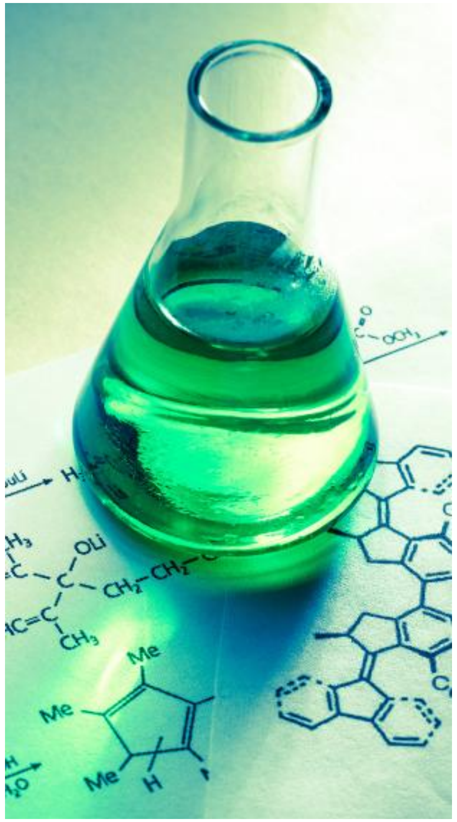
TURI

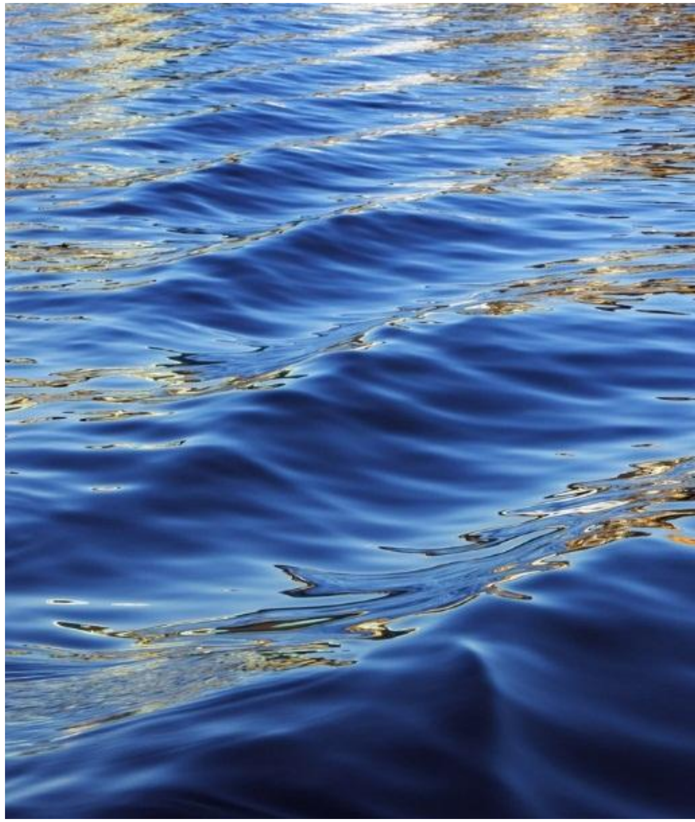
TOXICS USE REDUCTION INSTITUTE

Using Artificial Intelligence to Support Toxics Use Reduction (AI4TUR)

Dr. Gabriel Salierno – TURI

December 9th, 2025





Statistics, Neural Networks (NN), Machine Learning (ML), Artificial Intelligence (AI)

AI isn't magic: Statistics and information theory provides the foundation for understanding data and drawing interpretable conclusions.

Neural Networks and Machine Learning build upon statistics to create powerful prediction models, often sacrificing interpretability.

AI: Systems that replicate intelligent behaviors through learning from data and applying predefined rules and algorithms. Examples include task-specific Large Language Models (LLM), and logic-based systems.

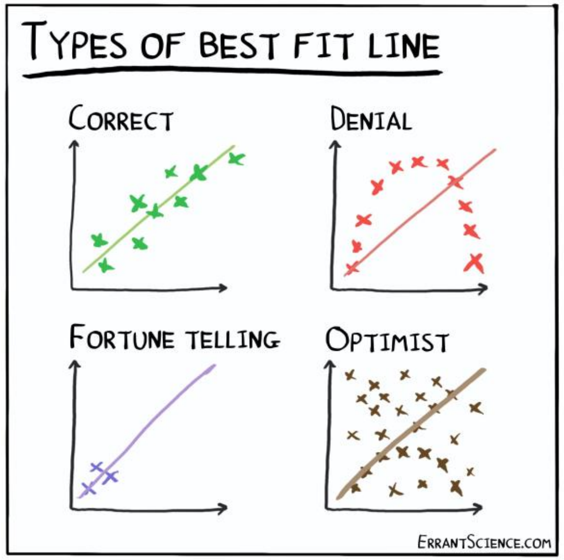
Generative AI learns from data to create new content: text, voices, images, video, and computer program codes.

Ultimate goal: Create machines that exhibit (super) human-like intelligence.

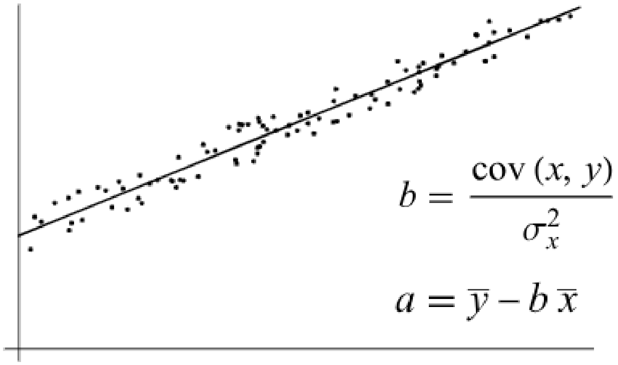
TURI
TOXICS USE REDUCTION INSTITUTE

Curve fitting: Traditional statistics vs ML

Model parameters with meaning are preferred



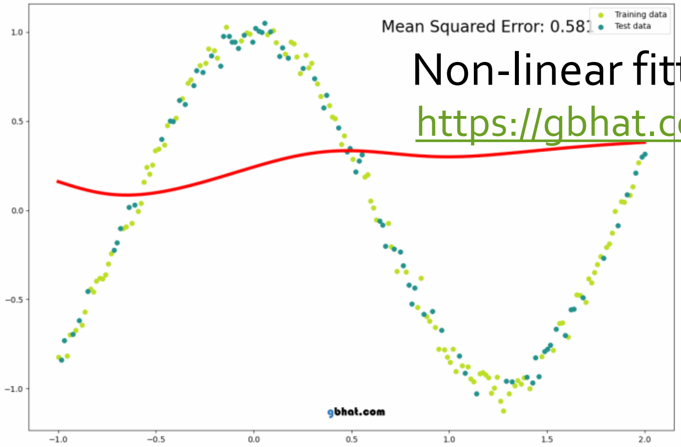
There is a concise formula for linear fitting to find the model with the least error.



$$b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2}$$

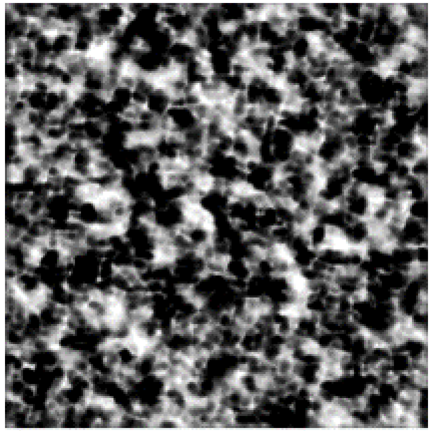
$$a = \overline{y} - b\,\overline{x}$$

For more complicated problems, we have to 'find' the model and its parameters



Non-linear fitting

https://gbhat.com/



Medical imaging

https://monai.io/

- ***Data quality outweighs quantity:*** Even massive amounts of data won't be helpful if noisy, biased, or irrelevant to the task.
- ***Parsimony:*** the more complex the model, the less meaningful and more risk of overfitting.

Model Verification (Fit) vs. Validation (Predict)

Verification (Fit -> Train):

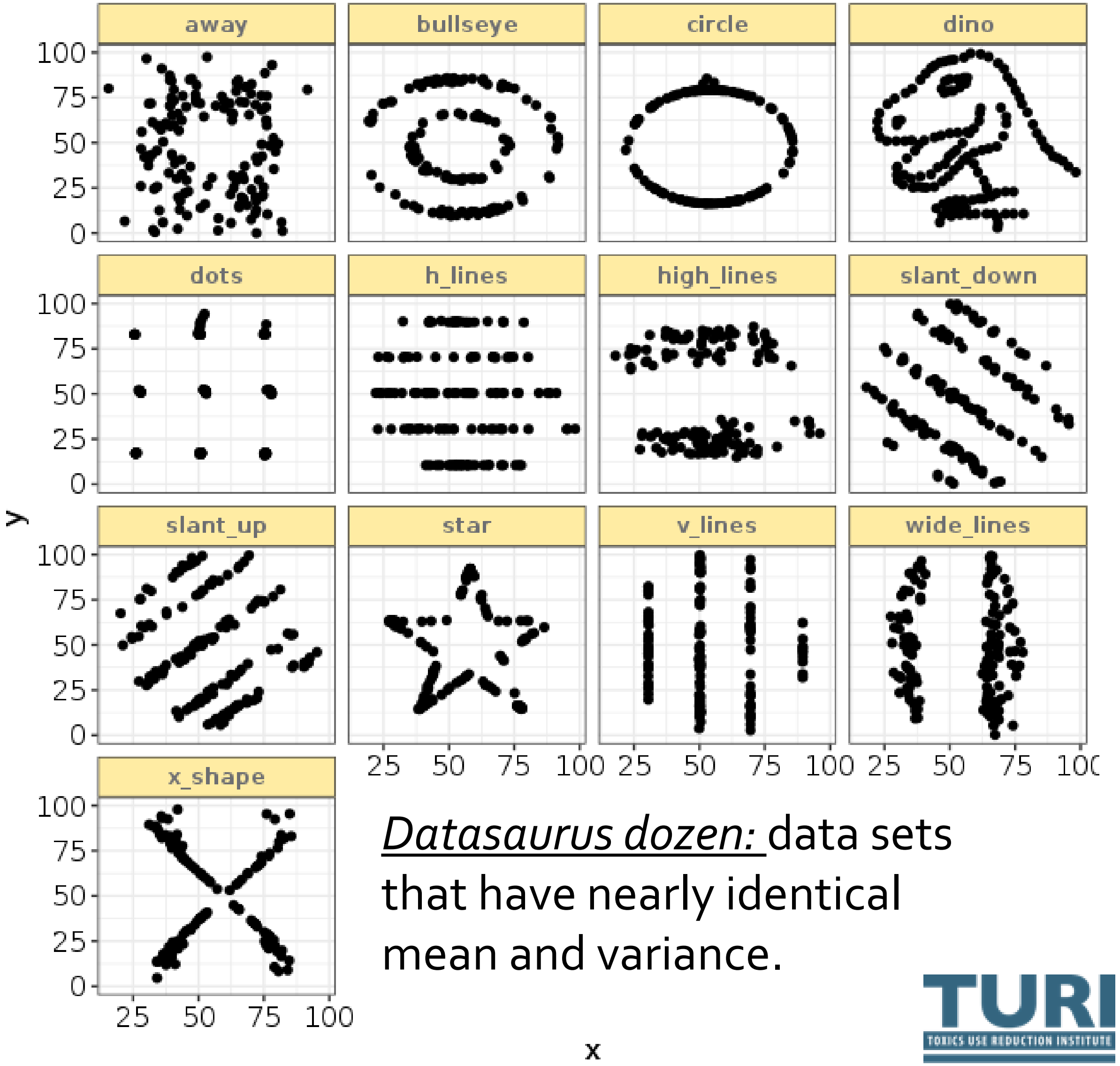
- Focuses on model accuracy with training data
- Tests if model reproduces known behaviors
- Evaluates internal consistency
- Maps model parameters to physical reality

Validation (Predict -> Test):

- Tests model's predictive capabilities
- Evaluates performance on unseen conditions
- Determines transferability and generalization
- Confirms practical utility and reliability

Avoiding correlation *mirages*:

- ***Correlation is not causality***
- ***Not everything that fits can predict***



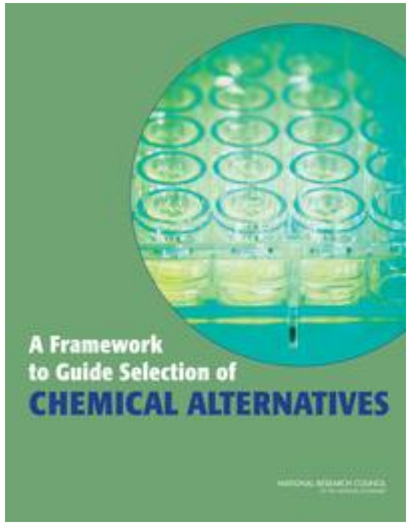
Datasaurus dozen: data sets that have nearly identical mean and variance.

Garbage in - Garbage out

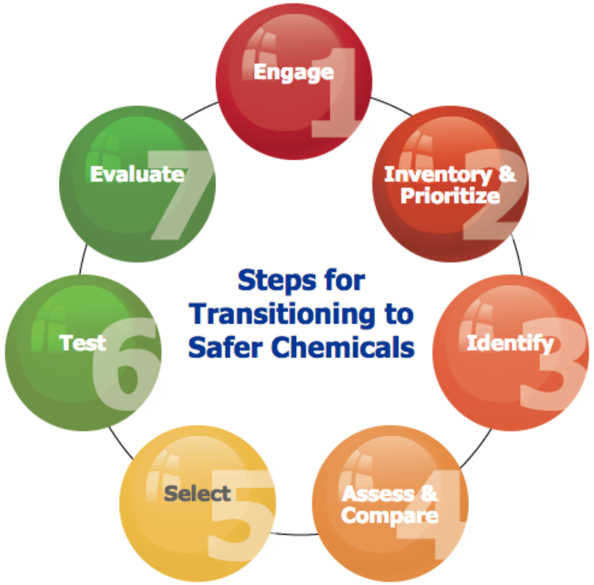
- **Models are as good as their data.**
  Poor data (e.g., inconsistent experimental conditions, missing physical features) leads to unreliable models, whether using classical statistics or machine learning.
  - **Classical Statistics:** Sensitive to outliers and assumes clean, structured data. Errors in data can skew coefficients and mislead interpretations.
  - **Machine Learning:** Can handle noise better but might be overfitting to artifacts in bad data (e.g., biased molecular descriptors).
- **Recursive training ("echo chambers"):** AI learning from its own outputs leads to (some) rapid improvement but also biases and uncontrolled evolution.
- **Hallucinations:** Incorrect or misleading outputs generated by AI models, often caused by insufficient training data or biases. These can lead to unreliable information and potentially harmful consequences.
- **Best practices:** Preprocess data rigorously, normalize features, and validate with a variety of data sources.

Alternatives Assessment

A process for identifying and comparing potential chemical, material, product or other alternatives that can be used as substitutes to replace chemicals of high concern.



NAS framework



Steps for Transitioning to Safer Chemicals

1 Engage
2 Inventory & Prioritize
3 Identify
4 Assess & Compare
5 Select
6 Test
7 Evaluate

OSHA guidance

Is it safer?

- Workers
- Community
- Customers
- Environment

Is it effective?

- Performance standards
- Sufficiency
- Impact on quality

Is it affordable?

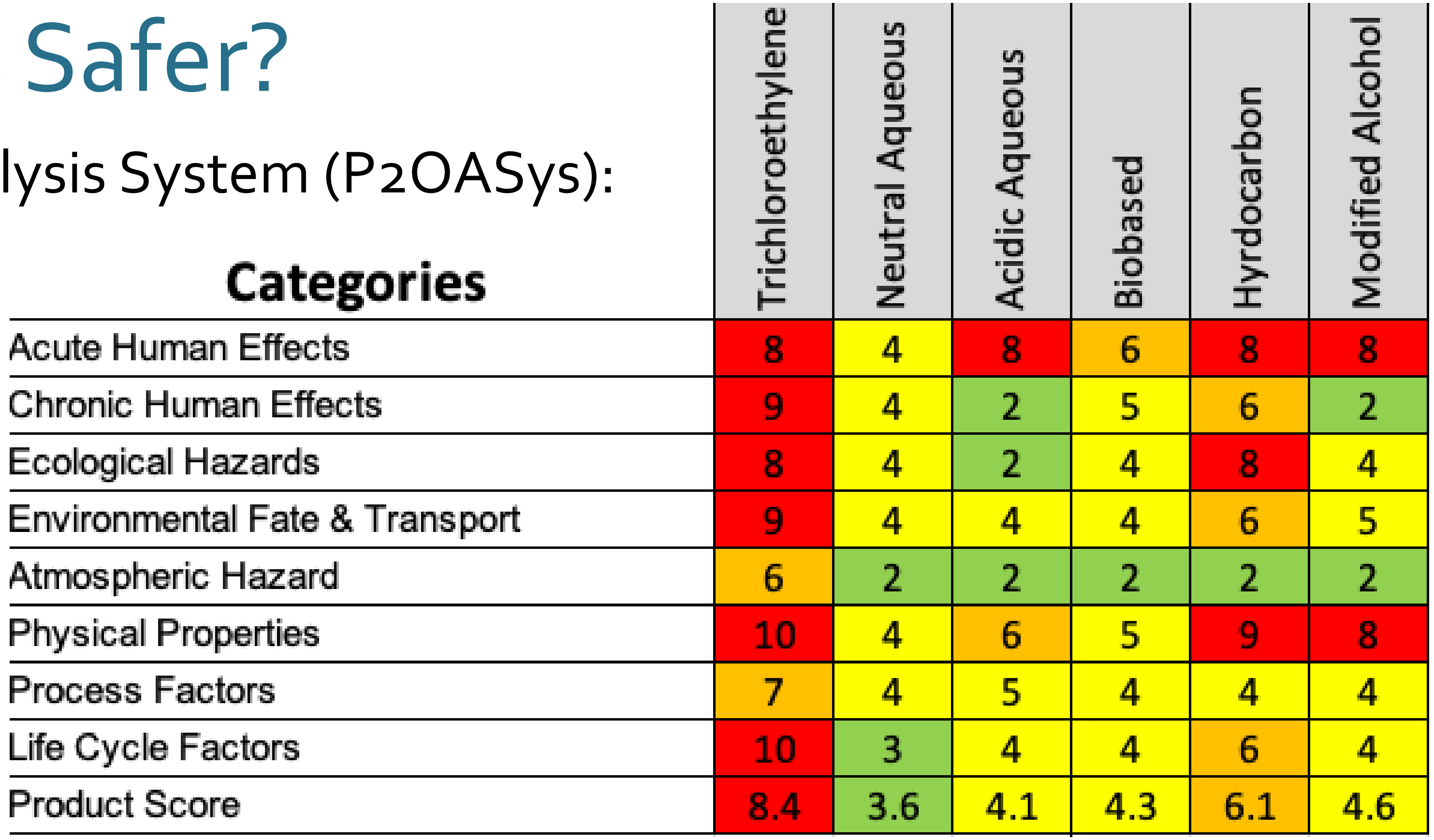
- Capital availability
- Ancillary costs
- External costs

TURI
TOXICS USE REDUCTION INSTITUTE

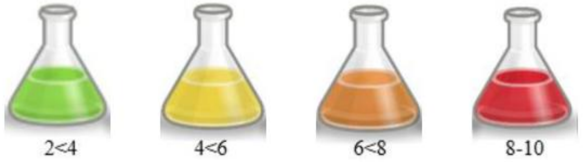
Are the Alternatives Safer?

Pollution Prevention Options Analysis System (P2OASys): https://p2oasys.turi.org

Compares potential Environmental Health and Safety (EHS) hazard categories:



Categories	Trichloroethylene	Neutral Aqueous	Acidic Aqueous	Biobased	Hyrdocarbon	Modified Alcohol
Acute Human Effects	8	4	8	6	8	8
Chronic Human Effects	9	4	2	5	6	2
Ecological Hazards	8	4	2	4	8	4
Environmental Fate & Transport	9	4	4	4	6	5
Atmospheric Hazard	6	2	2	2	2	2
Physical Properties	10	4	6	5	9	8
Process Factors	7	4	5	4	4	4
Life Cycle Factors	10	3	4	4	6	4
Product Score	8.4	3.6	4.1	4.3	6.1	4.6



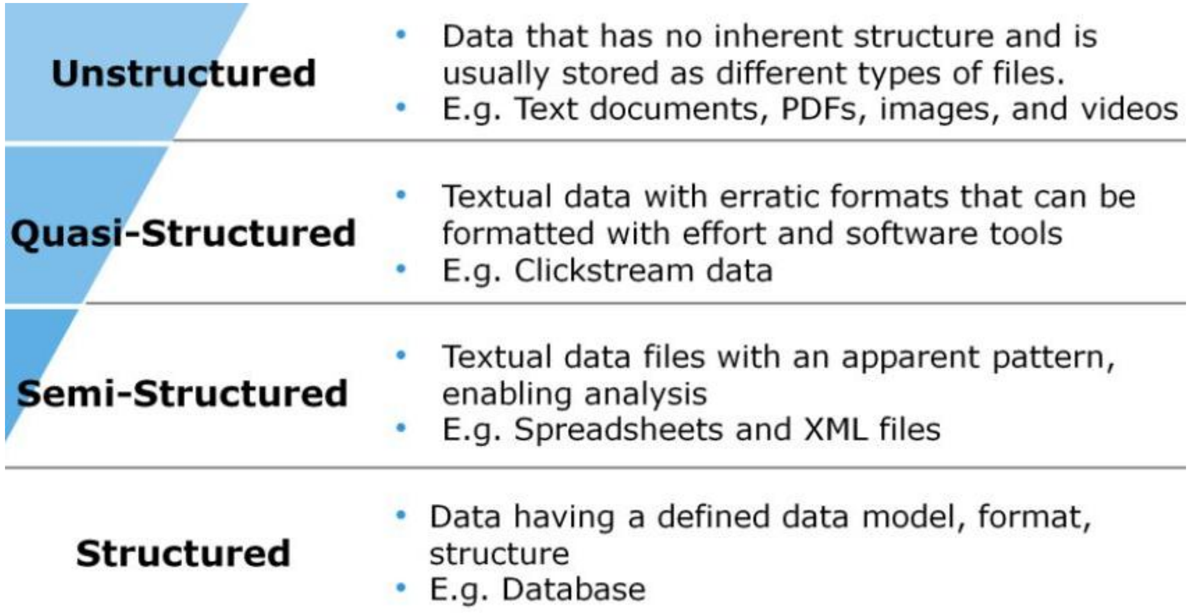
- Both quantitative data and qualitative input.
- Each category is rated using values, key phrases, GHS classifications, or other hazard designations.
- Depend upon available data (SDS, PubChem, ***computational toxicology***).

TURI
TOXICS USE REDUCTION INSTITUTE

AI4TUR: Document parsing

Natural Language Processing (NLP):

- AI models trained to understand the structure and information from text documents.
- Parsing: automatically extract data from SDS, webpages, and other hazard references.
- Why would AI be necessary? PDF documents are considered unstructured data.
- ***Why is it useful? Saving time gathering hazard information (e.g.: extract and update H-codes)***.
- *Risks of hallucination, important to supervise.*



Unstructured

- Data that has no inherent structure and is usually stored as different types of files.
- E.g. Text documents, PDFs, images, and videos

Quasi-Structured

- Textual data with erratic formats that can be formatted with effort and software tools
- E.g. Clickstream data

Semi-Structured

- Textual data files with an apparent pattern, enabling analysis
- E.g. Spreadsheets and XML files

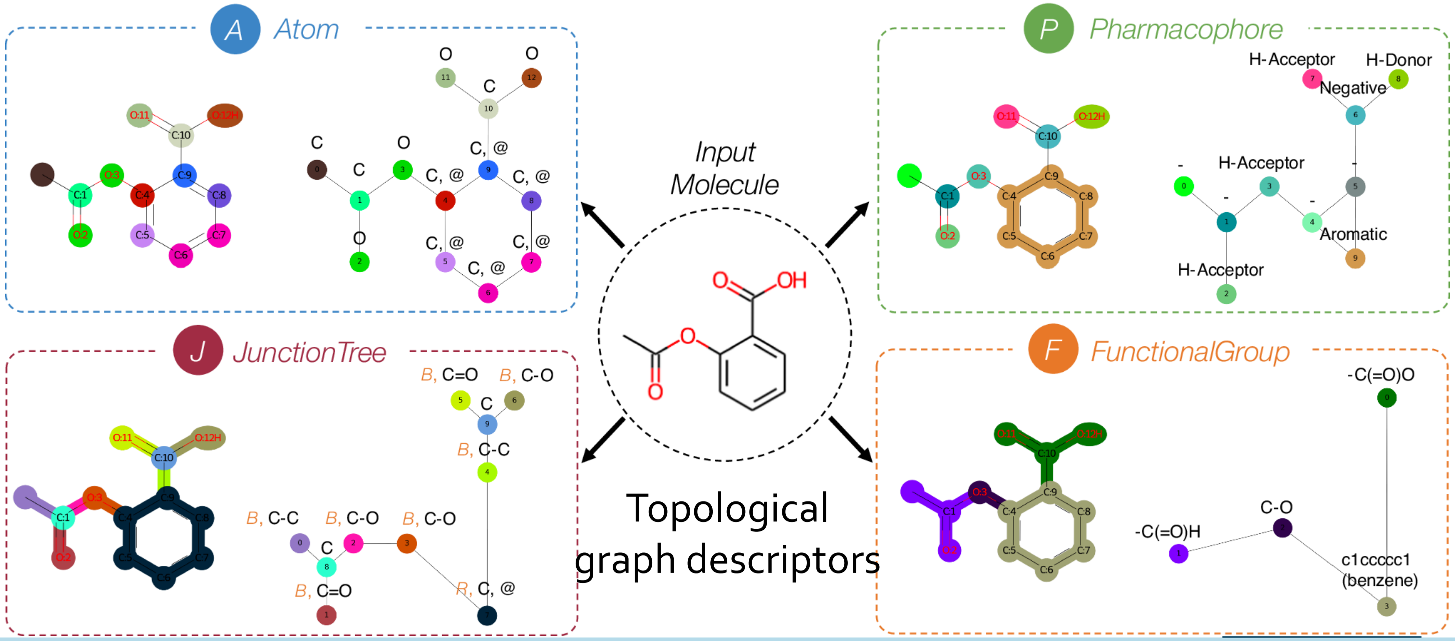
Structured

- Data having a defined data model, format, structure
- E.g. Database

- Challenges for automatic data lookup from scientific journals:
  - Scientific jargon: Scientific journals use specialized vocabulary and sentence structures.
  - Document format: Information can be scattered across sections.
  - Variability in document formats and writing styles.

Chemometrics: Statistics vs AI

- Traditionally, **Partial Least Squares (PLS)** fitting and **Principal Component Analysis (PCA),** identifying key variables, dominated low dimensional interpretable modeling.
- **Deep learning** can handle continuous and categorical data combined, but as an opaque ('black box') regarding how they arrive at their understanding.
- **Variational Autoencoders (VAEs):** A type of deep learning architecture that learns by deconstructing and reconstructing data through vectorization.
- **Support Vector Machines (SVMs):**
  - Identify clusters, constellations of molecules organized by motifs, or descriptors.
  - Analyze relationships between descriptors and properties.
  - Drug discovery.
  - QSAR decision trees.

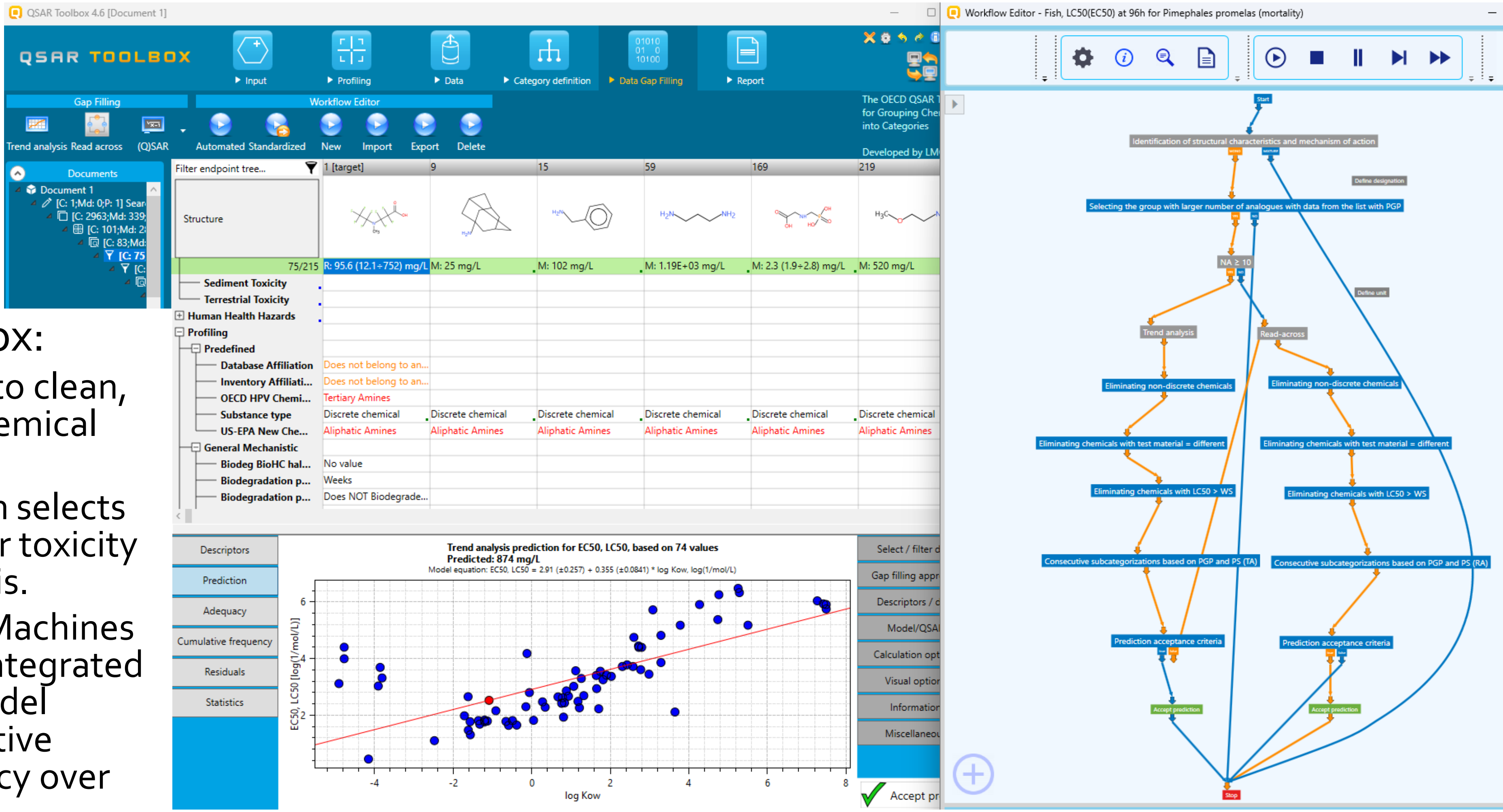


Topological graph descriptors

AI4TUR: Hazard data gap filling

Quantitative Structure-Activity Relationship (QSAR)

- OECD/EChA QSAR toolbox:
  - Non-AI Rule-based routines to clean, normalize, and transform chemical data into a suitable format.
  - Non-AI Rule-based algorithm selects closest structural relatives for toxicity read across and trend analysis.
  - AI/ML (e.g., Support Vector Machines or Random Forests) can be integrated by the user externally for model fitting, but QSAR toolbox native design prioritizes transparency over black-box methods.



QSAR Toolbox 4.6 [Document 1]

Workflow Editor - Fish, LC50(EC50) at 96h for Pimephales promelas (mortality)

Filter endpoint tree...	1 [target]	9	15	59	169	219
Structure						
75/215	R: 95.6 (12.1÷752) mg/L	M: 25 mg/L	M: 102 mg/L	M: 1.19E+03 mg/L	M: 2.3 (1.9÷2.8) mg/L	M: 520 mg/L
Sediment Toxicity						
Terrestrial Toxicity						
Human Health Hazards						
Profiling						
Predefined						
Database Affiliation	Does not belong to an...					
Inventory Affiliati...	Does not belong to an...					
OECD HPV Chemi...	Tertiary Amines					
Substance type	Discrete chemical	Discrete chemical	Discrete chemical	Discrete chemical	Discrete chemical	Discrete chemical
US-EPA New Che...	Aliphatic Amines	Aliphatic Amines	Aliphatic Amines	Aliphatic Amines	Aliphatic Amines	Aliphatic Amines
General Mechanistic						
Biodeg BioHC hal...	No value					
Biodegradation p...	Weeks					
Biodegradation p...	Does NOT Biodegrade...					

Trend analysis prediction for EC50, LC50, based on 74 values
Predicted: 874 mg/L
Model equation: EC50, LC50 = 2.91 (±0.257) + 0.355 (±0.0841) * log Kow, log(1/mol/L)

The OECD QSAR T
for Grouping Che
into Categories

Developed by LM

AI4TUR: Hazard data gap filling

- ML based Quantitative structure-property relationship (QSPR) models accurately forecast solubility parameters from physical properties.
- Quantitative property-consequence relationship (QPCR) can estimate ignition features (MIE, vapor cloud dimensions and concentrations).

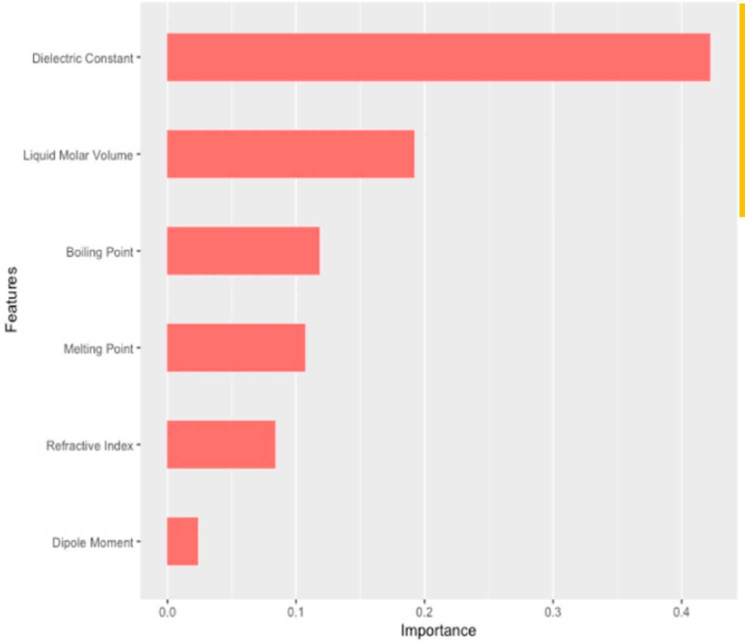
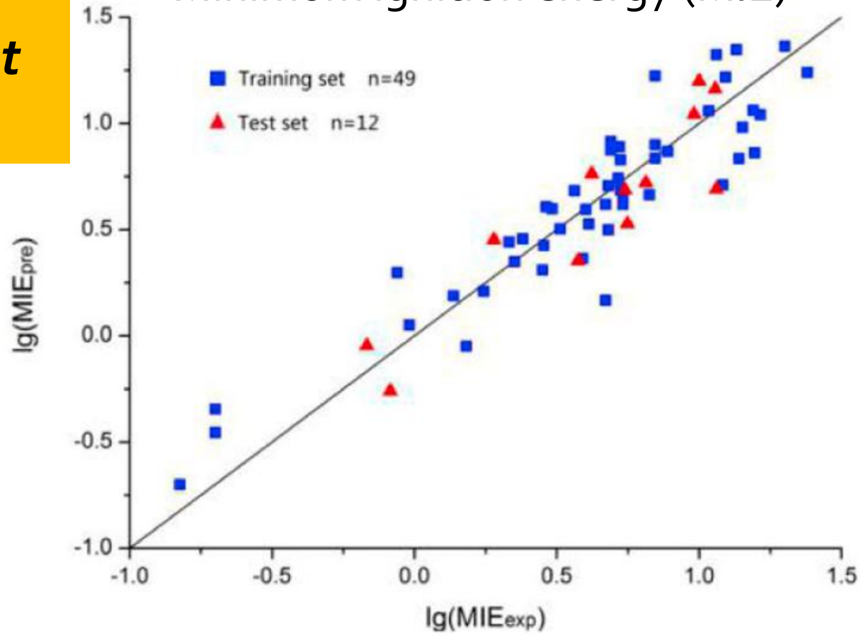


Figure 5. Importance of the solubility prediction

Hu et al.: Ind.Eng.Chem.Res.2021,60,11627–11635

Physics informed descriptors are most data efficient.

Minimum ignition energy (MIE)



Wang et al.: Ind. Eng. Chem. Res. 2017, 56, 47–51

TURI TOXICS USE REDUCTION INSTITUTE

HSPiP vs MIT ASKCOS for chemical property prediction

Feature	HSPiP	ASKCOS
Focus	Hansen Solubility Parameters	Synthesis optimization
Capabilities	Solubility predictions, polymer-solvent interactions, integration with qualitative experimental data	Solubility calculation from quantum descriptors, process optimization, integration with quantitative experimental data
Output	Comprehensive solubility parameter database, physical property predictions (boiling point, vapor pressure, melting point, critical point, viscosity, surface tension, refractive index …)	Solubility and additional features (NMR peaks, buyable look-up, retrosynthesis, forward synthesis, dipolar moment, quantum parameters)
Algorithms	Regression models, classification, Euclidean similarity, ***Genetic algorithms***	Genetic algorithms, simulated annealing, and gradient-based, deep learning methods
Data Requirements	Moderate	Much higher due to AI functionalities
Computer Power Usage	Any computer can do	Runs better in Ubuntu on a CPU>=4cores, needs RAM>=32GB

https://www.hansen-solubility.com/HSPiP/ https://askcos.mit.edu/

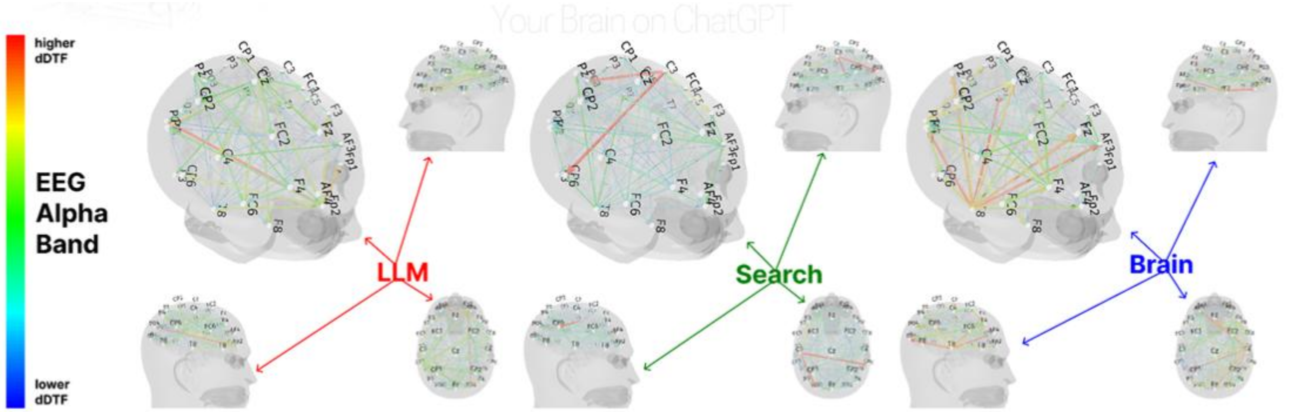
TURI FY26 Research Grant:
ML-Powered Solvent Alternatives Search at WPI

- Kickstarted in September 2025.
- Objective: Develop ML tool to identify non-toxic, low-cost solvents replacing NMP (TURA chemical) for dissolving PVDF in battery recycling.
- 3-pronged selection:
  - ML prediction of HSP and grid search of solvent blends (up to 3 components).
  - Evaluation of GHS hazards (acute/chronic toxicity, flammability, reactivity).
  - Bulk pricing comparison.
- Research (PI: Prof. Dr. Michael Tymko; Student: Muntasir Shahabuddin):
  - ML on molecular descriptors for HSP/solubility prediction.
  - Exhaustive screening of blends.
  - Lab Validation: High throughput turbidity experiments (modified 3D printer for dispensing precise blends).
- ***Broader Impact: Generalizable framework to accelerate any safer and affordable single solvent and solvent blend identification and selection.***

TURI
TOXICS USE REDUCTION INSTITUTE

Emerging AI Pitfalls: Brain Drain & Bust Pilots

- ***Cognitive Sedentarism*** (Kosmyna et. al, 2025):
  LLM overuse weakens neural connectivity and leads to "cognitive debt": Poorer memory recall (83% fail recent quotes), lower ownership of outputs, and skill atrophy over 4 months of AI over-reliance.



- ***AI Failure Rate*** (MIT NANDA, 2025):
  95% of genAI investments are yielding zero ROI: data/contextual gaps, scaling brittleness, and strategy mischiefs (e.g., overhyping sales tools vs. back-office wins).
  Real hits: Fast food companies ditched error-prone voice AI;
  e-commerce companies rehired after AI's "empathy" voids tanked service.

- ***Take home message:*** Thoughtful human oversight is essential for a reliable development safer alternatives using AI.

AI regulations: US vs. EU approaches

Feature	EU AI Act	US Regulations
Approach	Risk-based	Sector-specific, principles-based
Scope	Comprehensive, covering various AI applications	Primarily focused healthcare, autonomous vehicles
Risk Levels	Unacceptable, high, limited, minimal	No formal risk classification
Enforcement	Centralized authority	Multiple federal agencies, state-level regulations
Focus	Preventing harm, protecting human rights	Promoting innovation, addressing ethical concerns
Requirements	Risk assessments, transparency, accountability	Sector-specific guidelines, voluntary standards
Impact on Businesses	Strict compliance obligations	Varying requirements depending on sector

EU AI Act: centralized regulation categorizes AI systems into different risk levels:

- **Unacceptable Risk**
  - **Banned:** Posing a clear threat to fundamental rights.
  - **Examples:** Manipulative toys, social scoring, real-time remote biometric identification (with exceptions for law enforcement).
- **High Risk**
  - **Strict regulations:** AI systems impacting safety or fundamental rights.
  - Used in products under EU safety legislation (toys, cars, medical devices).
  - Involved in critical infrastructure, education, employment, essential services, law enforcement, migration, and legal interpretation.
  - Pre-market assessment, incident reporting, and consumer complaint rights required.
- **Limited Risk**
  - **Transparency obligations:** AI systems like generative AI (e.g., ChatGPT).
  - **Requirements:** Disclosure of AI-generated content must be clearly labeled, prevention of illegal content, and transparency about training data.
- **Minimal Risk**
  - **Minimal regulations:** Games, chatbots, spam filters, language translation…

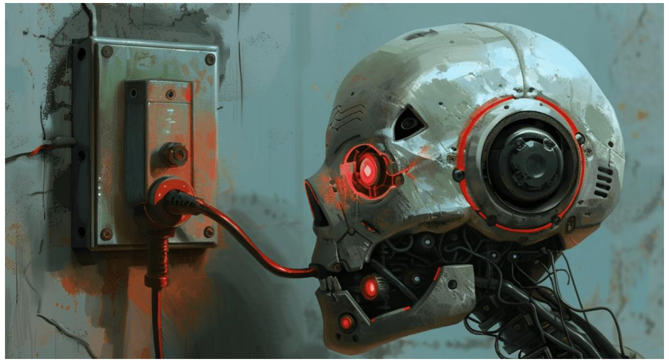
Data (and energy, and water) hunger games

Making an image with generative AI uses as much energy as charging your phone

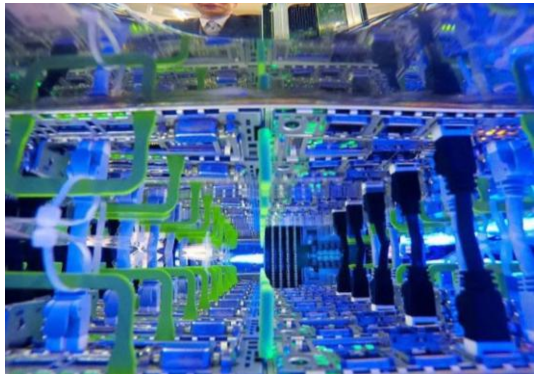
https://www.technologyreview.com/2023/12/01/1084189/making-an-image-with-generative-ai-uses-as-much-energy-as-charging-your-phone/

Data centers' electricity consumption in 2026 is projected to reach 1,000 terawatts, roughly Japan's total consumption.

https://e360.yale.edu/features/artificial-intelligence-climate-energy-emissions



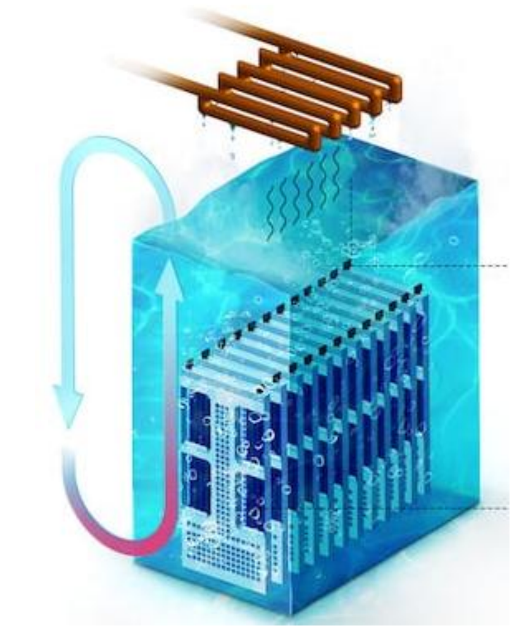
This is not water

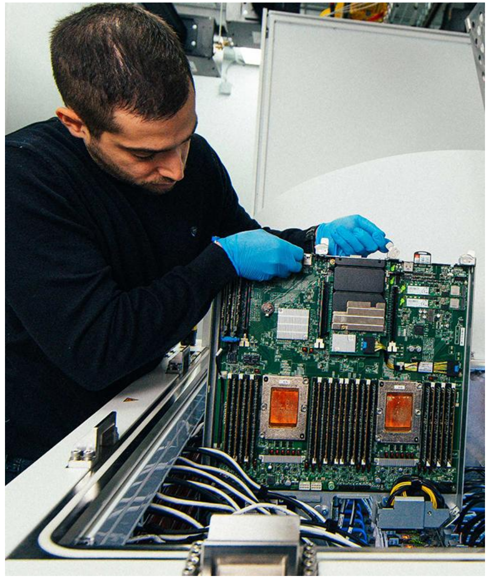


Water cooling of global AI may reach 4.2– 6.6 bn m^3 in 2027, which is more than the total annual water withdrawal of 4– 6 Denmark or half the UK

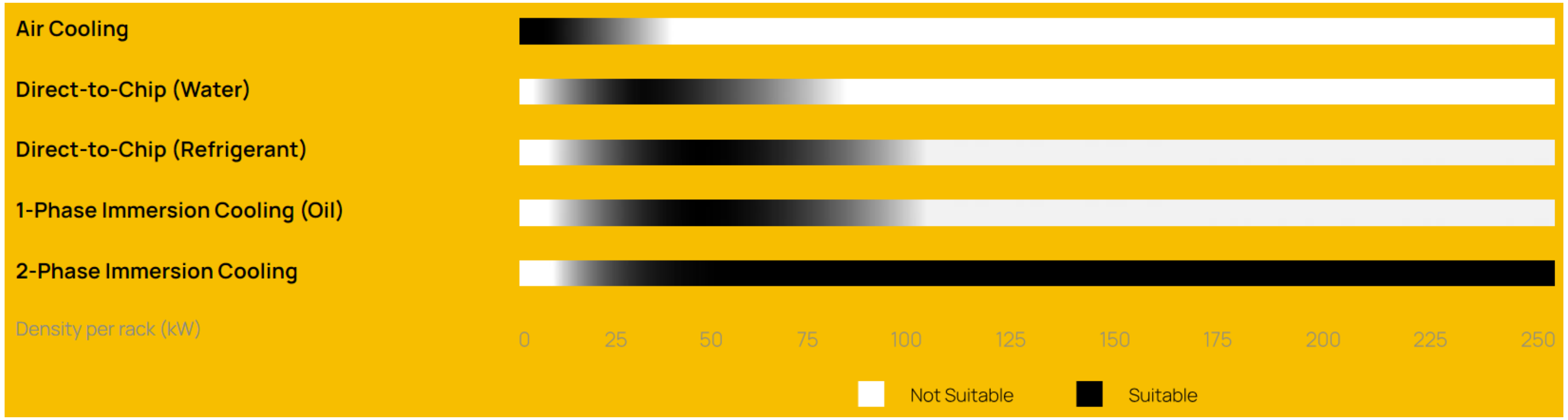
https://arxiv.org/abs/2304.03271

Boiling PFAS in the server room?

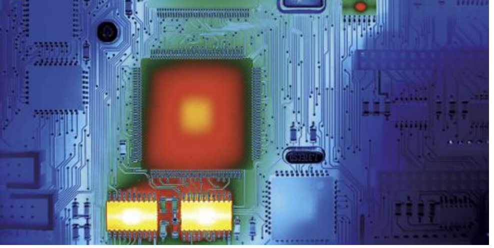




- Traditional air-cooling struggles with the high heat loads of modern High-Performance Computing (HPC) hardware.
- Two-phase cooling with fluorinated coolants offers efficient heat rejection.
  - **Benefit:** Highest heat rejection rate.
  - **Concern:** PFAS.
- Most non-PFAS volatiles are either flammable or material incompatible, making them unsuitable for two-phase cooling.
- Non-PFAS single-phase coolants require a different hardware configuration and might impact some user workflows (e.g., faster rack removal).



TURI
TOXICS USE REDUCTION INSTITUTE



HPC cooling strategies: PFAS-free tech available

$$PUE = \frac{Total\ energy}{Computing\ energy}$$

Cooling Method	Pros	Cons	Power Usage Effectiveness	Heat Rejection Rate
Air Cooling	Lower initial cost, Simpler implementation	Lower efficiency at high power densities, Increased energy consumption due to fan power	1.5 – 1.8	Up to 3 kW
Liquid Immersion Cooling	Higher efficiency, Reduced energy consumption, Better heat dissipation. PFAS-free.	Higher initial cost, Increased complexity, Potential for leaks	1.1 - 1.3	3 - 10 kW
Direct-to-Chip Cooling	Excellent heat removal. Minimal coolant use. PFAS-free options available	Very high initial cost, High complexity, Limited scalability	1.05 - 1.2	10 kW+
Two-Phase Cooling	High heat transfer capability	High complexity, higher potential for leaks, cavitation due to bubbling, PFAS reliance	1.1 - 1.2	5 - 200 kW

Sources:

DOE - Best practices Guide for Energy Efficient Data Center Design

Data Center Cooling Trends for 2025

Yuan et al., Energy and Buildings, 2021

Electronics Cooling

Immersion Cooling of Electronics in DoD Installations

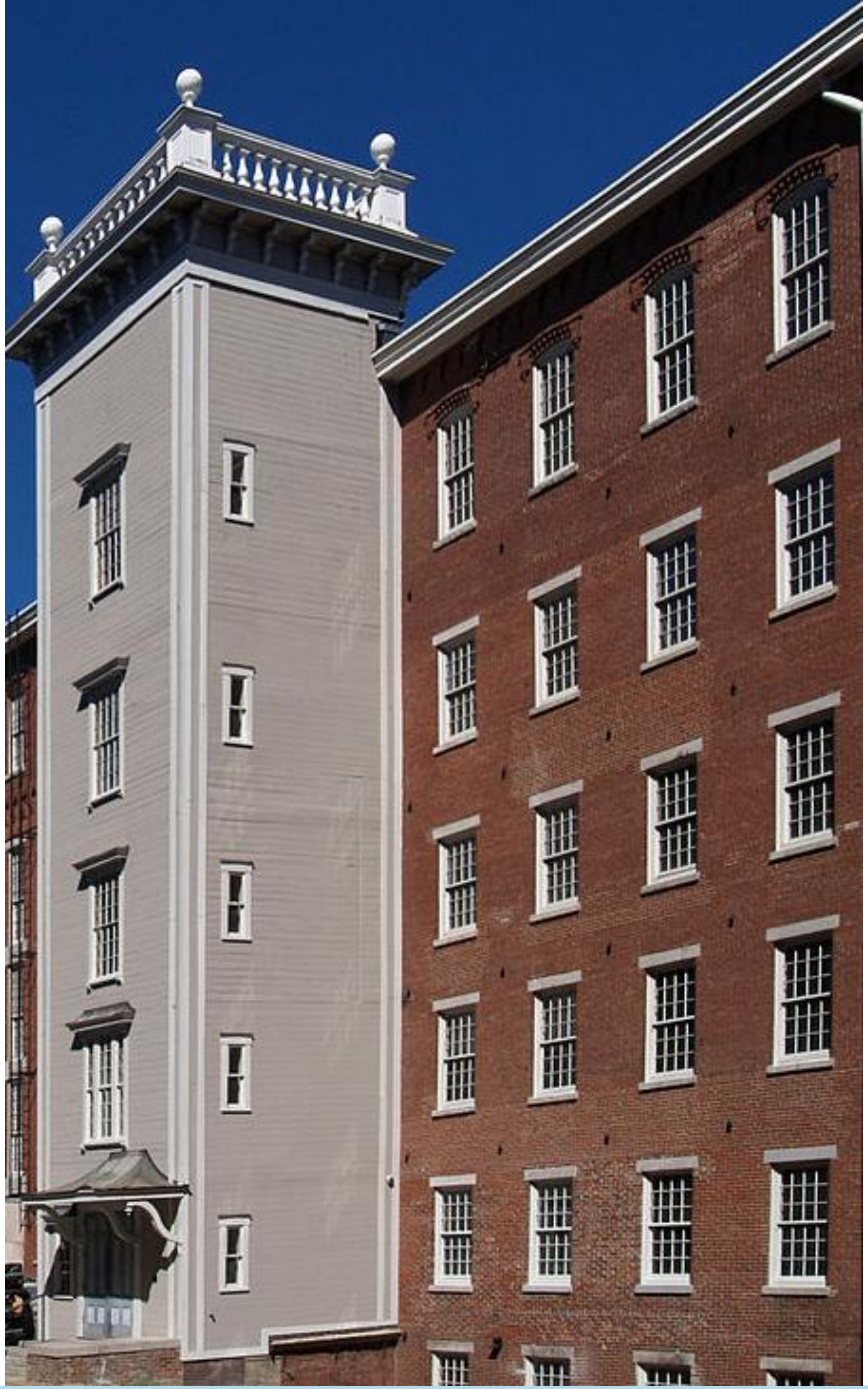
Alissa et al. Nature 641, 331–338 (2025)

TURI
TOXICS USE REDUCTION INSTITUTE

Opportunities and Challenges in AI4TUR

- AI helps TUR by:
  - Bridging data gaps (e.g.: QSAR predictions; document parsing).
  - Streamlining alternatives assessment.
- However,
  - Interpretable models are essential to ensure reliability.
  - AI relies on data centers that consume an enormous amount of energy and water.
  - Some data centers rely on PFAS for primary cooling.
  - Over-reliance can lead to cognitive sedentarism.

TURI
TOXICS USE REDUCTION INSTITUTE



TURI
TOXICS USE REDUCTION INSTITUTE

Thank you!

Visit our website www.turi.org for **free publicly available d**atabases, tools, and case studies:

- www.Cleanersolutions.org
- https://P2OASys.turi.org
- www.TURAdata.org
- https://www.turi.org/Our_Work/Resources

Contact us!

gregory_morose@uml.edu
gabriel_salierno@uml.edu
alicia_mccarthy@uml.edu
info@turi.org

Follow us:

@reducingtoxics