National Institute of Environmental Health Sciences
Division of Translational Toxicology

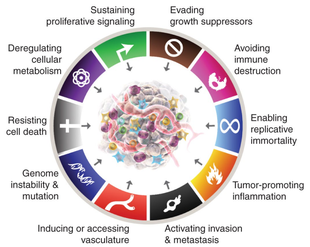
Deep Learning Profile QSAR Modeling to Impute In Vitro Assay Results and Predict Chemical Carcinogenesis Mechanisms

A. Borrel[1], A.L. Karmaus[1*], G. Tedla[1], K. Mansouri[2], T. Luechtefeld[3], R. Lunn[4], A. Wang[4], D.G. Allen[1], N. Kleinstreuer[2]

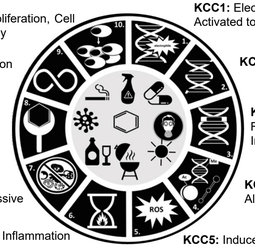
[1]Inotiv, RTP, NC; [2]NIH/NIEHS/DTT/NICEATM, RTP, NC; [3]Insilica LLC, Bethesda, MD; [4]NIH/NIEHS/DTT/IHAB, RTP, NC

1. How Does a Normal Cell Become a Cancer Cell?

Carcinogenesis is a multi-step process in which normal cells are transformed into cancer cells by acquiring properties that allow them to form tumors or malignant cancers. These properties, which distinguish cancer cells from normal cells, have been classified as a series of 10 Hallmarks of Cancer (HMC) (Hanahan, 2022).



The World Health Organization's International Agency for Research on Cancer (IARC) and collaborators have identified and extensively characterized the mechanisms of a set of carcinogens from their monograph program. By focusing specifically on these chemicals, they have defined a set of Key Characteristics of Carcinogens (KCC) (Smith et al., 2016).



KCC10: Alters Cell Proliferation, Cell Death or Nutrient Supply

KCC1: Electrophile or Can Be Activated to Electrophiles

KCC9: Causes Immortalization

KCC2: Genotoxic

KCC8: Modulates Receptor-mediated Effects

KCC3: Activates Mutagenic DNA Repair & Promotes Genomic Instability

KCC7: Immunosuppressive

KCC4: Induces Epigenetic Alterations

KCC6: Induces Chronic Inflammation

KCC5: Induces Oxidative Stress

There is no mapping between HMC and KCC, and a carcinogen can have several KCC and exhibit several HMC.

2. Carcinogenicity Modeling Challenges

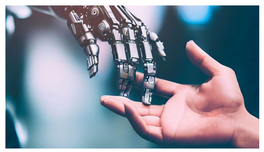
Developing a predictive model for carcinogenicity is challenging because the model should:

- Not be limited to a unique prediction (carcinogenicity overall) but be able to predict several KCC at once.
- Account for interaction among mechanisms/targets.
- Have the ability to combine diverse data type (in vitro assays, molecular descriptors, etc.) to cover most of the mechanisms involved in the KCC.
- Perform effectively with sparse data sets.

Current models available for carcinogenicity, which are based on quantitative structure–activity relationships (QSARs), are limited to a specific type of carcinogen, e.g., liver carcinogens (Li et al., 2021), or are focused on one KCC, such as genotoxicity (Toma et al., 2020).

3. Can Modern AI Help Build a Better Model?

- In this project, we leveraged modern artificial intelligence (AI) techniques utilizing in vitro data to predict chemical carcinogenicity.
- Our approach involved imputing data from high-throughput ToxCast/Tox21 assays to create a carcinogenicity profile for each chemical based on a scoring system by KCC.
- This modeling allows us to incorporate a large amount of data that cannot be handled in a classic QSAR modeling.

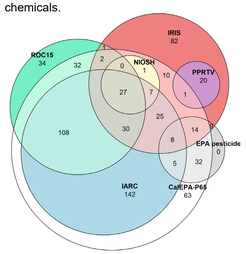


3. Sets of Carcinogens and Noncarcinogens

We compiled collections of carcinogen classifications from U.S. and international organizations including:

- National Toxicology Program Report on Carcinogens (ROC)
- U.S. Environmental Protection Agency (EPA) programs:
  - Integrated Risk Information System (IRIS)
  - Provisional Peer-Reviewed Toxicity Values (PPRTV; Superfund Program)
  - Pesticide Program (EPA pesticide)
  - The Proposition 65 List at California EPA (CalEPA-P65)
- National Institute for Occupational Safety and Health (NIOSH)
- International Agency for Research on Cancer (IARC)

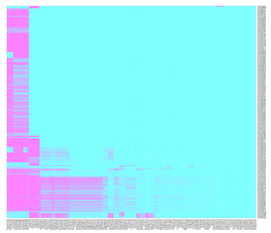
The carcinogen set only included chemicals with clear evidence of being human carcinogens, such as IARC Group 1 chemicals.



771 Carcinogens
496 included in the Tox21 chemical library

4. Tox21 Program Assays

The U.S. federal interagency Tox21 program has tested approximately 10,000 chemicals in up to 2000 assays to gain mechanistic insights into chemical toxicity. The challenge in using this large data set for modeling is that the data are notably skewed toward compounds with no activity (Richard et al., 2021).



No tested data
Tested data

The skewed data presents a significant challenge for conventional QSAR modeling.

5. Assay Mapped to the KCC

We mapped Tox21 assay data to KCCs using assay gene targets and expert opinion. The mappings are available within NICEATM's Integrated Chemical Environment.

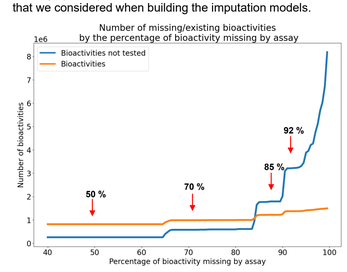
Integrated Chemical Environment

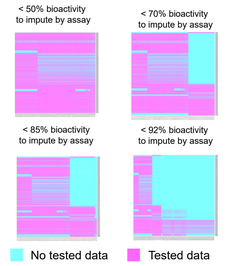
https://ice.ntp.niehs.nih.gov/

KCC	Number of Assays Mapped
KCC2: Genotoxic	17
KCC3: Activates Mutagenic DNA Repair & Promotes Genomic Instability	3
KCC5: Induces Oxidative Stress	14
KCC6: Induces Chronic Inflammation	48
KCC8: Modulates Receptor-mediated Effects	142
KCC10: Alters Cell Proliferation, Cell Death or Nutrient Supply	204

6. Iterative Imputation

We initially examined the amount of data available in ToxCast/Tox21 by looking at the percentage of bioactivity missing by assay versus the number of bioactivities available in the whole dataset. We identified four plateaus that we considered when building the imputation models.





No tested data
Tested data

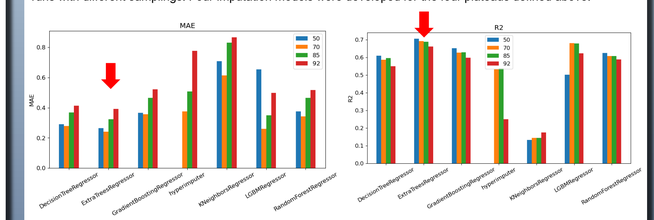
We developed a regressor iterative imputer for four subsets of the ToxCast/Tox21 assays, one for each plateau, using seven types of machine learning models and molecular descriptors. The models were run on the National Institutes of Health's Biowulf high-performance computing server, with each run allocated 32 CPUs, 100MB of memory, and 100 hours of computation.

BIOWULF AT THE NIH

https://hpc.nih.gov/

7. Imputation Performance on Regression

Performance results are presented below on a test set that included 15% of the available AC50 data, randomly chosen, utilizing the average of Mean Absolute Error (MAE) and the average of the R-Squared for five repeated runs with different samplings. Four imputation models were developed for the four plateaus defined above.



In general, the ExtraTreeRegressor performed the best. Increasing the amount of data to impute up to 92% by assay did not drastically decrease the performance.

8. KCC Scores

Summary scores (KC_{score}) for each KCC were computed using ToxPi by first finding the normalized proportion of assays in which a chemical was active for a given KCC, and then scaling the value based on the lowest activity concentration among those positive responses.

$$s_i = \frac{(number\ of\ positive\ responses)}{(maximum\ number\ of\ postive\ responses\ for\ that\ KCC)}$$

$$f = 1 - cdf(minimum\ positive\ response\ value\ of\ that\ chemical)$$

$$KCC_{score} = s_i + \frac{f}{3} * (1 - s_i)$$

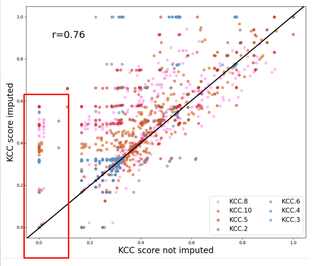
s_i: normalized proportion of active assays for each chemical specific to an individual KCC.
cdf: percentile of the minimum positive response value for each chemical relative to all other positive response values across all chemicals active against assays mapping to that KCC.

This scoring equation ensured that both potency and frequency were factored in and the values remained between 0 and 1. Dividing the concentration factor by 3 ensured a balance between the contribution of potency and the contribution of activity across multiple targets. A higher score characterizes more positive assays mapped on the KCC, as well as a lower AC50 for the assays mapped.

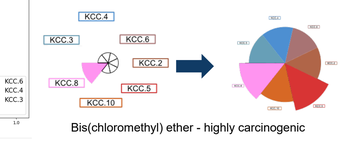
Example of KCC profile for lindane, a carcinogen tested in most of the assays mapped on KCC. Each KCC has a score between 0 to 1.

9. KCC Scores for Carcinogens

Comparison of the KCC_{scores} calculated from the imputed and not imputed Tox21/ToxCast results with the best model.



- Good correlation between the KCC_{scores} from imputed and actual data gave confidence that we were not drastically changing the KCC score with the imputed data.
- We completed the carcinogenic profile by computing a KCC_{scores} for all of the KCC even when no data were present (red box).



Bis(chloromethyl) ether - highly carcinogenic

10. Build a Profile Not Limited to the ToxCast/Tox21 data

Modern iterative imputers allow us to integrate more data mapped on each KCC to build the most robust imputation model and carcinogenic profile. We are using BioBricks.ai to bring more data into the modeling.

A Bioinformatics Data Registry
Import data-dependencies for your own projects with a single line of code. Use common data-science tools to analyze 40+ life science databases. Deploy your own databases or machine learning models to the platform.

https://biobricks.ai/

We are developing a Cancer Harmony Repository, where we consolidate pertinent data for carcinogenicity modeling from various databases, including ChEMBL, Gene Ontology, and the Kyoto Encyclopedia of Genes and Genomes (KEGG). You can access the repository under development on GitHub at **https://github.com/biobricks-ai/cancerharmony**.

Conclusion and Future Directions

In this project, we developed:

- A set of carcinogens based on available regulatory datasets.
- A competitive carcinogenicity model that can take into consideration all 10 KCC at once, based on iterative imputation modeling.

Next, we are working on:

- Further validation of the results, such as using literature evidence.
- Evaluating confidence of the imputation modeling.
- Improving the KCC scoring in order to take into consideration more data mapped on a specific KCC.
- Completing the assay mapping on the KCC using new sources of data with biobricks.ai.

References

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov 12(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059
Li, T., et al. (2021). DeepCarc: Deep Learning-Powered Carcinogenicity Prediction Using Model-Level Representation. Front Artif Intell 4. https://doi.org/10.3389/frai.2021.757780
Toma, C., et al. (2020). QSAR Models for Human Carcinogenicity: An Assessment Based on Oral and Inhalation Slope Factors. Molecules 26(1). https://doi.org/10.3390/molecules26010127
Richard, A. M., et al. (2020). The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. Chem Res Toxicol 34(2), 189–216. https://doi.org/10.1021/acs.chemrestox.0c00264
Smith, M. T., et al. (2016). Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. Environmental Health Perspectives, 124(6), 713–721. https://doi.org/10.1289/ehp.1509912

Acknowledgements

Authors thank the National Institutes of Health high-performance computing resources that run the imputation models.
This project was funded in part with federal funds from the NIEHS, NIH under Contract No. HHSN273201500010C. The authors declare there exists no conflict of interest.
*A.L. Karmaus is currently affiliated with Syngenta, Greensboro, NC