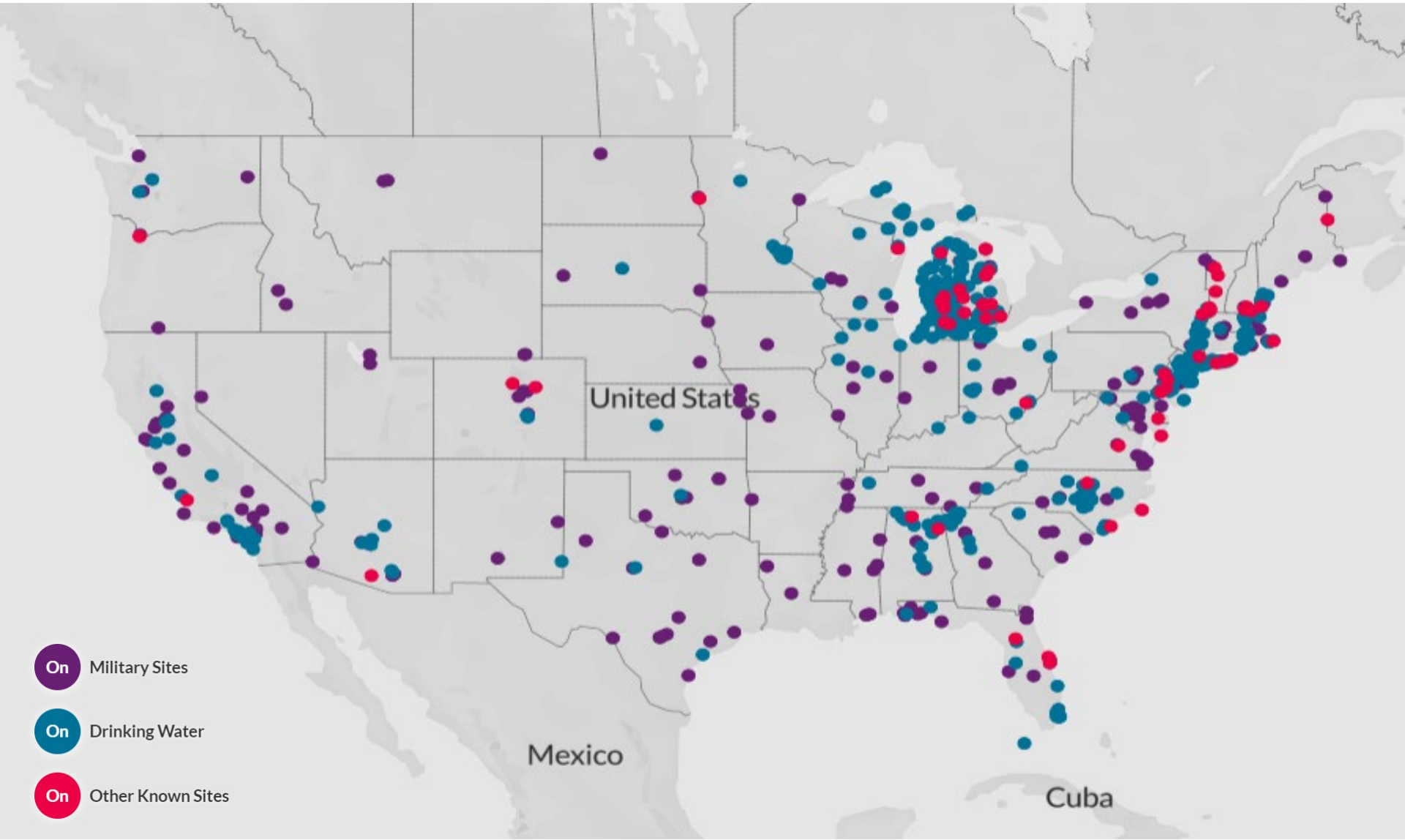


PFAS exposure and the health effects



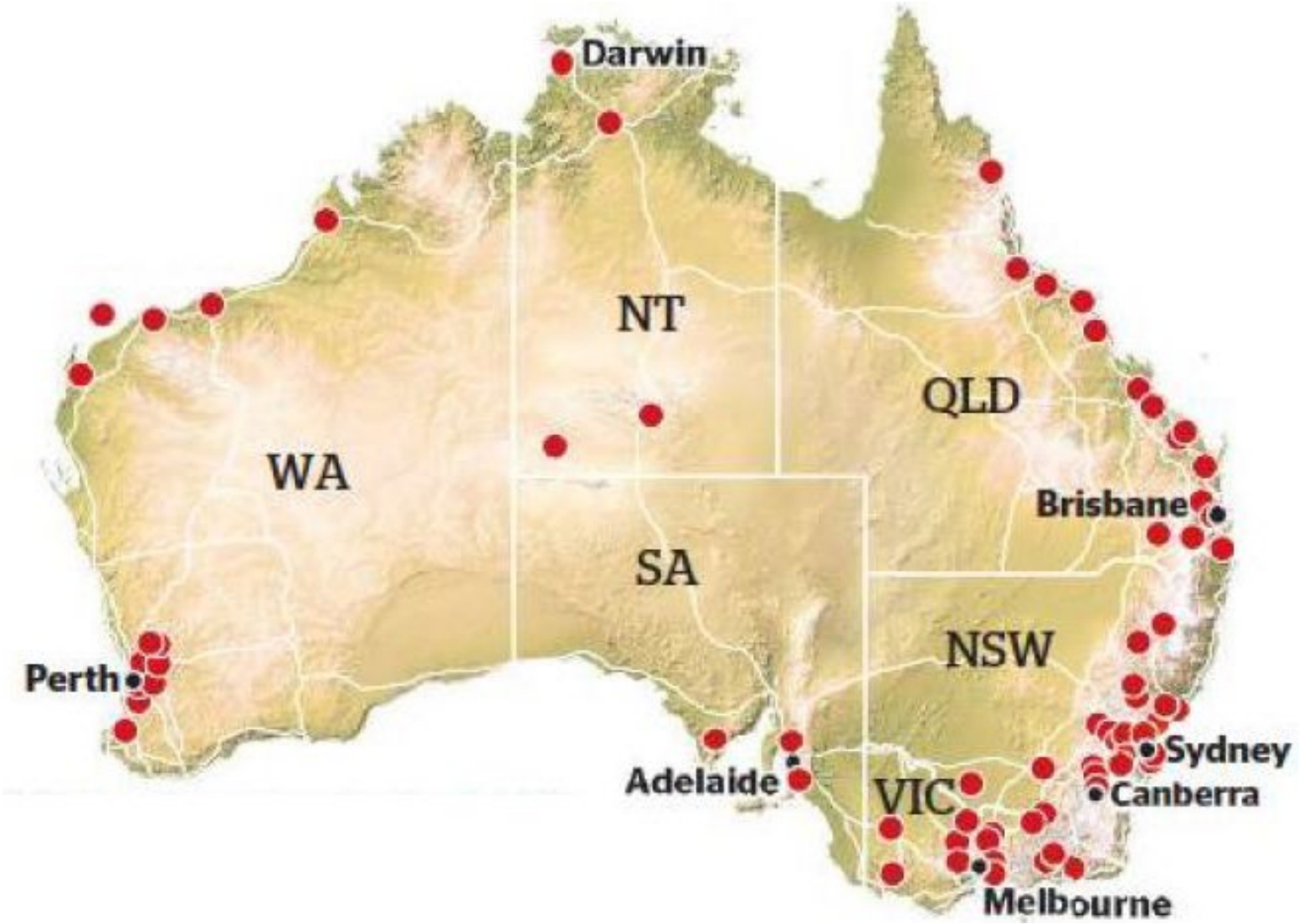
Information sourced from Agency for Toxic Substances and Disease Registry

It has been detected in many site in the US



https://www.ewg.org/

And PFAS can be found in other countries too !!!



https://www.miltoncaine.com

Private wells around the Chemours Chemical Plant are contaminated with PFAS

- During 2017-2018, PFAS (including GenX), were detected in more than 75% of 803 private water supply wells near the Chemours facility

[NC DEQ, 2017-18, Scruggs, 2019]

GenX concentrations vary widely from well to well

- 803 wells tested
  - 23% > health goal
  - 23% non-detect

- Map key
  - Red = > 140 ng/L
  - Yellow = 0- 140 ng/L
  - Green = non-detect

- Figure courtesy of DEQ

Objectives

- Overall: To uncover **factors** influencing the risk of PFAS contamination in water supply wells near the Chemours plant

  - Determine why **some wells are contaminated** and others are not. What features of the wells, landscape, geology, weather, and geographic location influence risks to wells?

  - Build model to **prioritize** private wells for future testing.

  - Develop **user-friendly web site** to help private well owners assess risks.

Methods

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Many variables may influence the PFAS showing up in a private water well

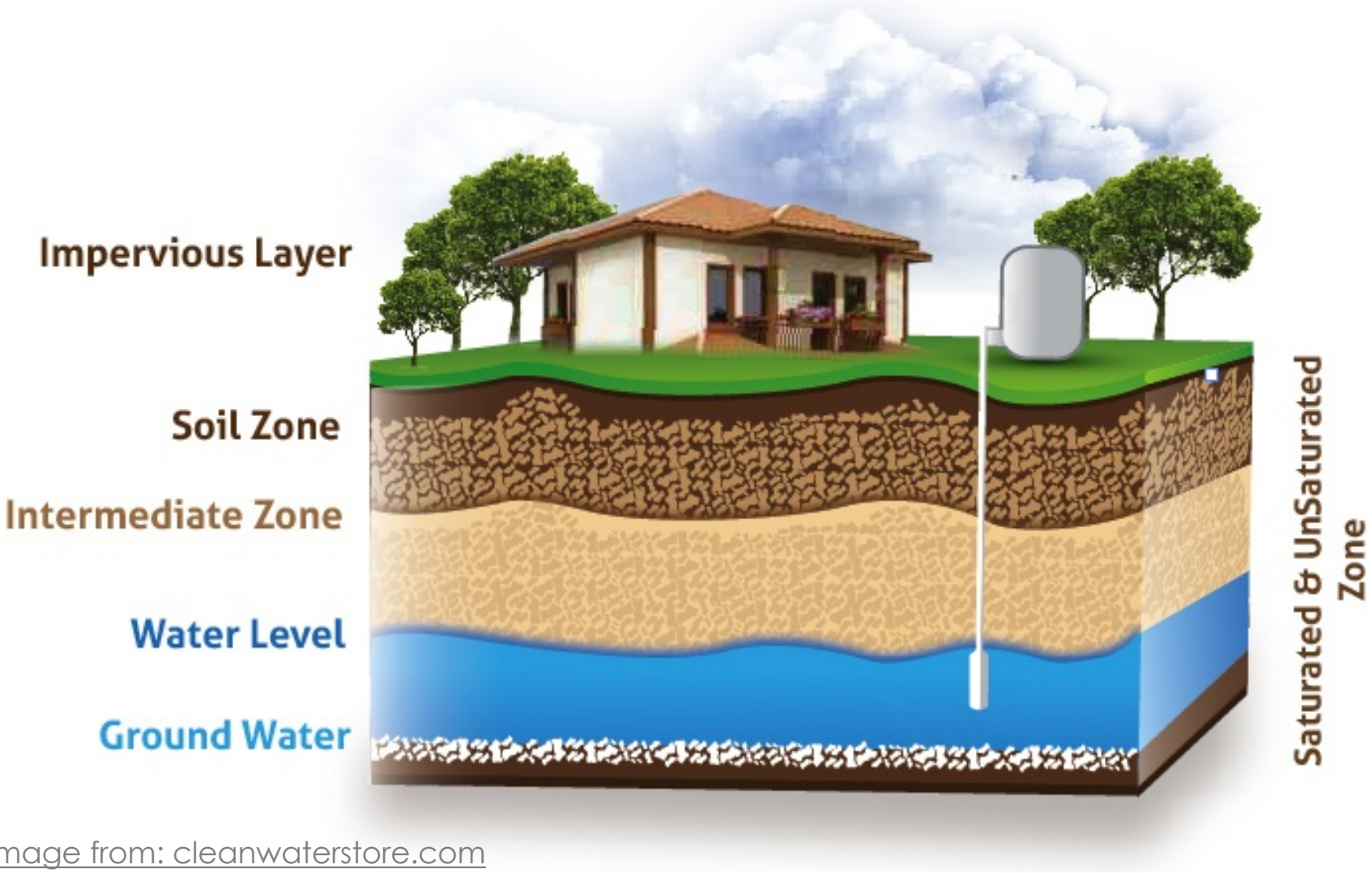
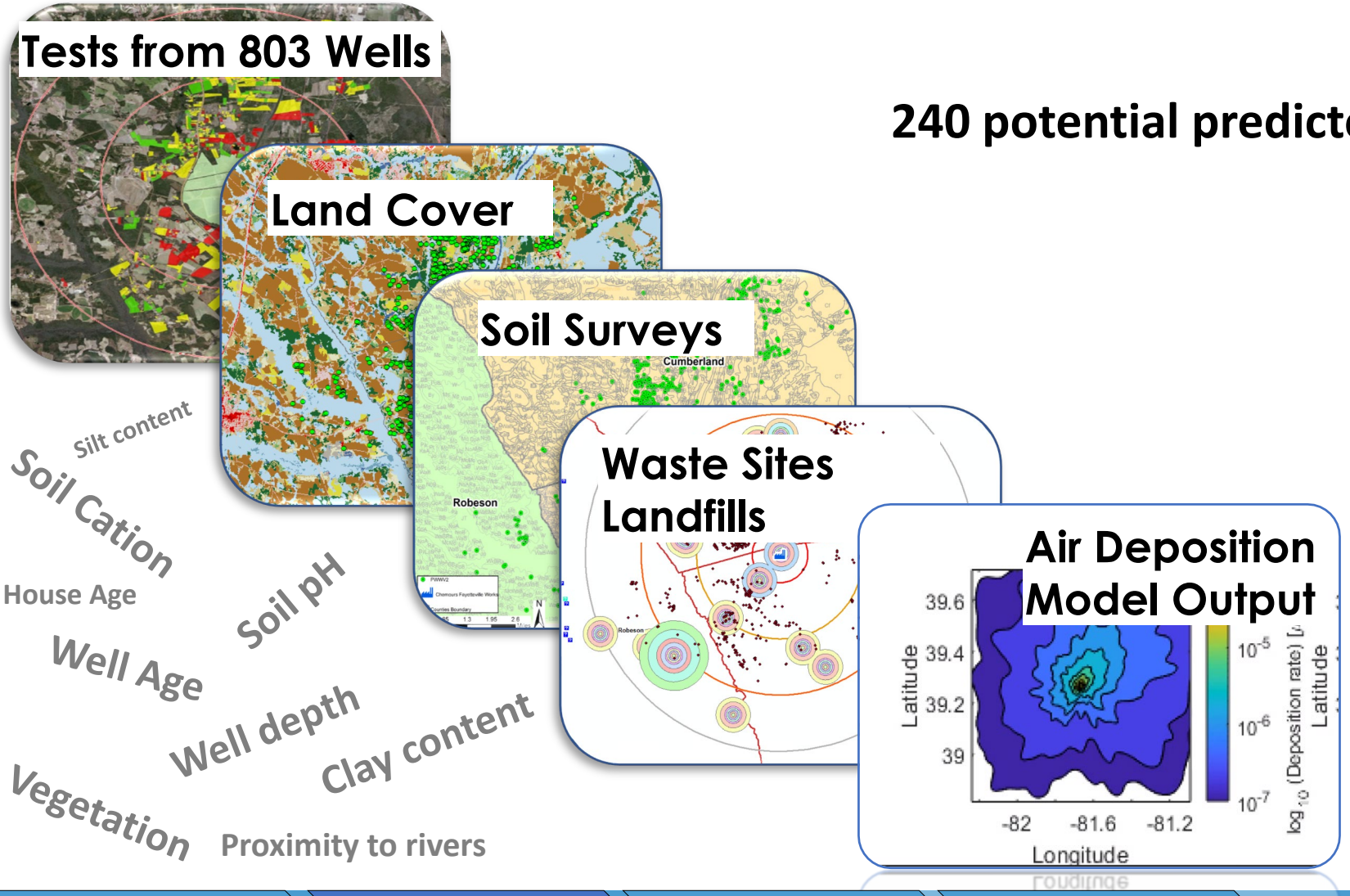


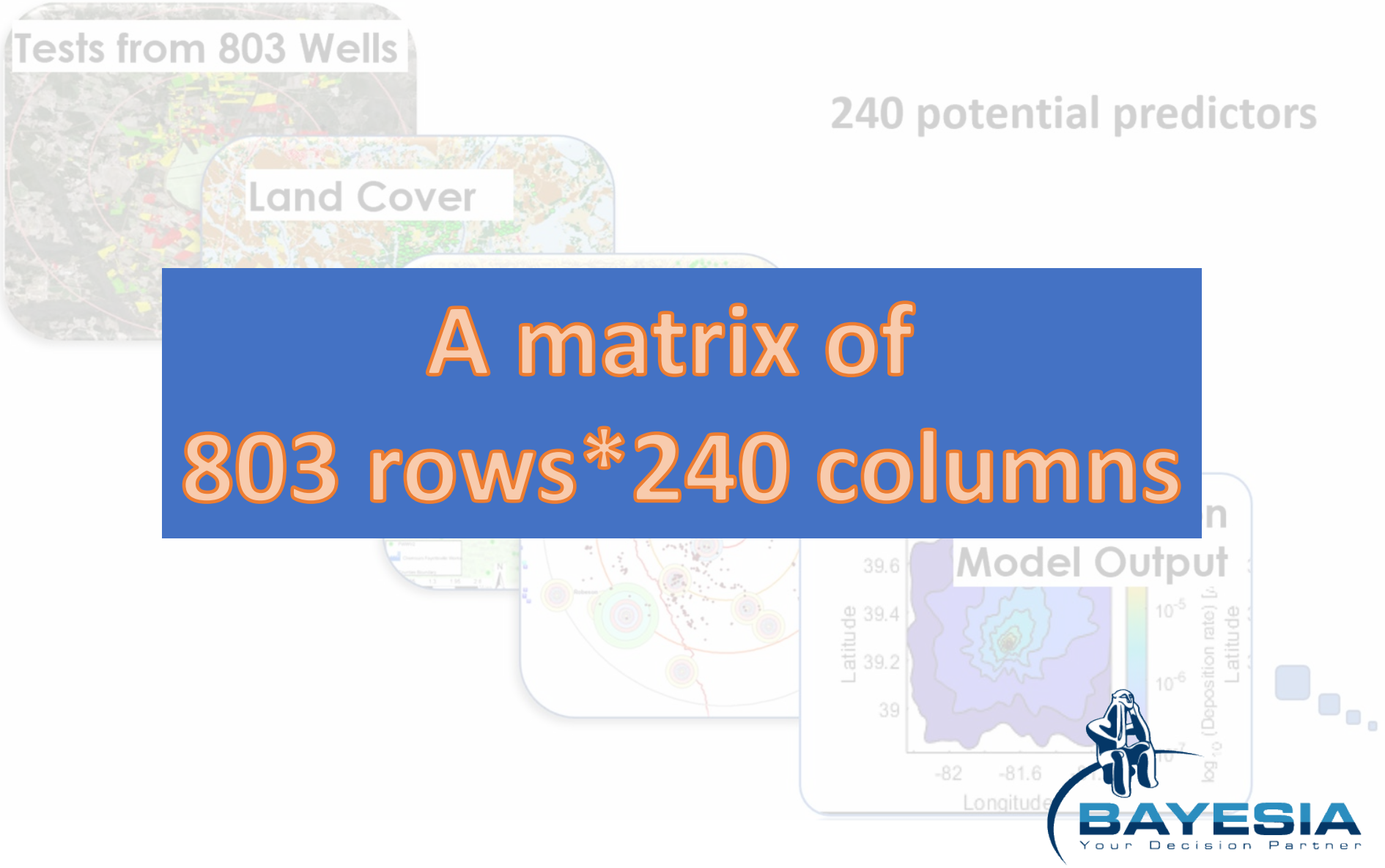
Image from: cleanwaterstore.com

We built a database of multiple factors that might influence GenX in well water



240 potential predictors

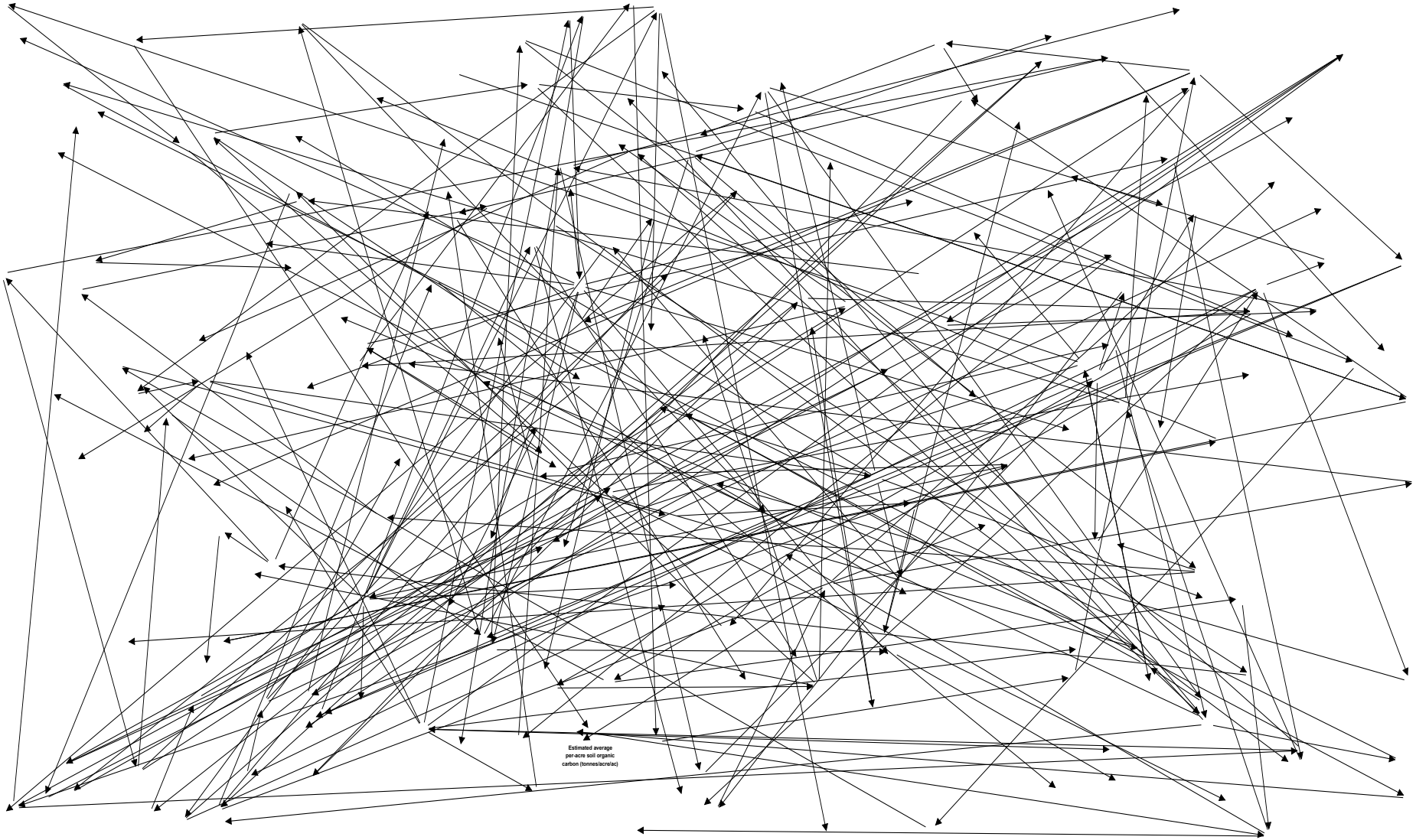
Our curated data set imported to BayesiaLab



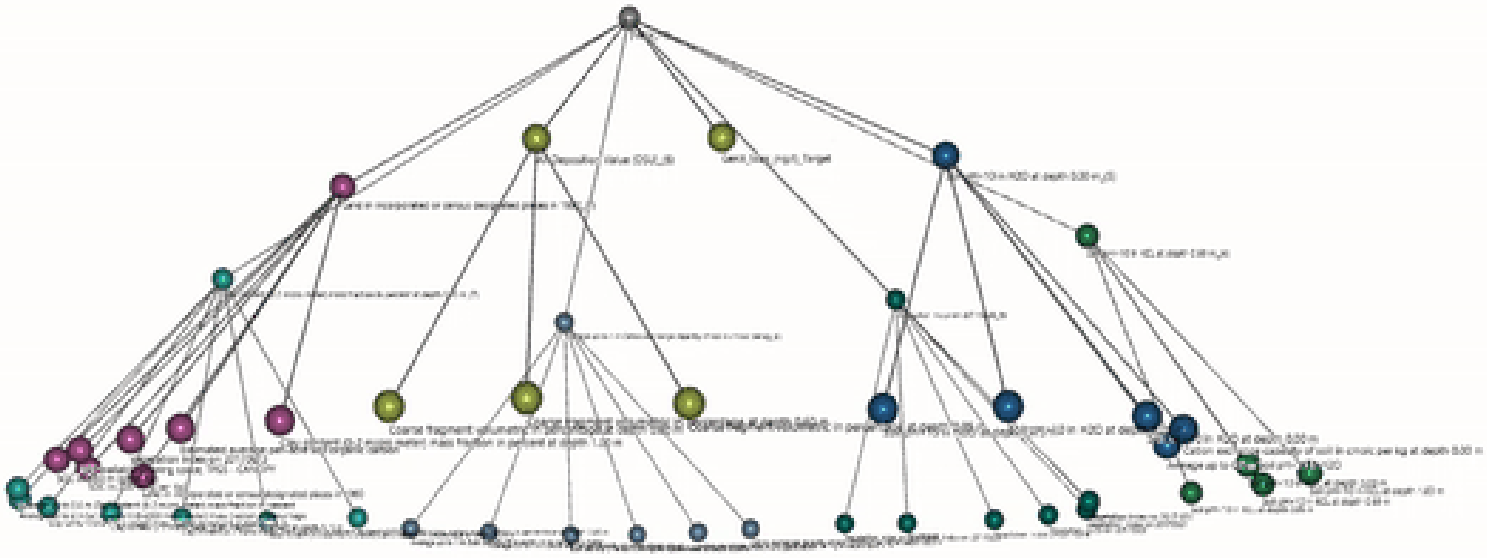
A matrix of
803 rows*240 columns

BAYESIA
Your Decision Partner

Finding influencing variables among this data set could be complicated



We used Bayesian network to cluster the variables and find pattern in the data set

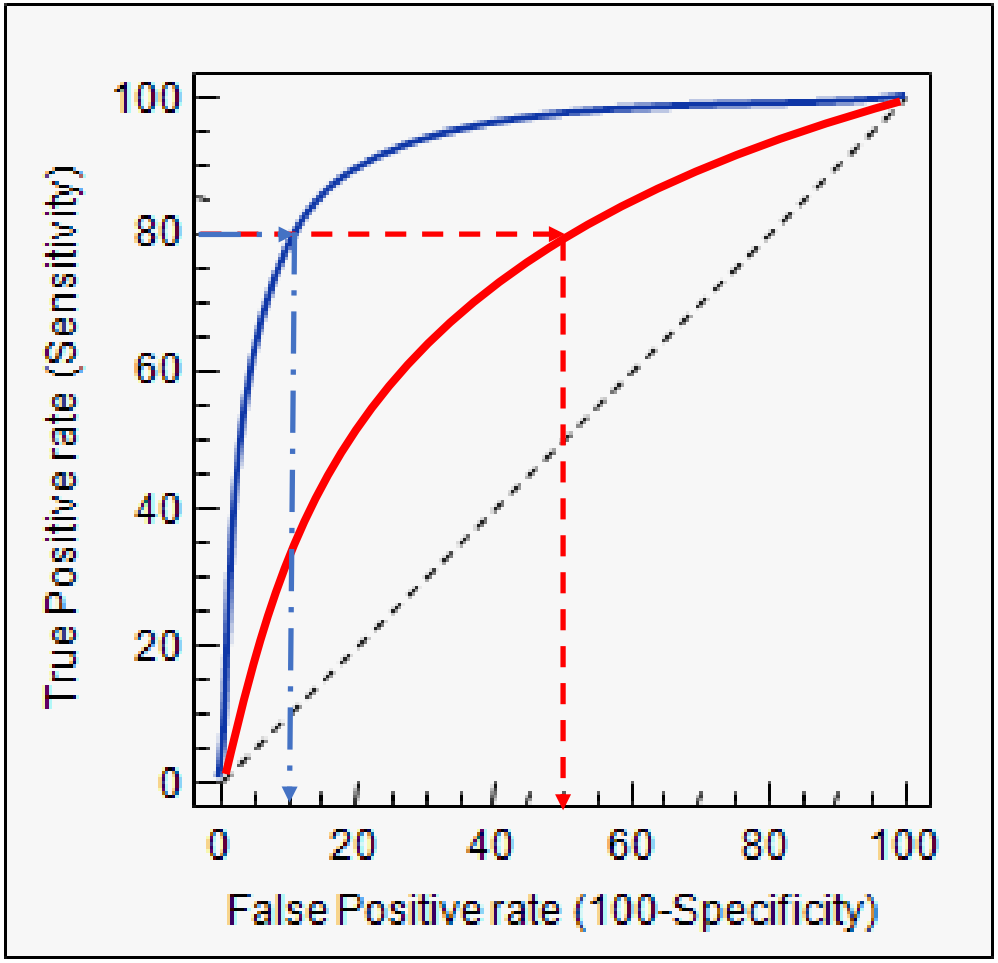


Multiple models built and tested Using BayesiaLab software

- Modeling goals:
  - Predict which untested wells are at risk of having GenX above NC health goal (140 ng/L)
  - Identify factors most influencing risk

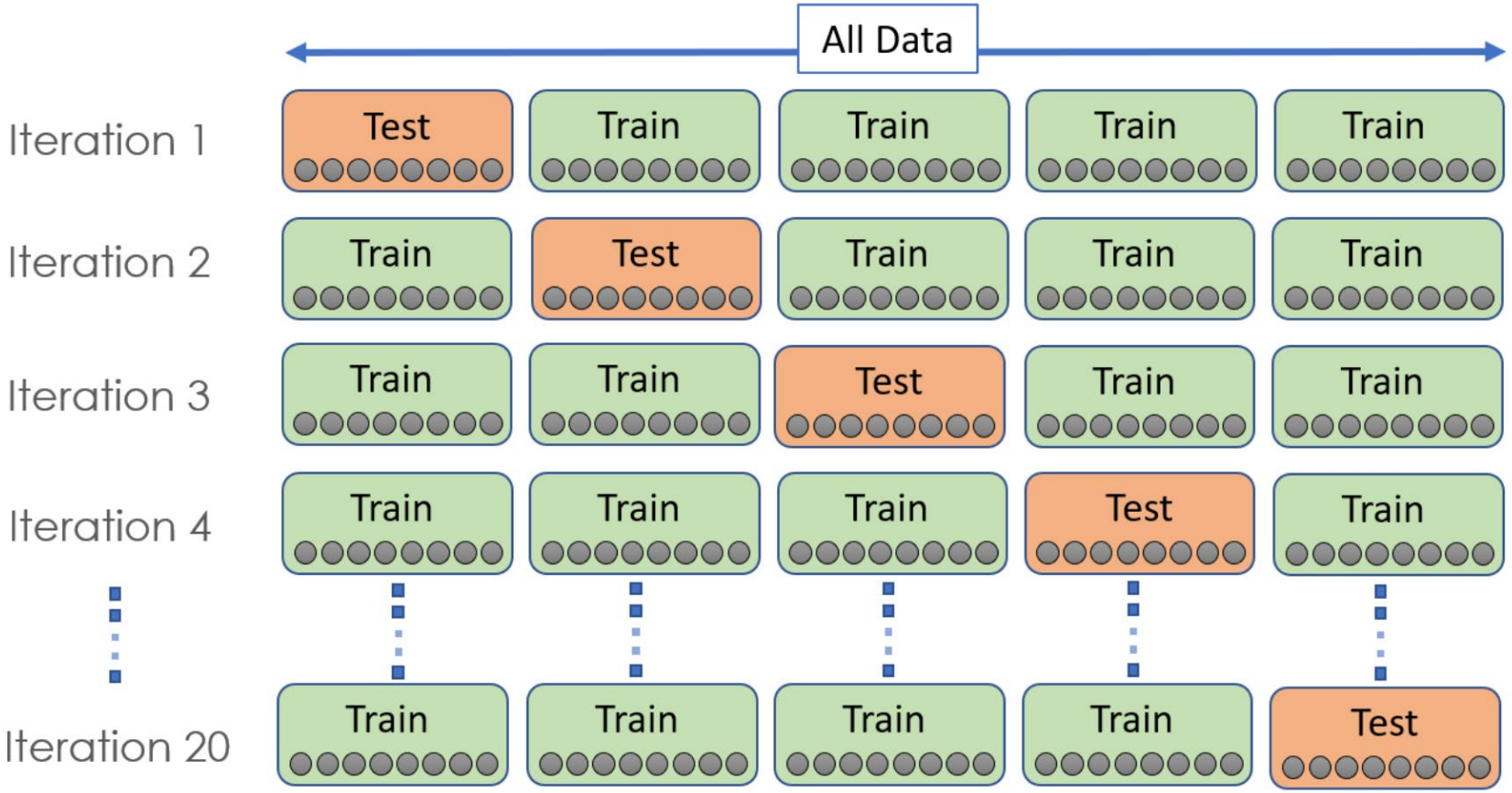
- Model building steps (in BayesiaLab)
  1. "**Unsupervised learning**" and "**variable clustering** "to discover which variables are most closely related to GenX
  2. **Elimination** of unrelated variables
  3. "**Supervised learning**" to build a predictive model
     - Augmented naïve Bayes algorithm often used for classification problems

Area Under Receiver-Operating Characteristics (ROC) curve used as performance metric



Prediction Accuracy = Area Under Curve (1.00 is perfect)

Accuracy tested in cross-validation



- Five-fold cross Validation
- Repeated 20 times
  - Different random train/test split each time

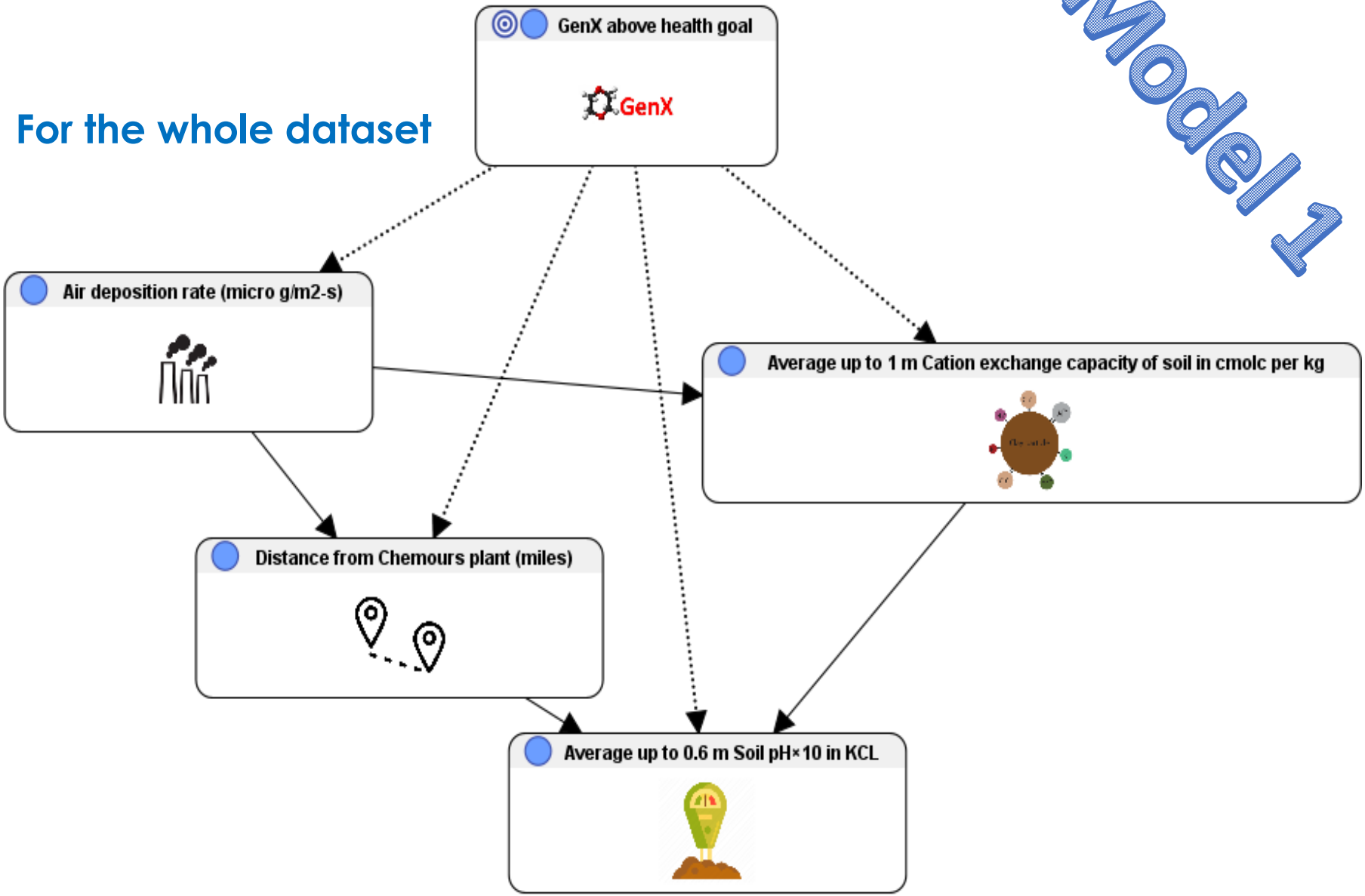
Results

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

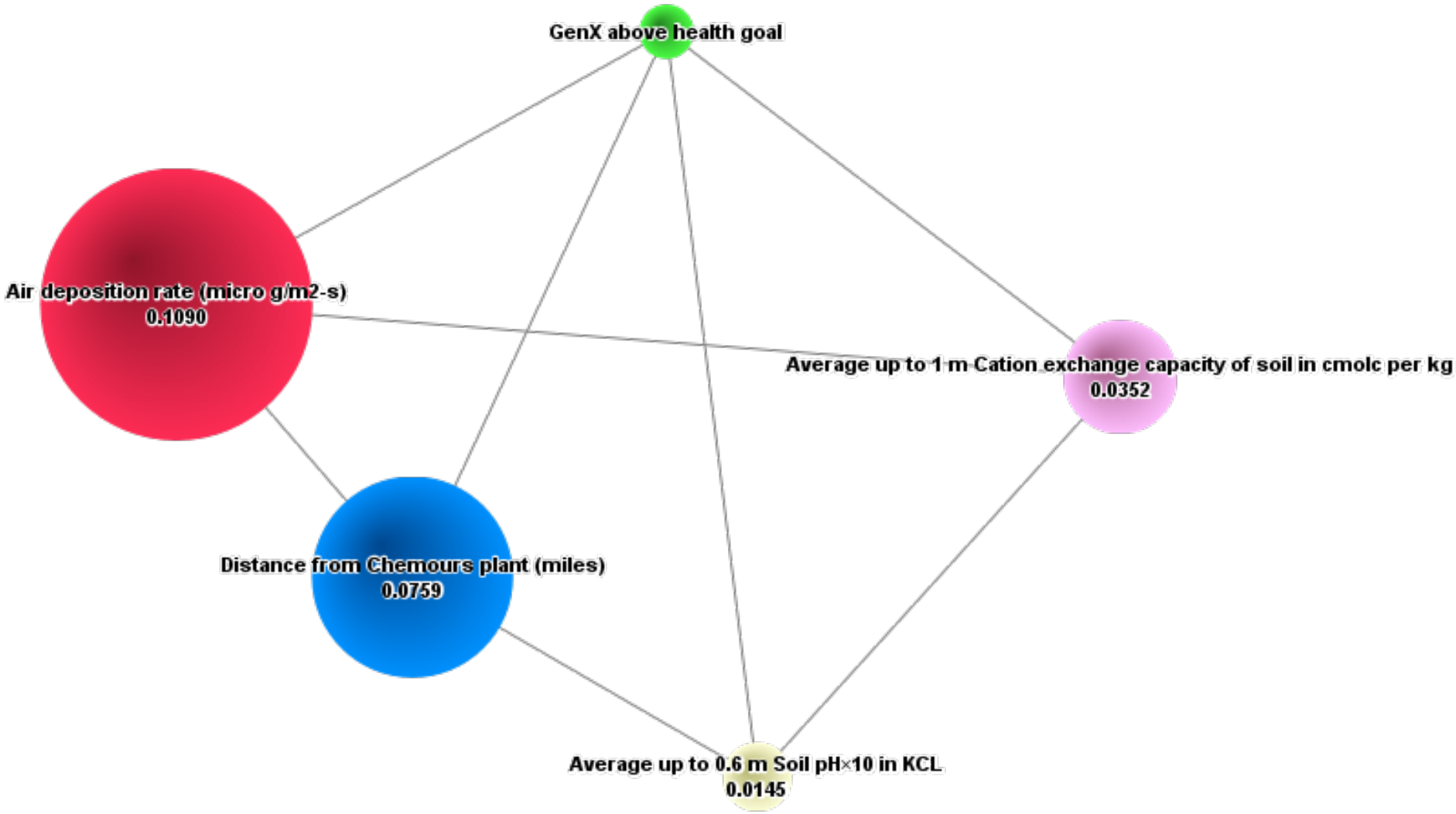
Main variables that influencing higher risk of GenX

Model 1

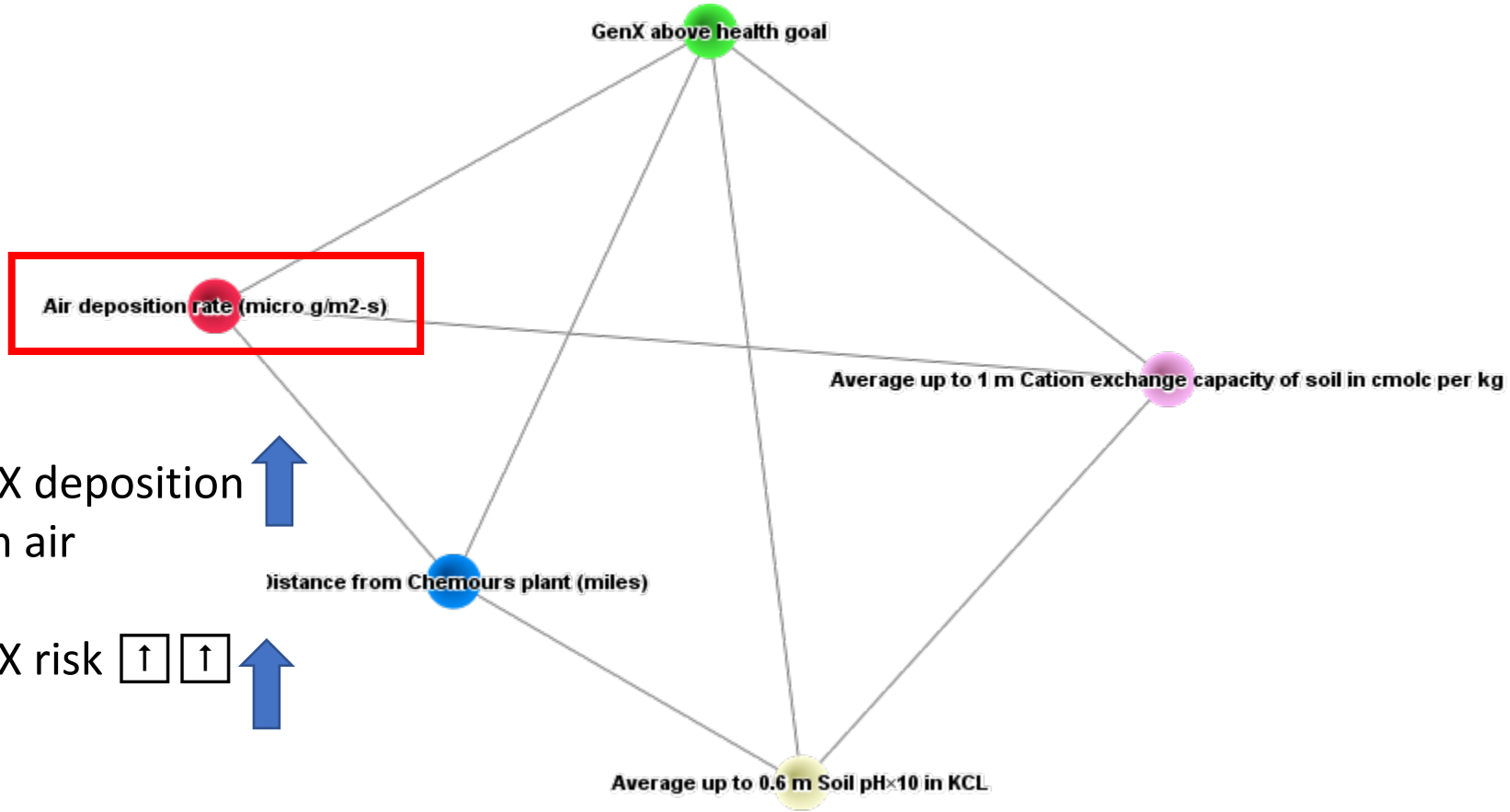
- For the whole dataset



Mutual information with the GenX as target node



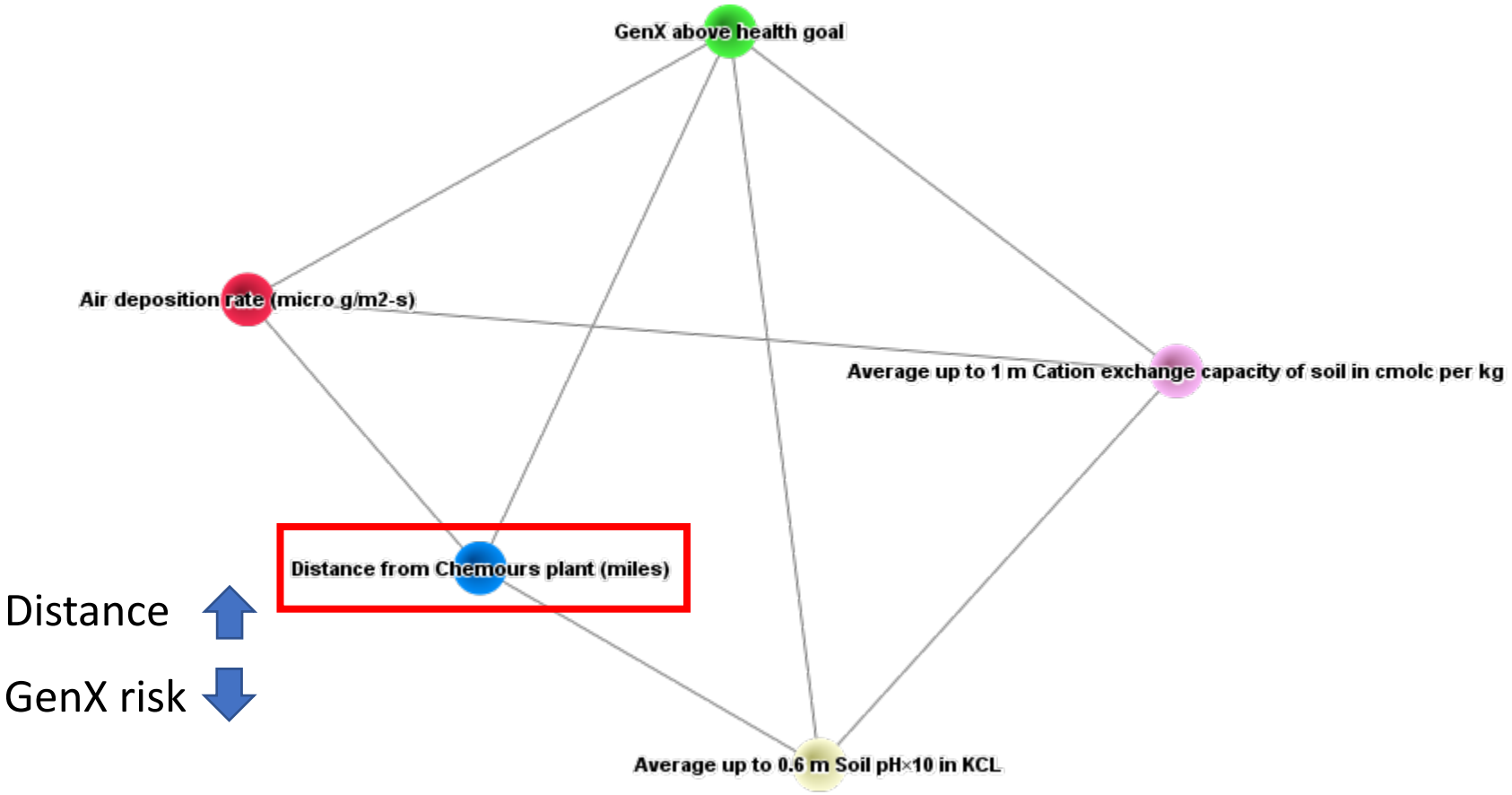
Knowing "Air Deposition Rate" provides the most information about GenX risk



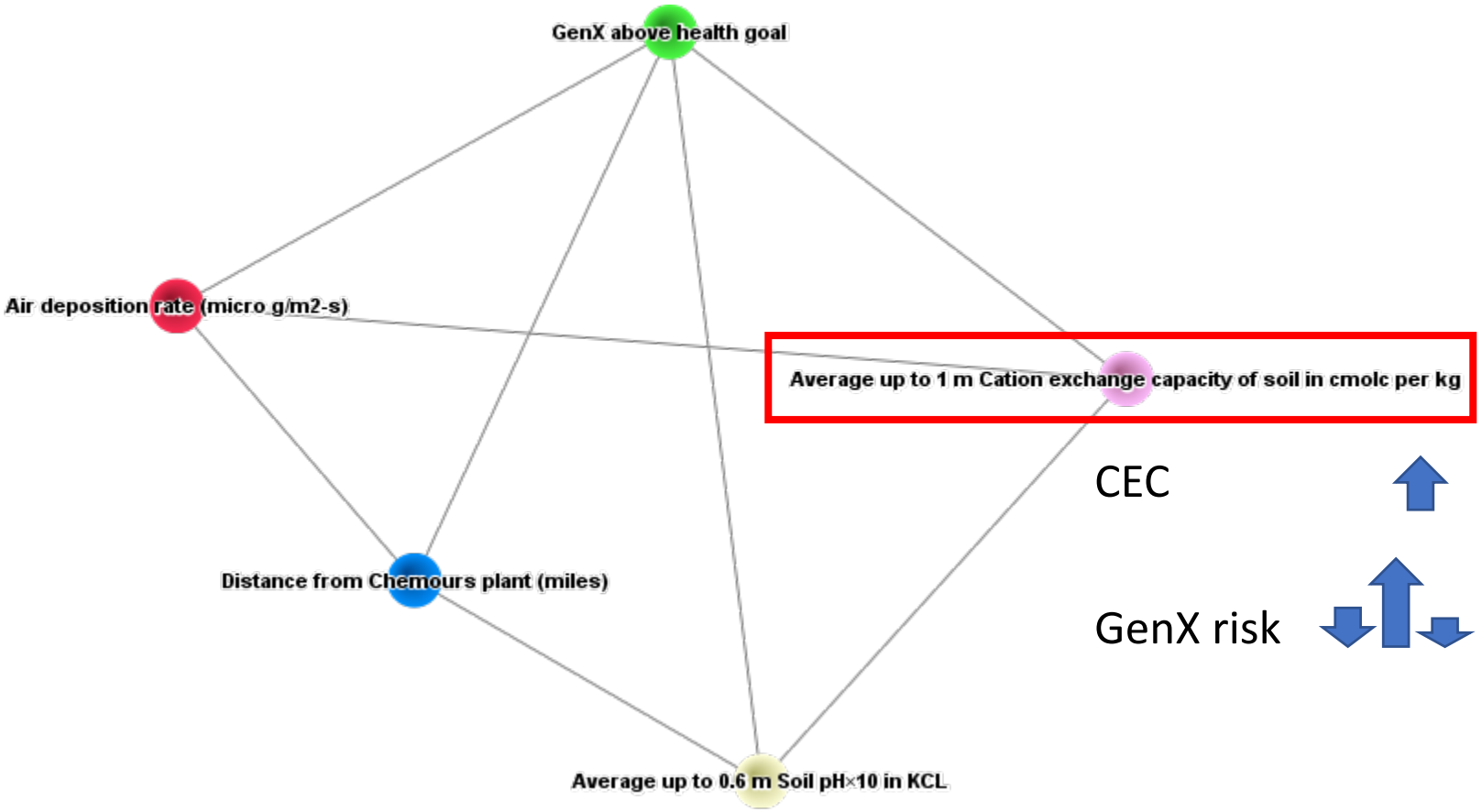
GenX deposition from air ↑

GenX risk ↑ ↑

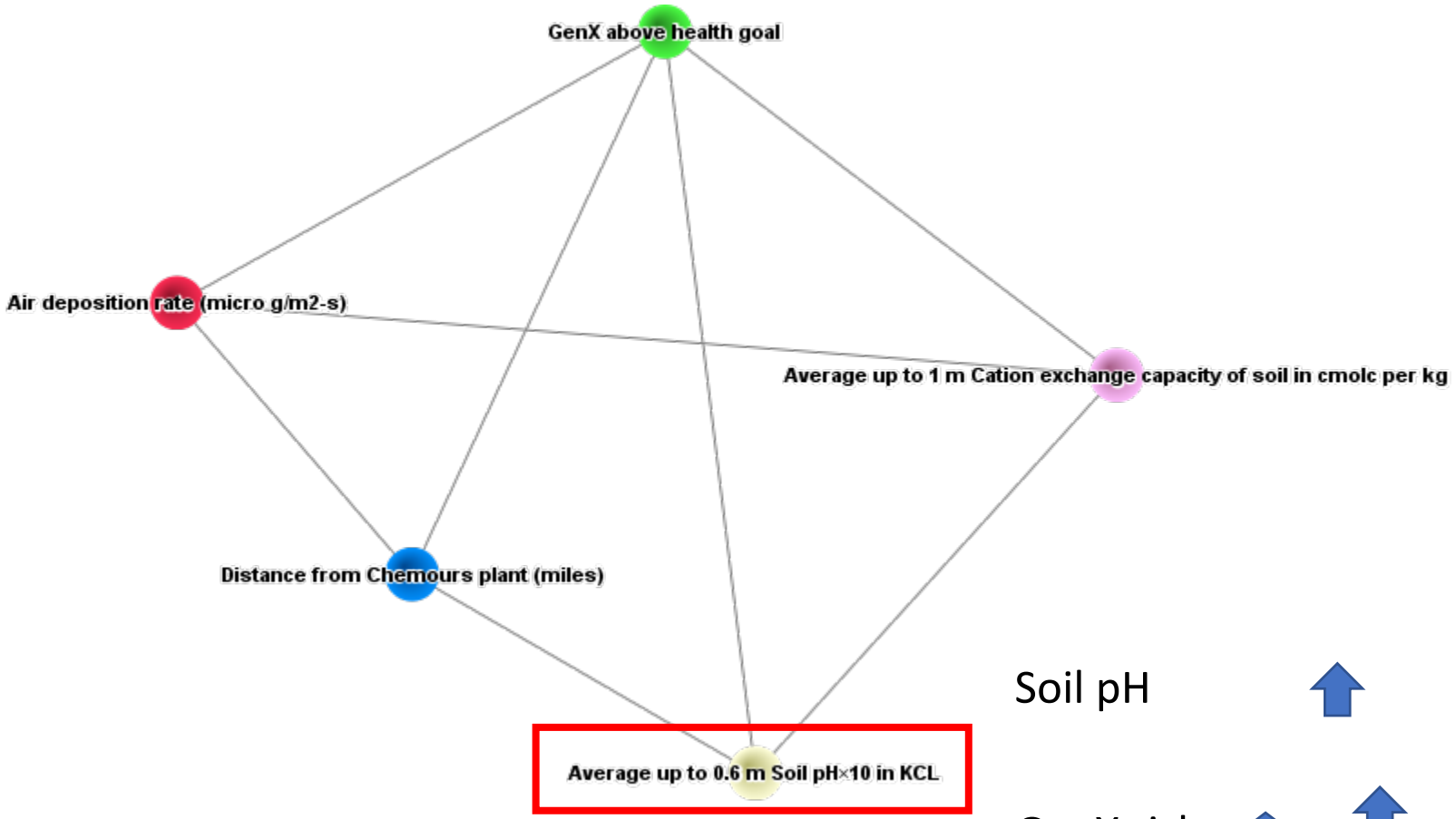
Distance from Chemours is second-most important predictor



Cation Exchange Capacity is third-most important predictor



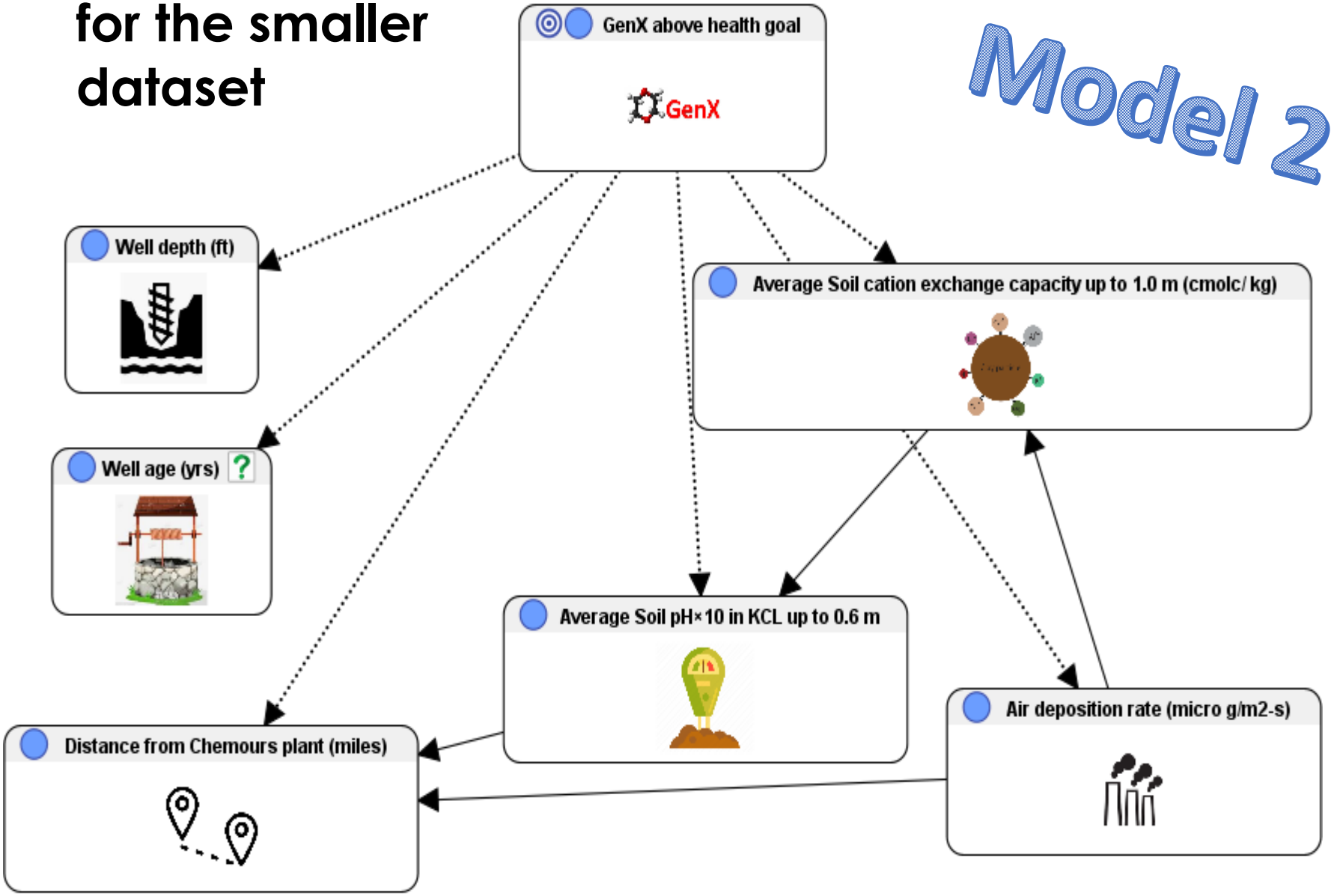
Soil pH in KCL is forth-most important predictor



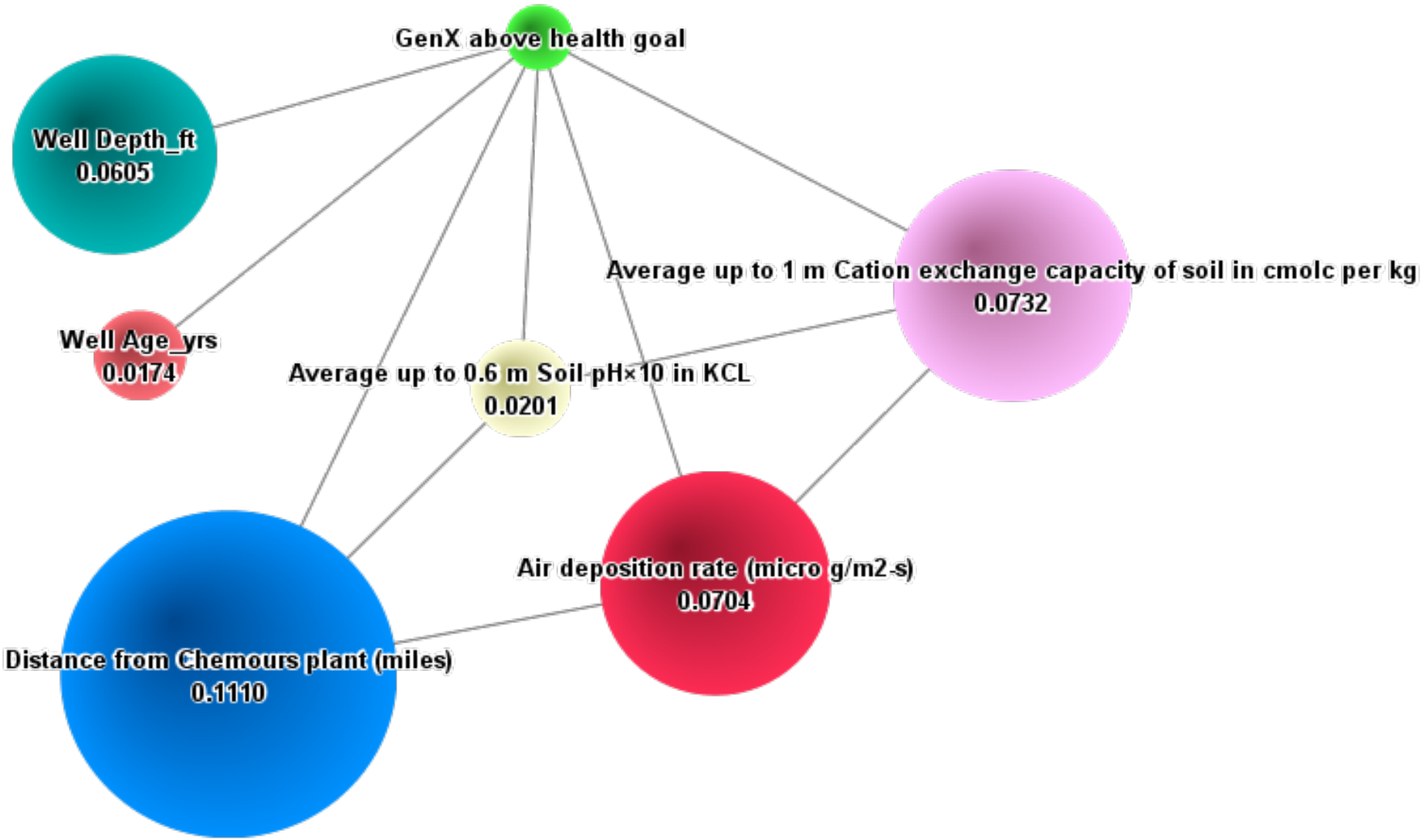
For 25% of houses we have additional information

- ✓ Well depth
- ✓ Year of construction
- ✓ Method of construction

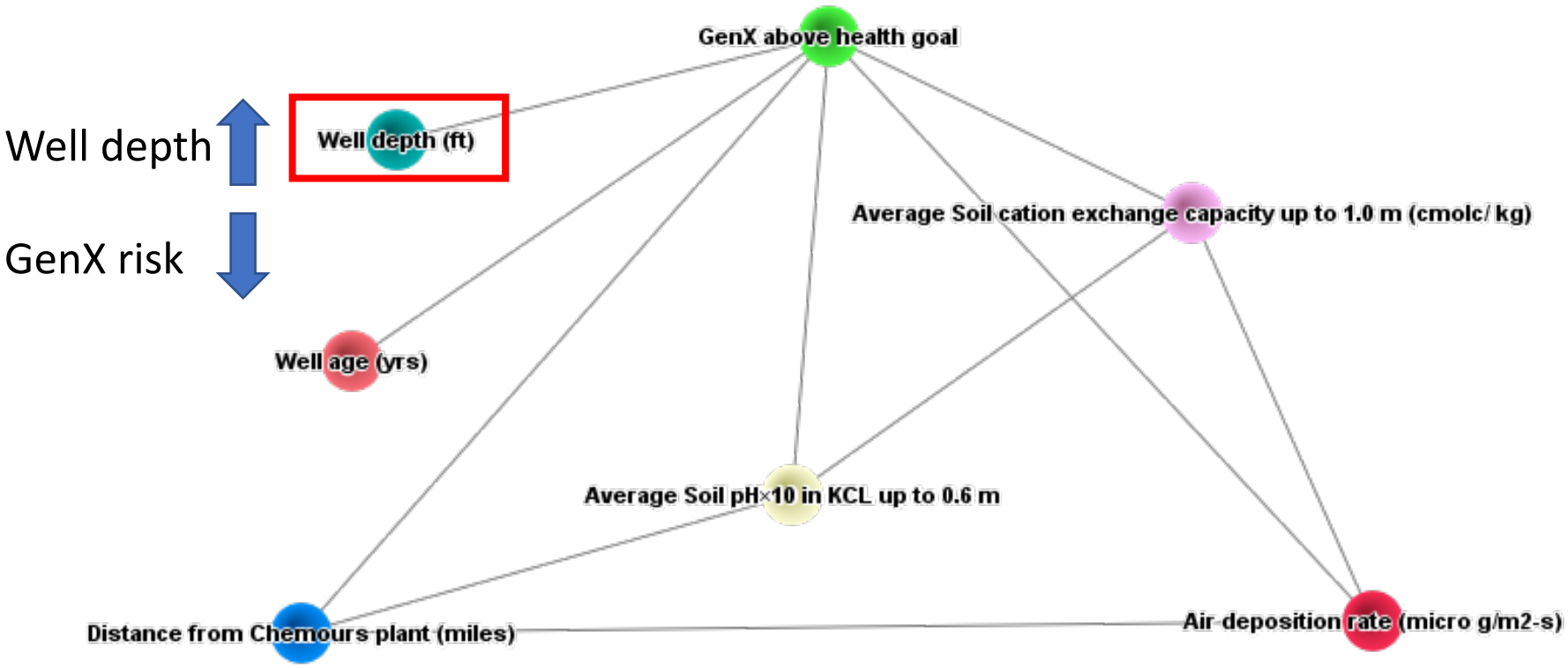
Main variables that influencing higher risk of GenX for the smaller dataset



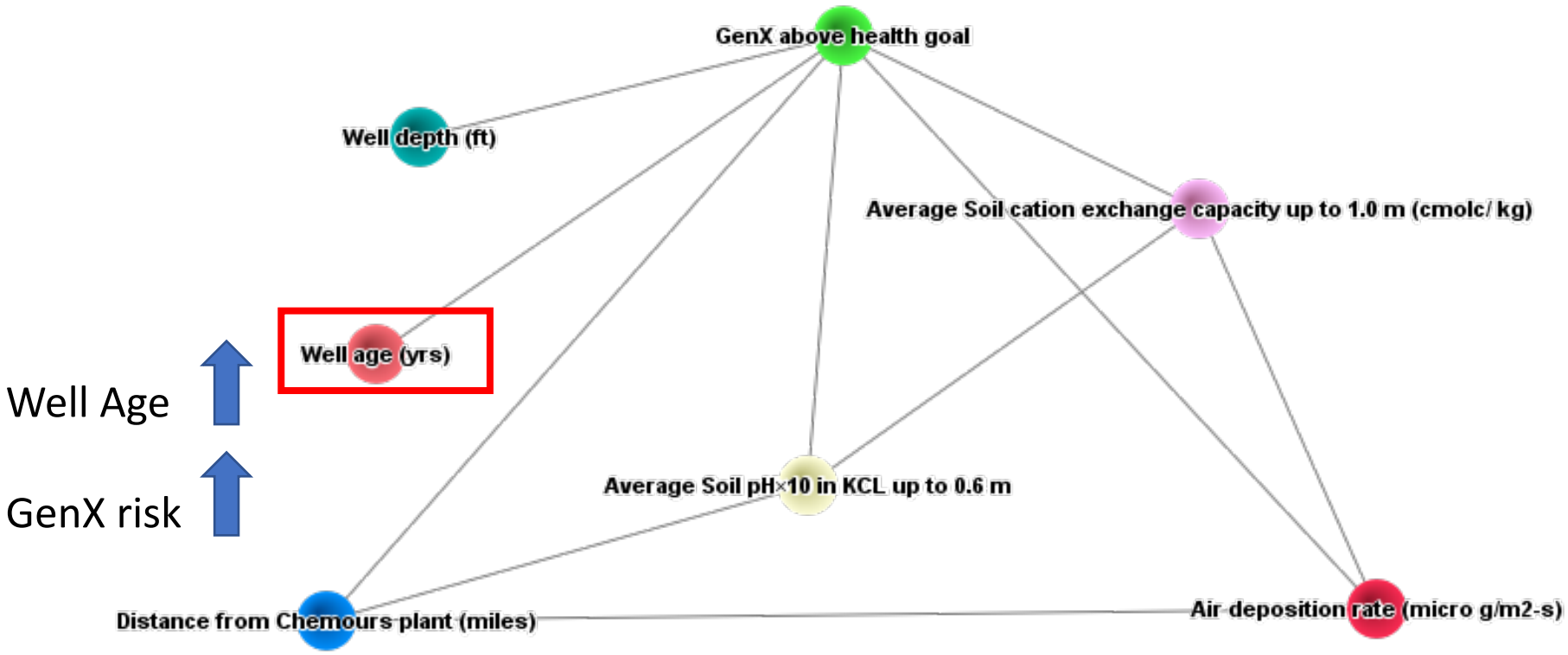
Mutual information with the GenX as target node



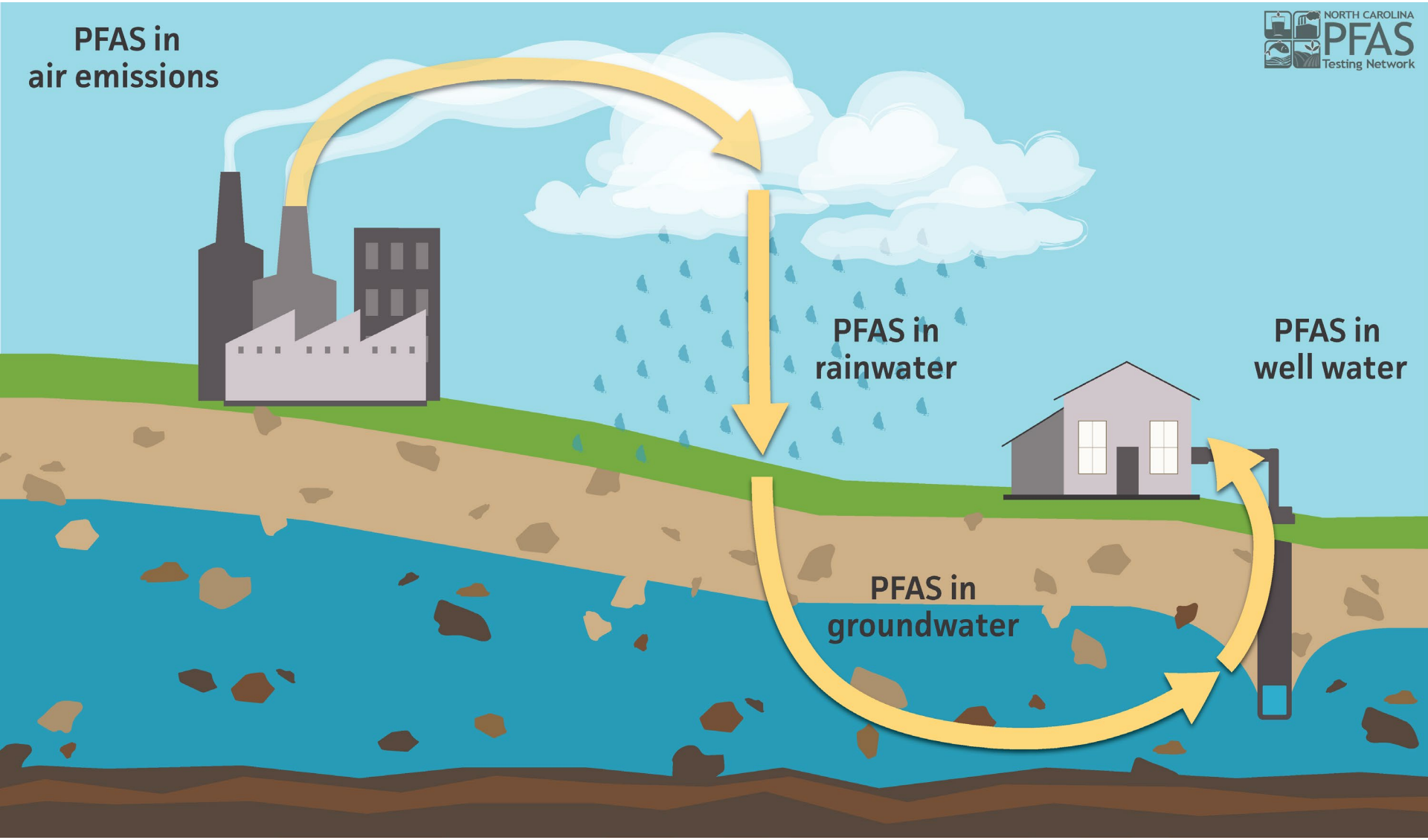
Well depth can reduce the risk of Genx contamination



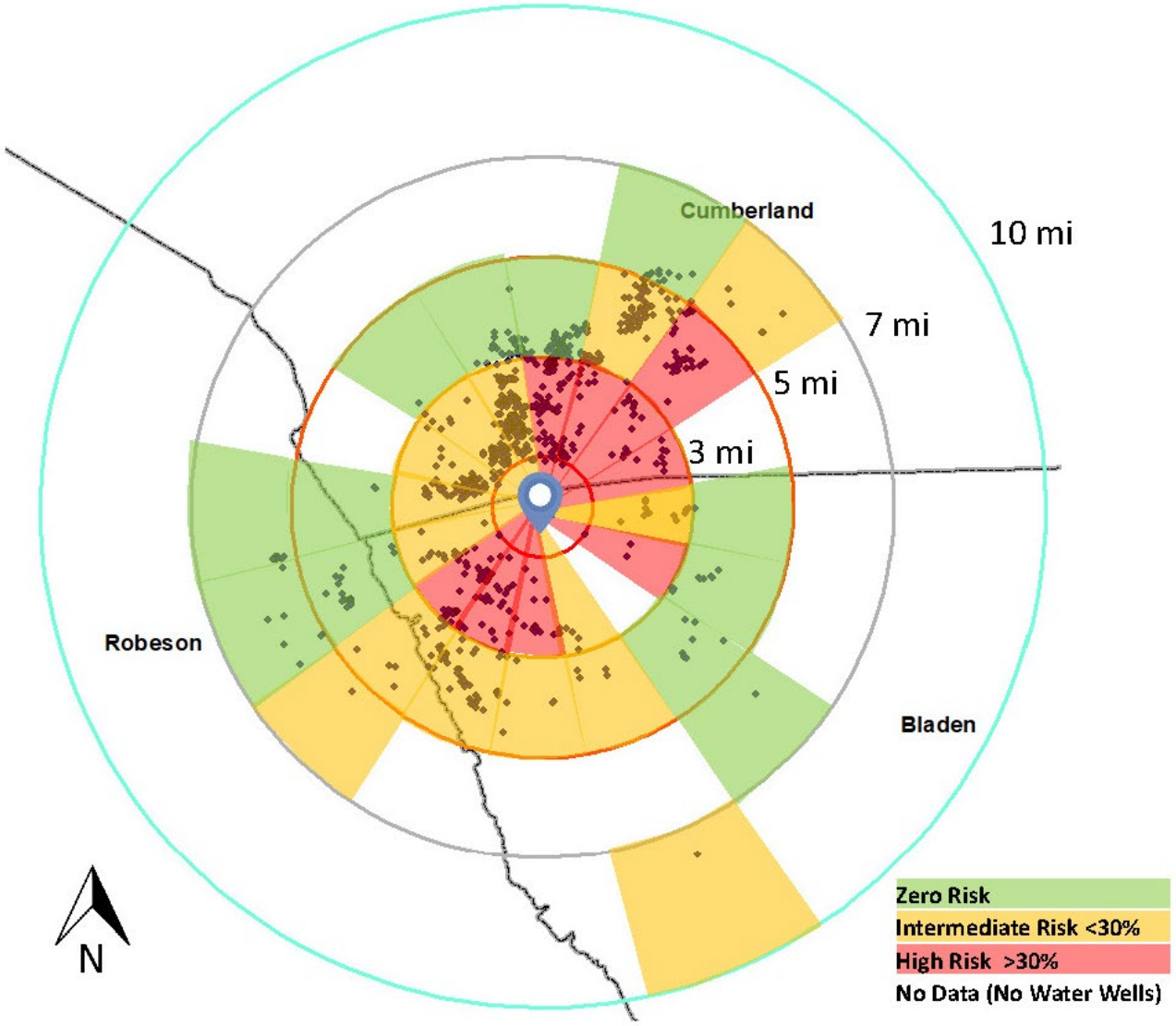
Increase in Well age may cause higher risk of Genx



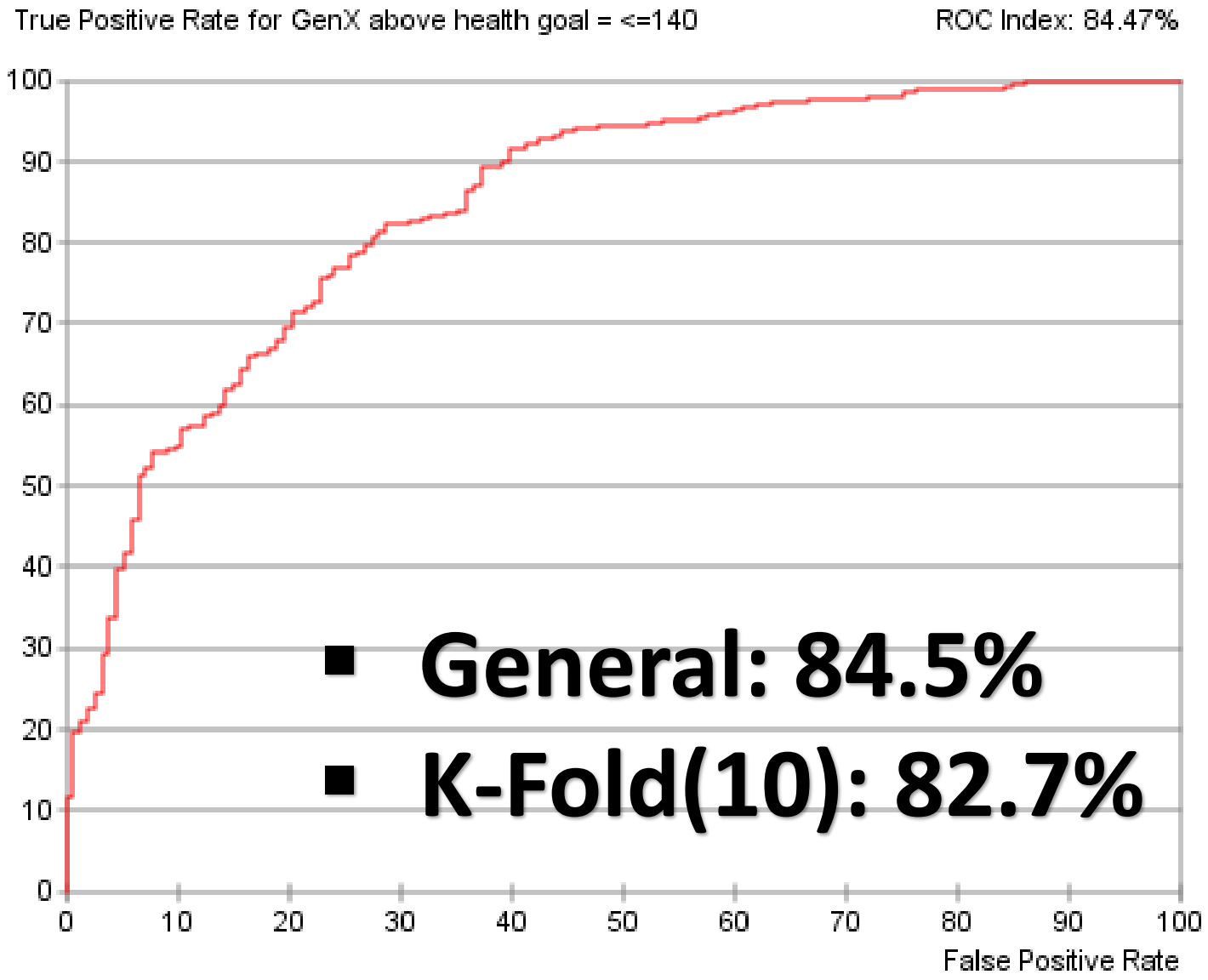
Our ML model confirmed the source and how it reaches to the houses



Predictive risk in the area of study

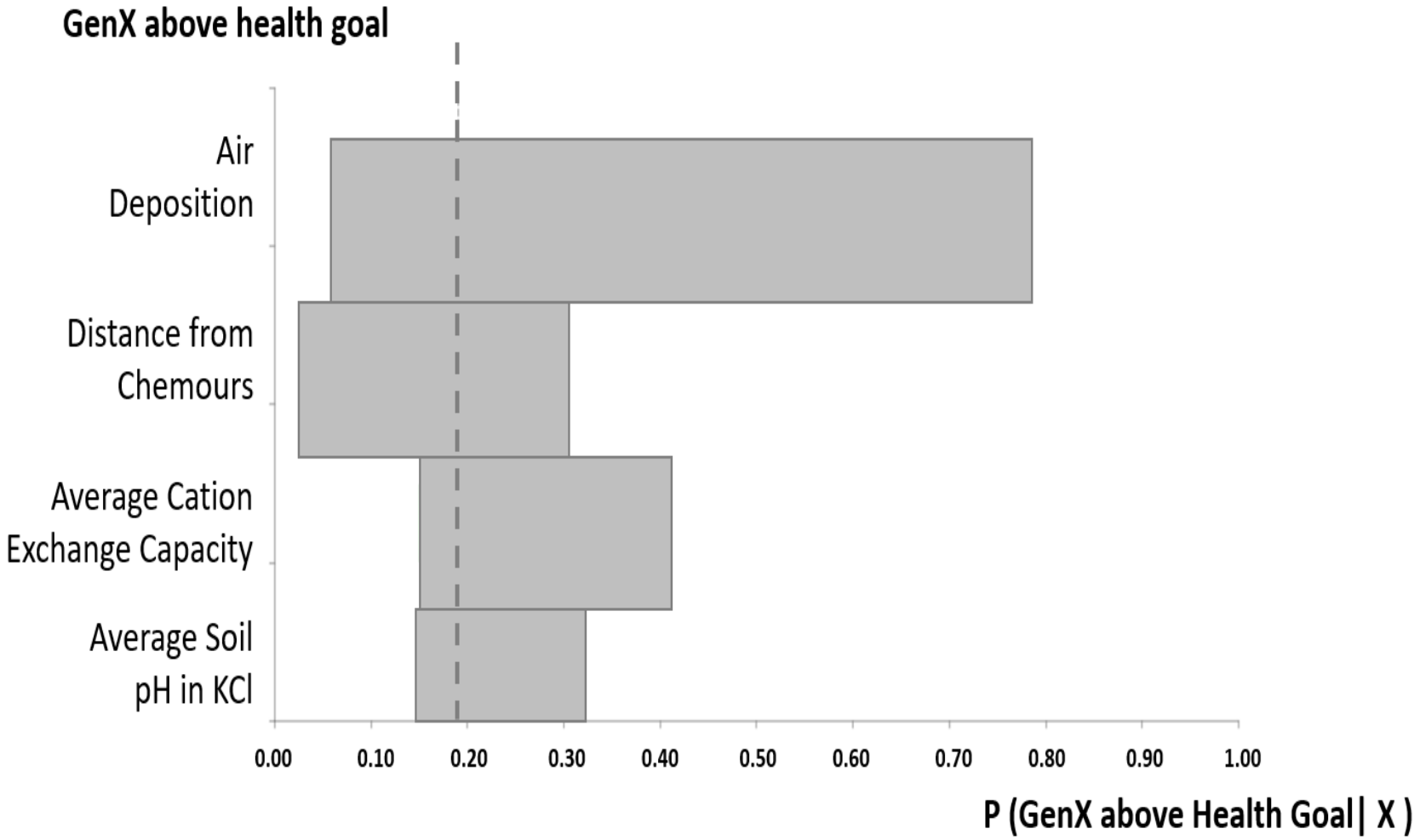


ROC curve for model 1 – whole dataset

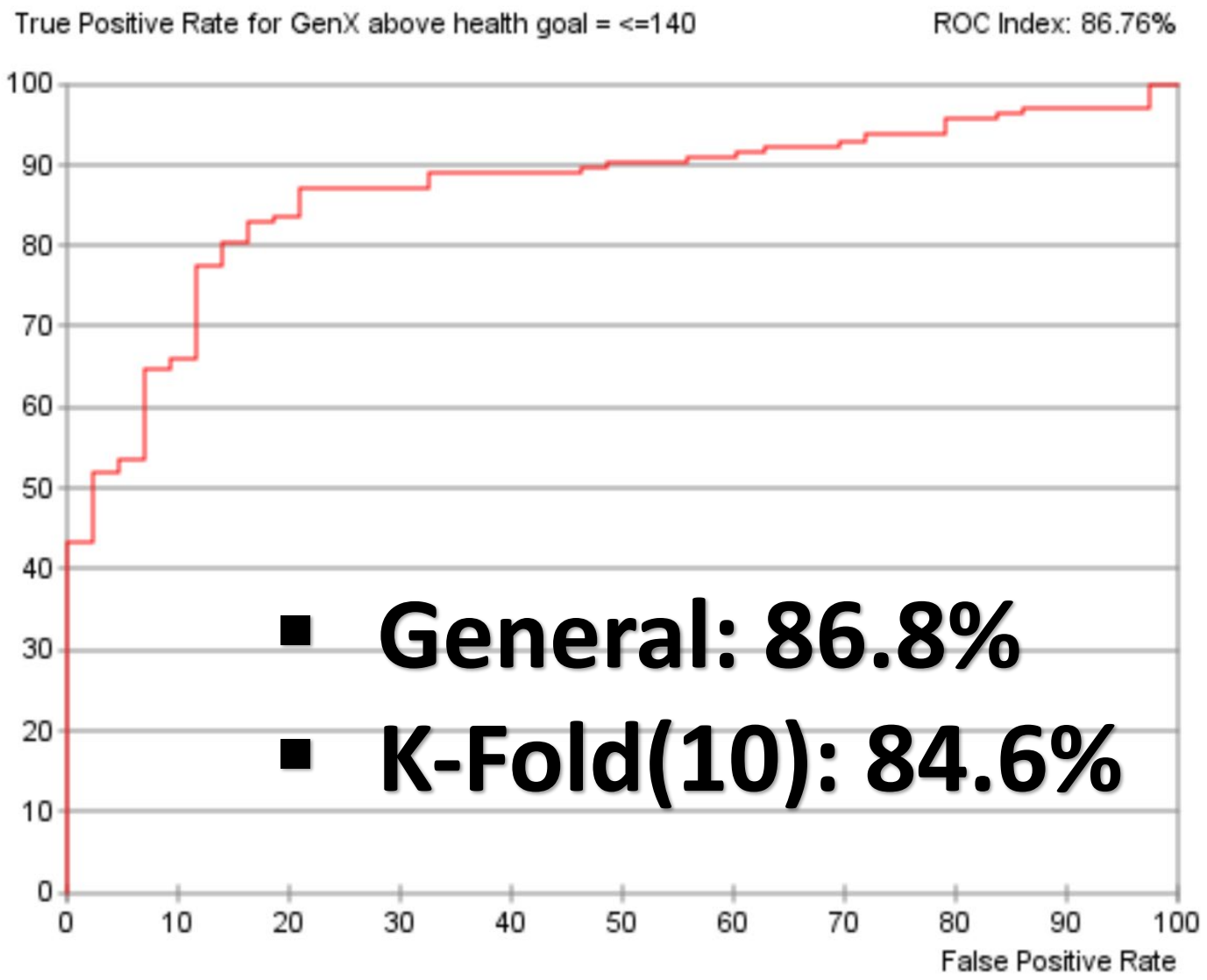


- General: 84.5%
- K-Fold(10): 82.7%

Tornado graph shows variables with most influence on risk – model 1

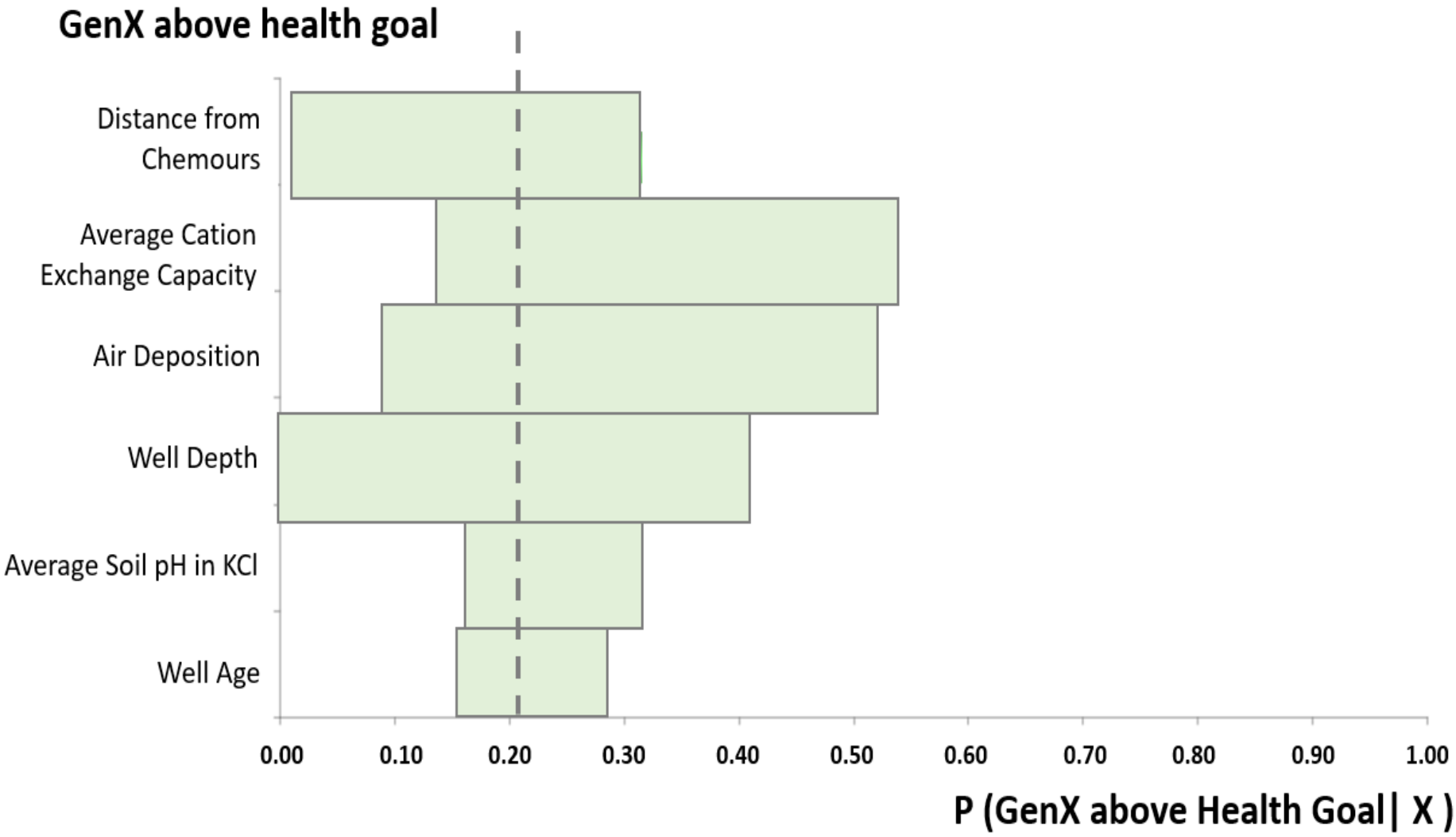


ROC curve for model 2 – 25% of data set

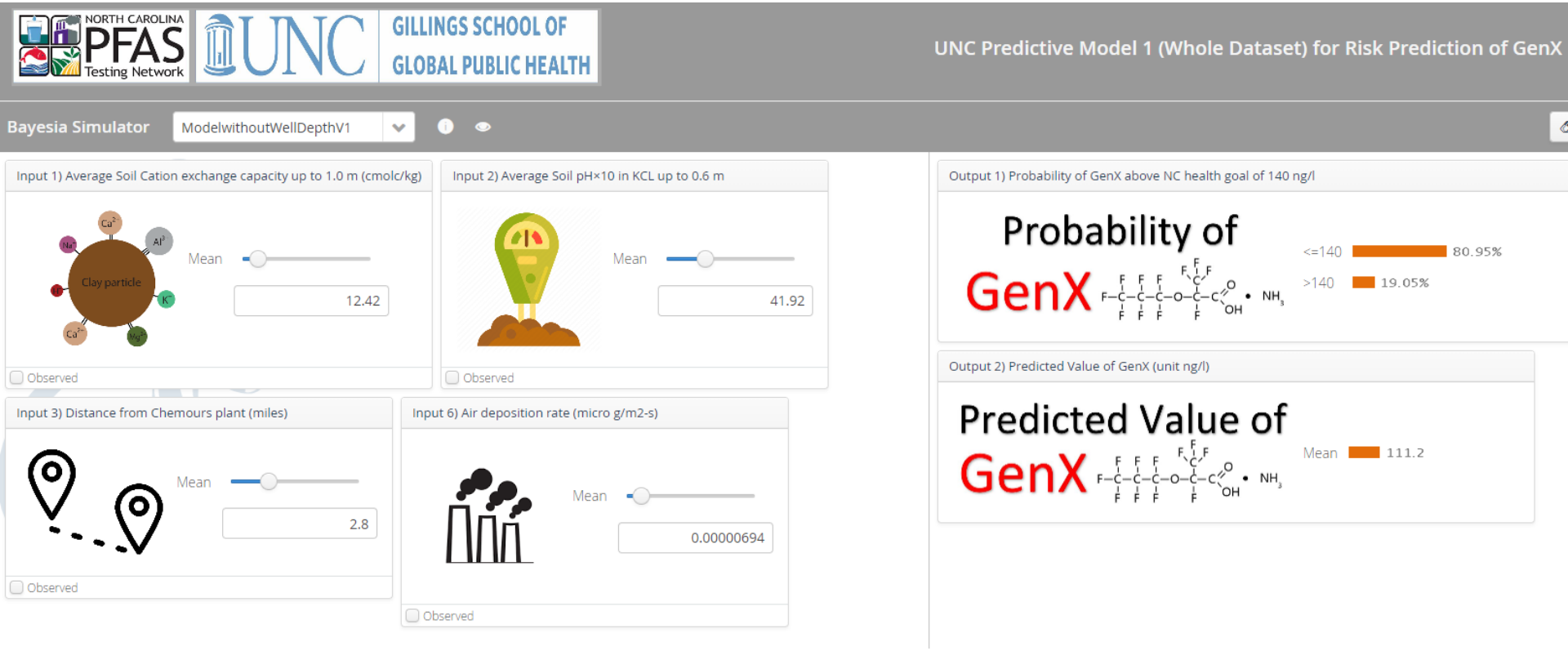


- General: 86.8%
- K-Fold(10): 84.6%

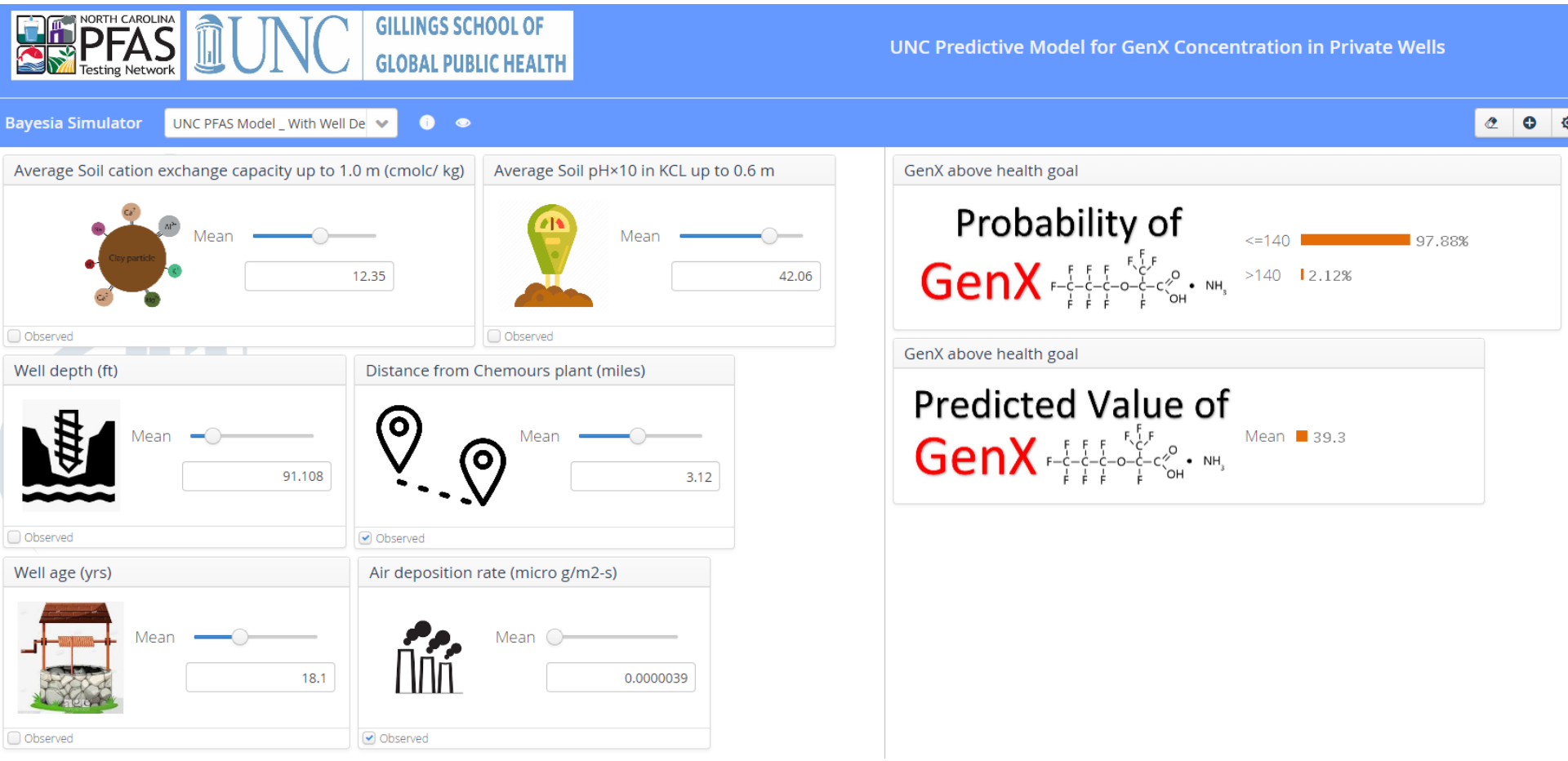
Tornado graph shows variables with most influence on risk – model 2



The web simulator for model 1



The web simulator for model 2



Conclusion

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

Conclusion

- Bayesian network identified **factors** associated with the GenX Concentration
  - Proximity to Chemours Plant
  - Air Deposition
  - Soil pH
  - Cation Exchange Capacity

- BN model can classify the private wells with above **average sensitivity (84%)** and reasonable false-positive rate.

- Adding more well permit **data and more measurement** to the model can help increase ROC index.

Main takeaways

- Although PFAS issue is complicated, we have **tools** and datasets to build ML models which can help us understand PFAS better.

- Bayesian network modeling has the capability to show us the **influencing variables** in PFAS contaminations

- The **accuracy** of the ML model is reasonably good, however **more data** (such as well-depth, age, and measurements) can help it increases the accuracy.

Acknowledgement

❑ Funding Source

- North **Carolina** Policy Collaboratory

❑ Collaborators

- Dr. Detlef Knappe, NCSU
- Dr. David Genereux, NCSU
- Riley Mulhern, UNC
- Dr. Frank Stillo, UNC
- Ariel Lam, UNC
- Alex Shoaf, UNC

❑ Organizations

- Local Health Department at Cumberland, Robeson and Bladen County
- NC DEQ

Questions?

UNC | GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

References

S. Conrady;, Bayesian Networks & bayesialab. 2015.

SHEENA SCRUGGS, 2019, PFAS — a problem in North Carolina drinking water, https://factor.niehs.nih.gov/2019/3/feature/2-feature-pfas/index.htm

Urban Forest Data for North Carolina
https://www.nrs.fs.fed.us/data/urban/state/?state=NC

Soil survey ,
http://www.ncmhtd.com/arcgis/rest/services/NRCS/NRCS_SoilData/MapServer/4

National Land Cover Database (NLCD)
https://www.usgs.gov/centers/eros/science/national-land-cover-database?qt-science_center_objects=0#qt-science_center_objects

Fire Detection GIS Data
https://fsapps.nwcg.gov/gisdata.php