ED-Screen
User's guide

About the application **2**
Insight on Chemoinformatics components **2**
Estrogen receptor relative binding affinity model (IRFMN) 2
Estrogen Receptor-mediated effect (IRFMN-CERAPP) 2
Androgen Receptor-mediated effect (IRFMN/COMPARA) 2
Thyroid peroxidase inhibition activity (Oberon) 3
Transthyretin binding model (not available in VEGA yet) 3
Thyroid receptor alpha activation model (VEGA/NRMEA) and Thyroid receptor beta activation model (VEGA/NRMEA) in VEGA 3
DIO1 inhibition (IRFMN/ED-Screen) 3
Steroidogenesis perturbation (OBERON/ED-Screen) 3
Insight on Bioinformatics components **4**
Models 4
Input Preparation 4
How to install and run the tool **5**
Windows users 5
Linux users 5
How to use the tool **5**
References and acknowledgments **10**

About the application

The ED-SCREEN is a software enabling the prioritization of lists of chemicals as a function of their propensities to disrupt EATS modalities. The only input required by the software are SMILES strings (and/or transcriptomics data if available) for the given compounds. The propensities for disruption will be predicted by chemoinformatics approaches (i.e., QSAR models) and bioinformatics approaches (i.e., QGexAR models).

The results will be provided in the form of scores for Estrogen (PS-E), Androgen (PS-A), Thyroid (PS-T) and Steroidogenesis (PS-S) activity. These scores summarize the perturbation potential, and range between -1 (lack of endocrine interference sustained by experimental evidence) to +1 (experimental evidence for interference).

These scores are used to derive a single overall score i.e. the ED index, indicating the perturbation potential for all macro-targets combined. It ranges between 0 (lack of any endocrine interference sustained by experimental evidence) to +1 (experimental evidence for interference).

Insight on Chemoinformatics components

The chemoinformatics models (QSARs) implemented in the chemoinformatics part of the calculation workflow are the following:

Estrogen receptor relative binding affinity model (IRFMN)

The model is a QSAR classification model based on a classification and regression tree (CART) based on Dragon descriptors (Roncaglioni, 2008).

Estrogen Receptor-mediated effect (IRFMN-CERAPP)

The model provides a qualitative prediction for Estrogen Receptor-mediated effect classification model for endocrine disruptor screening. The model has been built as a set of rules, extracted with Sarpy software (59 structural alerts) from a dataset obtained from a collection of high-quality estrogen receptor (ER) signaling data (1529 chemicals) from the ToxCast program (Judson 2015). The model has been developed within the framework of the Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (Mansouri, 2016).

Androgen Receptor-mediated effect (IRFMN/COMPARA)

The model provides a qualitative prediction for Androgen Receptor (AR) effects mediated through the AR pathway. The data were used to generate binary classification models to discriminate active (both agonists and antagonists) compounds from inactive ones.

It is a two-steps model developed with SARpy. In the first step, SARpy was used to model the two classes, identifying a set of 127 rules (17 for active and 110 for inactive). Then, a second set of 22 rules identifying active compounds is applied to unpredicted compounds only. Compounds are labeled as inactive if matched either by an inactive SA in the first step or not matched by the latter step. More precisely: If a compound matches the active class

when applying the first ruleset the output is Active; If a compound matches the inactive class when applying the first ruleset the output is Inactive; If a compound matches the active class when applying the second ruleset the output is Probably Active; If a compound does not match any class when applying the first ruleset nor the second one the output is Probably Inactive

Thyroid peroxidase inhibition activity (Oberon)

The model is a QSAR classification model based on the k-NN algorithm and using Dragon descriptors (Gadaleta, 2021).

Transthyretin binding model (not available in VEGA yet)

The model is developed from a data set of 200 compounds, 68 positives and 132 negatives. The model is based on Gradient boosted tree algorithm and Dragon Descriptors.

Thyroid receptor alpha activation model (VEGA/NRMEA) and Thyroid receptor beta activation model (VEGA/NRMEA) in VEGA

These two models are the implementation of the NRMEA models inside VEGA to permit the use of the applicability domain index associated with the prediction.

NRMEA models are classification models based on Structural alerts. They are made using data collected from two open free libraries, including ToxCast/Tox21 (https://www.epa.gov/chemical-research/toxcast-dashboard) and ChEMBL (https://www.ebi.ac.uk/chembl/).

TRα model is developed on a data set of 6075 compounds, 99 agonists, 26 antagonists and 5950 negatives.

TRβ model is developed on a data set of 6101 compounds, 114 agonists, 37 antagonists and 5950 negatives.

DIO1 inhibition (IRFMN/ED-Screen)

DIO1 inhibition is a consensus model between a QSAR model based on Gradient boosted trees algorithms on 19 Dragon descriptors and a kNN model based on structural similarity. Data are taken from the ToxCast database and is composed of 1646 compounds, 209 positives and 1437 negatives.

Steroidogenesis perturbation (OBERON/ED-Screen)

Steroidogenesis perturbation model is a consensus model between a QSAR model based on Random Forest algorithms on 11 Dragon descriptors and a kNN model based on structural similarity. Data are taken from ToxCast database and EPA database. Data set is composed of 1547 compounds, 557 positives and 990 negatives.

Insight on Bioinformatics components

Models

The bioinformatics component consists of six *Qualitative Gene Expression Activity Relationship (QGexAR)* machine learning models. These models were developed for three endpoints related to endocrine disruption: estrogen receptor binding, androgen receptor binding, and thyroid receptor antagonism. Each endpoint is represented in two cell lines: MCF7 (breast cancer) and A549 (human alveolar basal epithelial adenocarcinoma).

The models use transcriptomics data associated with chemical exposure to predict the risk of endocrine (EAT) disruption.

The models were developed using transcriptomics data from the LINCS project (Subramanian et al., 2017), specifically the LINCS L1000 dataset. This project experimentally measured 978 informative transcripts, referred to as “landmark genes,” and includes over 1.3 million gene expression profiles covering 34,419 chemicals.

Transcriptomics data generated in MCF7 and A549 cells (tested at 10 µM for 24 hours) were cross-referenced with known endocrine-disrupting chemicals (EDCs) using the following resources: CERAPP (for estrogen receptor) (Mansouri et al., 2016), CoMPARA (for androgen receptor) (Mansouri et al., 2020), and the dataset from Gadaleta et al. (for thyroid receptor) (Gadaleta et al., 2021).

For more information on the QGexAR models, please refer to this article (Ollitrault et al., 2024).

When no transcriptomics data are available for a given compound, a KNN-QGexAR approach is applied. This method selects the six most structurally similar compounds (VEGA similarity) with available transcriptomics data in the LINCS database and bases the prediction on the mean probability of activity derived from these compounds’ transcriptomic profiles.

Please note that for KNN-QGexAR predictions, the applicability domain weight is set as the value of the similarity for the most similar compound (equal to 1 if the target compound is found in the training set, indeed the provided prediction is actually the experimental value).

Input Preparation

Transcriptomic Z-score profiles should be provided with compounds in rows and gene identifiers in columns. The columns must correspond to the 978 LINCS landmark genes (using LINCS-specific gene IDs, available here: https://github.com/guillaumeolt/ED-SCREEN_tool/blob/main/data/geneinfo_beta.txt).

Users may compute Z-scores independently using their own data and workflows, for example, using DESeq2-derived Z-scores (Love et al., 2014). However, it is essential to include all landmark genes. For genes that were not measured, a value of 0 can be assigned. Please note that the model performance was evaluated using LINCS data, and performance may differ when using other protocols.

How to install and run the tool

ED-Screen is a Docker based application, designed to work as a standalone tool on a local machine. To use it, you should have installed Docker (https://www.docker.com/) or, on Windows machines, Docker Desktop (https://www.docker.com/products/docker-desktop/).

Windows users

Before using the application, please run install.bat and wait until its termination (this will install on your machine all the needed Docker container images for the application).

To use the application, launch run.bat. This will start the application backend, and after that it should automatically open your default web browser with the ED-Screen interface. If this does not happen, please open your web browser and go to http://localhost:7968/. Please DON'T close the command line window opened by run.bat. To close the application, go in such window and type "Q" as reported, the application will be shut down.

Linux users

Install the images for the application with the following commands:

```
docker load -i eds_backend.tar
docker load -i eds_db.tar
docker load -i eds_frontend.tar
docker load -i eds_python.tar
```

When this is done, start the application with the following command from the application folder:

```
docker compose up -d
```

When the application will be correctly running, open any web browser and go to the local address http://localhost:7968

Please note that the application will be still running in the background even if you close your browser; to shut down the application run, again from the application folder:

```
docker compose down
```

How to use the tool

Once the application is running, the initial screen (Fig. 1) should be visible in the browser.

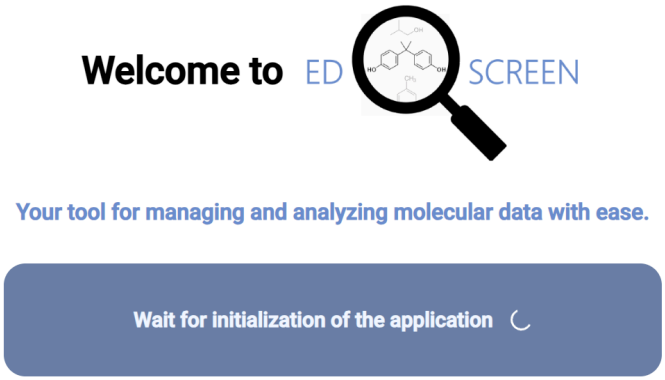


Figure 1. The waiting screen displayed until the application is ready

The application could require a few seconds before being ready, please wait until the "loading" button is ready, then click it. The "Molecule" section (Fig. 2) will be then displayed.

In this view is displayed the molecule list created by the user. At startup the list is empty, the user can add a single molecule, providing Molecule ID (a user defined identifier for a molecule) and the SMILES. The molecule is added to the list by clicking the "+" button.
To provide a SMILES there are two additional ways:

- "Draw molecule": it is possible to draw the molecule structure and, when the "Insert button" is pressed, the corresponding SMILES is reported in the input field (Fig. 3).
- "Search on PubChem": the user directly types the molecule name or CAS number in the input box and the application searches the corresponding SMILES on the public online PubChem database. That SMILES is shown and the user can choose to maintain or discard it. If maintained, the SMILES is reported in the input field.

Molecules can be imported from a local file with the import function (the green "up arrow" button). Files should be provided in .csv format, with two columns: molecule ID and SMILES. The names in the header are not relevant.
When the button is clicked the import window is shown, the user has to choose the file to be imported, the separator used in the file, if a header row is present, and if the imported molecules should replace the current data set or should be appended to it.

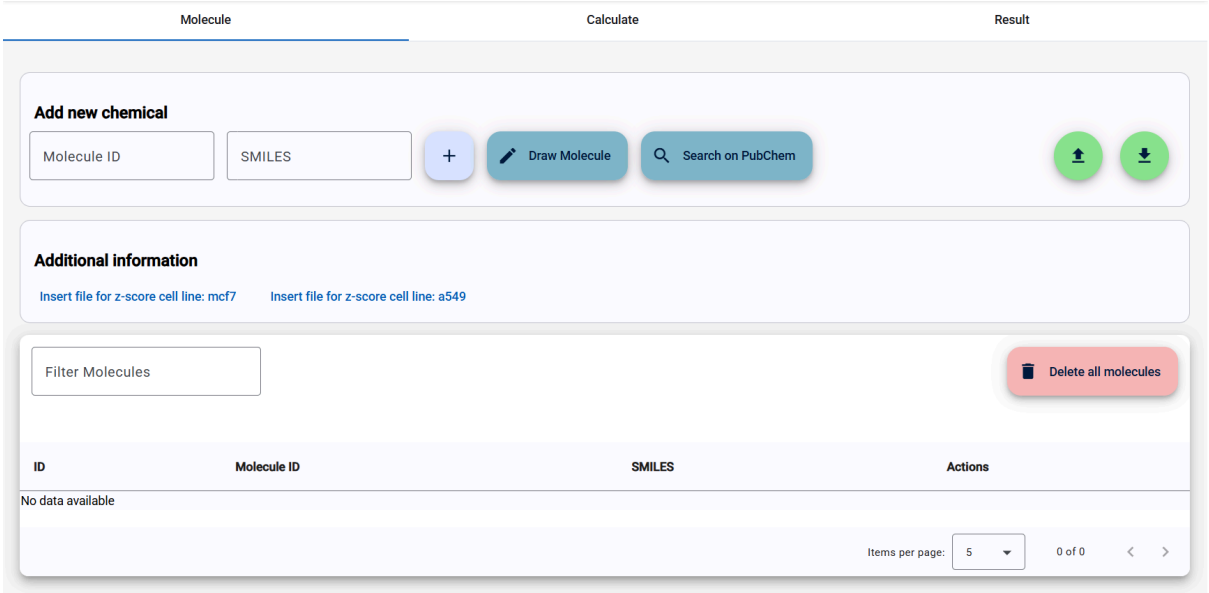


Figure 2. The Molecule page

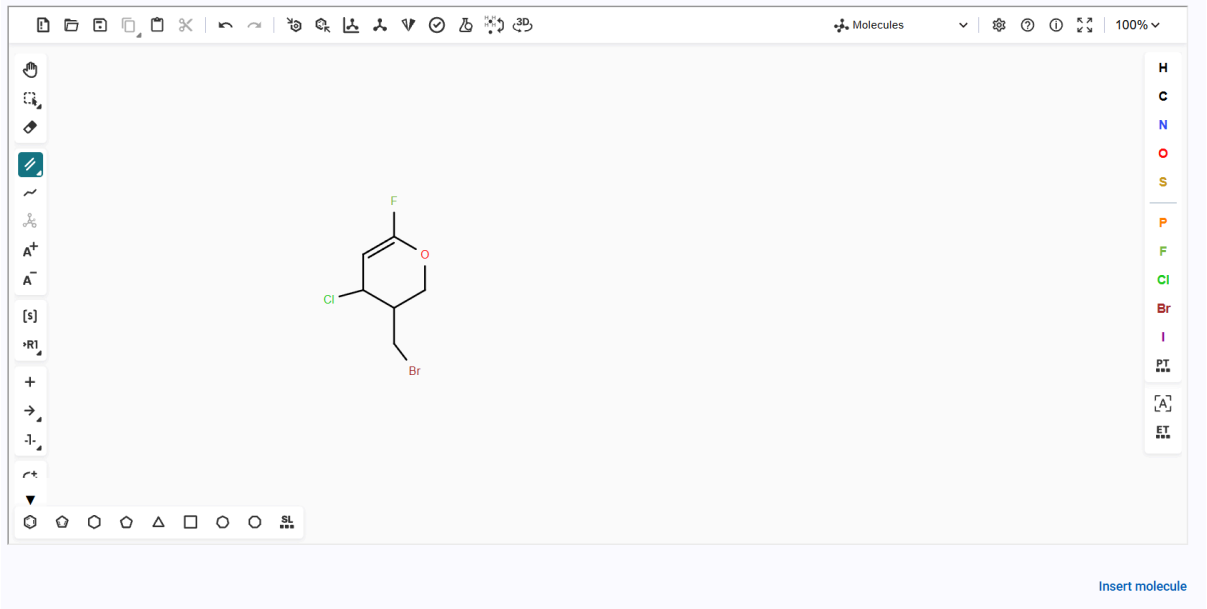


Figure 3. The drawer to design a molecule. In this tool can be loaded molecules previously drawn.

The green down-arrow button exports the current dataset of compounds as a csv file.
In the “Additional Information” panel there are two buttons, needed to load the z-score files in order to run the transcriptomic models; please refer to the “Insight on Bioinformatics components” for details about the format of the files to be used.
Each entry in the molecules table can be deleted with the red trash button; this will delete only the correspective row, while the red “Delete all molecules” button, at the top of the table, will delete all the molecules, subject to user confirmation.

The second page is about the list of models that can be selected for the computation. The models are divided by endocrine endpoints. If z-score files are not provided, the models that can't be calculated are automatically disabled. Figure 4 shows the page of models.
The "Start Calculation" button starts the computation of the models, please note that this operation could take a long time, depending on the number of molecules and selected models. During the calculation the computation can be stopped through the button in over impression.
When the calculation is performed, a message is displayed at the bottom of the page and the results will be available in the "Result" page.

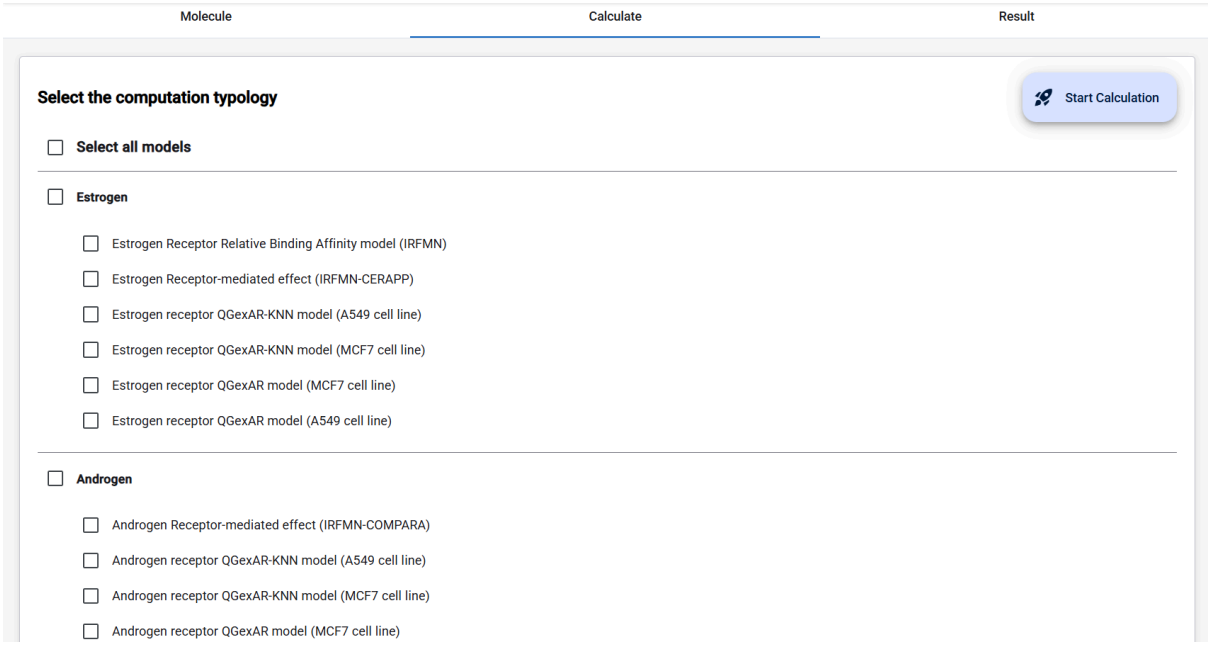


Figure 4. The models list in the "Calculate" page.

The last section is called "Result", Figure 5. A main table is available there, reporting the list of target molecules. Each entry reports the ID, SMILES, ED-index, the Property Scores based on which models the user selects and the "Inspect" button for a target compound. Above the table there is the search box, where the user can insert a key word to search among the table entries. On the right side there are the three export buttons:

- CSV: it exports a comprehensive file with all molecules and theirs results
- XLS: it exports a comprehensive excel file
- PDF: it exports a single PDF file with all the endpoints values and the relative graphs

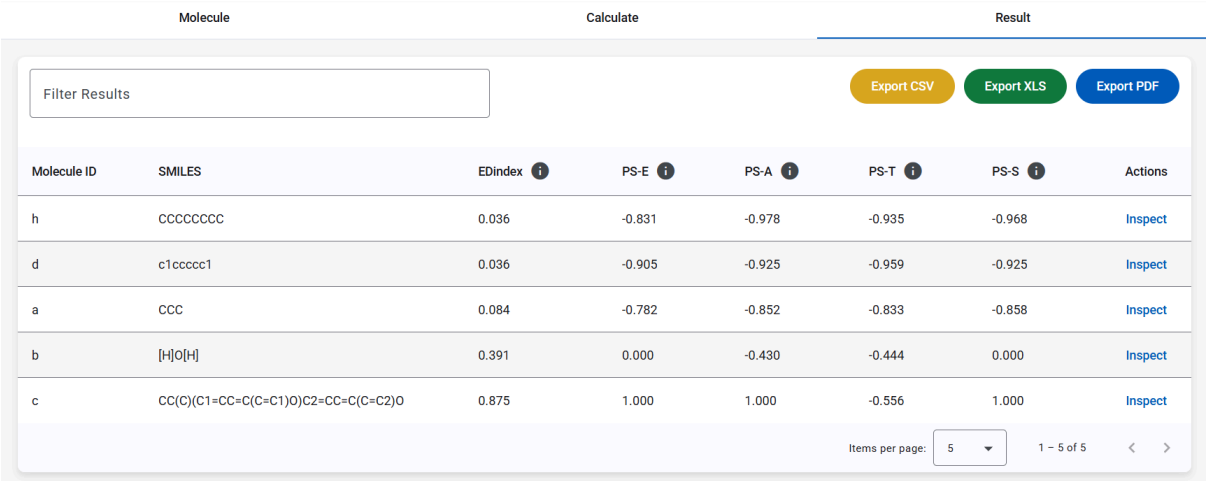


Figure 5. The result page

When the “Inspect” button is clicked, a form with the details of the calculated models and scores for that target compound is shown. There are different tabs, the first one reports the details of the calculated models for each endpoint. In the other tabs, different plots allowing a better visualization and analysis of the results are reported.

References and acknowledgments

The ED-SCREEN project (ANSES-21-EST-131) was funded by the French National Research Program for Environmental and Occupational Health and supervised by the French Agency for Food, Environmental, and Occupational Health and Safety (Anses)

Gadaleta, D., d'Alessandro, L., Marzo, M., Benfenati, E., & Roncaglioni, A. (2021). Quantitative Structure-Activity Relationship Modeling of the Amplex Ultrared Assay to Predict Thyroperoxidase Inhibitory Activity. *Frontiers in Pharmacology*, *12*, 713037. https://doi.org/10.3389/fphar.2021.713037

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ Health Perspect* **2016**, *124* (7), 1023–1033. https://doi.org/10.1289/ehp.1510267.

Mansouri, K., Kleinstreuer, N., Abdelaziz, A. M., Alberga, D., Alves, V. M., Andersson, P. L., Andrade, C. H., Bai, F., Balabin, I., Ballabio, D., Benfenati, E., Bhhatarai, B., Boyer, S., Chen, J., Consonni, V., Farag, S., Fourches, D., García-Sosa, A. T., Gramatica, P., … Judson, R. S. (2020). CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives*, *128*(2), 27002. https://doi.org/10.1289/EHP5580

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., Gould, J., Davis, J. F., Tubelli, A. A., Asiedu, J. K., Lahr, D. L., Hirschman, J. E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I. C., Lam, D., … Golub, T. R. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, *171*(6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049

Ollitrault G., Marzo, M., Roncaglioni, A., Benfenati, E., Mombelli, E., Taboureau, O. (2024). Prediction of endocrine-disrupting chemicals related to estrogen, androgen and thyroid hormone (EAT) modalities using transcriptomics data and machine learning. Toxics, 12(8), 541. https://doi.org/10.3390/toxics12080541.