Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2025)16

Unclassified

English - Or. English

14 October 2025

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Case Studies for the Integrated Approaches for Testing and Assessment in the Application of Combined Bioinformatics Approaches for Cross Species Extrapolation of Toxicity Knowledge to inform Chemical Safety. Tenth Review Cycle (2024).

Series on Testing and Assessment No 415

JT03574284

Please cite this publication as:

OECD (2025), *Case Studies for the Integrated Approaches for Testing and Assessment in the Application of Combined Bioinformatics Approaches for Cross Species Extrapolation of Toxicity Knowledge to inform Chemical Safety. Tenth Review Cycle (2024),* OECD Series on Testing and Assessment, No 415, OECD Environment, Health and Safety, Paris, https://one.oecd.org/official-document/ENV/CBC/MONO(2025)16/en

Contact us:

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

E-mail: ehscont@oecd.org

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 countries in North and South America, Europe and the Asia and Pacific region, as well as the European Union, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several Partner countries and from interested international organisations attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in twelve different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (https://www.oecd.org/en/topics/chemical-safety-and-biosafety.html).

Foreword

OECD member countries have been making efforts to expand the use of alternative methods in assessing chemicals. The OECD has been developing guidance documents and tools for the use of alternative methods such as (Q)SAR, chemical categories, and Adverse Outcome Pathways (AOPs) as a part of Integrated Approaches for Testing and Assessment (IATA). There is a need for the investigation of the practical applicability of these methods/tools for different aspects of regulatory decision-making, and to build upon case studies and assessment experience across jurisdictions.

The objective of the IATA Case Studies Project is to increase experience with the use of IATA by developing case studies, which constitute examples of predictions that are fit for regulatory use. The aim is to create common understanding of using novel methodologies and the generation of considerations/guidance stemming from these case studies.

This case study was developed by United States Environmental Protection Agency (US EPA) for illustrating practical use of IATA and submitted to the 2024 review cycle of the IATA Case Studies Project. This case study was reviewed by the project team.

This case study is an illustrative example, and its publication as an OECD monograph does not translate into direct acceptance of the methodologies for regulatory purposes across OECD countries. In addition, this case study should not be interpreted as official regulatory decisions made by the authoring member countries.

Table of contents

About the OECD 3

Foreword 4

Acronyms 8

Executive summary 11

1 Introduction 12
- Background 12
- Endpoints Addressed 13
- Type of Approach 14

2 PURPOSE 15
- Regulatory Context 15
- Data Interpretation Process 15
- Regulatory Application 15

3 CHEMICAL DOMAIN 17
- Limitations 17
- Applicability Domain 17

4 DESCRIPTION OF WORKFLOW 19

5 BASIS FOR IATA WORKFLOW 21
- MODULES 21
- INFORMATION SOURCES IN EACH MODULE 21
- What is Required? 22
- DATA SELECTION/GATHERING 23
- DATA EVALUATION 23
- DATA INTEGRATION AND INTERPRETATION 23
- EXPERT JUDGEMENT AND ALTERNATIVE INTERPRETATIONS 24
- DESCRIPTION OF UNCERTAINTIES 24

6 REFERENCES 25

Annex A. REPORTING INDIVIDUAL INFORMATION SOURCES for Genes to Pathways – Species Conservation Analysis 28
NAME OF THE INFORMATION SOURCE 28
RELEVANCE OF INFORMATION SOURCE 28
DESCRIPTION 28
RESPONSE(S) MEASURED 29
PREDICTION MODEL 30
METABOLIC COMPETENCE 31
STATUS OF DEVLEOPMENT/STANDARDIZATION/VALIDATION 31
TECHNICAL LIMITATIONS/LIMITATIONS WITH REGARD TO APPLICABILITY 31
WEAKNESSES AND STRENGTHS 32
RELIABILITY 33
PREDICTIVE CAPACITY 33
PROPRIETARY ASPECTS 34
PROPOSED REGULATORY USE 34
REFERENCES 35

Annex B. REPORTING INDIVIDUAL INFORMATION SOURCES for Sequence Alignment to Predict Across Species Susceptibility 36
NAME OF THE INFORMATION SOURCE 36
PREDICTION MODEL 41
METABOLIC COMPETENCE 42
STATUS OF DEVELOPMENT/STANDARDIZATION/VALIDATION 42
TECHNICAL LIMITATIONS/LIMITATIONS WITH REGARD TO APPLICABILITY 43
WEAKNESSES AND STRENGTHS 43
RELIABILITY 45
PREDICTIVE CAPACITY 45
PROPRIETARY ASPECTS 51
PROPOSED REGULATORY USE 51
REFERENCES 52

Annex C. REPORTING IATA OUTCOMES 55
PURPOSE 56
DESCRIPTION 57
DATA REVIEW 57
DATA INTEGRATION AND INTERPRETATION 58
EXPERT JUDGEMENT 58
UNCERTAINTIES SPECIFIC TO THE CHEMICAL ASSESSED 63
OUTCOME OF THE ASSESSMENT 64
Linking pathway Data to AOPs: 66
DISCUSSION OF OTHER INTERPRETATIONS 66
GUIDE TO THE COMBINED USE OF NAMs: SeqAPASS AND G2P-SCAN 67
REFERENCES 83

Annex D. READ ME to accompany ANNEX C Data for Case Studies 85

FIGURES

Figure 4.1. Application of -OMICs and bioinformatics approaches to enhance species extrapolation in decision making. 19

TABLES

Table C.1. Expert Judgment Summary 60
Table C.2. Uncertainties in Workflow Summary 62
Table C.3. Uncertainties Specific to the Chemical Assessed Summary 63
Table C.4. SeqAPASS query information for the 3 case examples (adapted from Schumann et al., 2024) 79

Acronyms

% – percent

(Q)SAR – Quantitative Structure Activity Relationship

AChE – acetylcholinesterase

ADME – absorption, distribution, metabolism, and elimination

ANOVA – analysis of variance

AO – adverse outcome

AOP-DB – adverse outcome pathway database

AOPs – Adverse Outcome Pathways

API – Application Programming Interface

AR – androgen receptor

BLASTp – Basic Local Alignment Search Tool for proteins

C-score – confidence score

CASP – Critical Assessment of protein Structure Prediction

CDD – Conserved domain database

COBALT – Constraint-based Multiple Alignment Tool

DES – diethylstilbestrol

DNA – deoxyribonucleic acid

DTSXID – DSSTox substance ID

E2 – Estradiol

EcR – ecdysone receptor

EDSP – Endocrine Disruptor Screening Program

ERA – ecological/environmental risk assessment

ERα – estrogen receptor-α

ESA – Endangered Species Act

ESR1 – estrogen receptor 1

F50V – Proline in position 50 to valine

G2P-SCAN – Genes-to-Pathways Species Conservation Analysis

GABA – (gamma-aminobutyric acid)

GABRA1 – Gamma-aminobutyric acid type A receptor subunit alpha

GABRA1gamma-aminobutyric acid type A receptor subunit alpha

HSD – honestly significant difference

HTS – high throughput screening

I-TASSER – Iterative Threading ASSEmbly Refinement

IATA – Integrated Approaches for Testing and Assessment

IC50 – Inhibitory concentration

ICACSER – International Consortium to Advance Cross Species Extrapolation in Regulation

ID – Identifier

KE – key event

KER – key event relationship

KEs – Key events

LC50 – Lethal concentration to 50% of the population

LDOs – least divergent orthologs

LFABP – human liver fatty acid–binding protein

MAPK – mitogen-activated protein kinases

MCODE – Molecular Complex Detection

MD – molecular dynamics

MIE – molecular initiating event

MIEs – molecular initiating events

MOAtox – US EPA Mode of Action and Toxicity

mRNA – messenger ribonucleic acid

MW – molecular weight

N – no

NADPH – nicotinamide adenine dinucleotide phosphate

NAMs – new approach methodologies

NCBI – National Center for Biotechnology Information

NIEHS – National Institute of Environmental Health Sciences

NR2F2 – Nuclear Receptor Subfamily 2 Group F Member 2

OCSPP – Office of Chemical Safety and Pollution Prevention

OECD – Organisation of Economic Cooperation and Development

OSRI – Other Scientifically Relevant Information

PDB – Protein Data Bank

PFAS – per- and polyfluoroalkyl substances

PFNA – Perfluorononanoic acid

PFOA – Perfluorooctanoic acid

PPARα – peroxisome proliferator activated receptor alpha

PPI – protein-protein interaction

PPP – Plant Protection Products

RBH – reciprocal best hit

RMSD – root mean square deviation

RNA – ribonucleic acid

RPS – Reversed Position Specific

RUNX1 – runt-related transcription factor 1

SeqAPASS – Sequence Alignment to Predict Across Species Susceptibility

SSDs – species sensitivity distributions

STRING – Search Tool Retrieval of INteracting Genes/Proteins

TD – toxicodynamic

tDOA – taxonomic domain of applicability

TEST – Toxicity Estimation Software Tool

THSD – thyroid hormone system disruption

TK – toxicokinetic

TM-align – template modeling-align

ToxCast – Toxicity Forecaster

US EPA – United States Environmental Protection Agency

WOE – weight of evidence

Y – yes

Executive summary

The Organisation for Economic Co-operation and Development (OECD) Integrated Approaches for Testing and Assessment (IATA) demonstrates how bioinformatics can be strategically applied to extrapolate chemical toxicity knowledge across species. It emphasizes the integration of two complementary tools, the United States Environmental Protection Agency's Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS; https://seqapass.epa.gov/seqapass/) and Genes-to-Pathways Species Conservation Analysis (G2P-SCAN), to assess biological pathway and protein target conservation, supporting chemical safety assessments while reducing reliance on animal testing.

Global regulatory frameworks increasingly prioritize new approach methodologies (NAMs) to evaluate chemical safety in a more ethical, efficient, and scientifically robust manner. Cross-species extrapolation is central to this goal, allowing data from a few well-studied organisms to inform protection goals across diverse species, including threatened and endangered wildlife.

The web-based SeqAPASS tool evaluates protein sequence and structural conservation across thousands of species to predict potential chemical-protein interactions, yielding predictions of chemical susceptibility. The SeqAPASS tool can initiate a cross-species extrapolation from any species with empirical evidence of a chemical-protein interaction and available protein sequence information. The G2P-SCAN analyzes conservation of pathway conservation using human gene inputs. Individually, each tool can evaluate conserved biology across species but together they strengthen the weight of evidence by combining molecular depth with pathway breadth to define the tDOA within the adverse outcome pathway framework.

The integrated approach informs hazard assessment, supports prioritization during early screening, and enhances interpretation of mechanistic data. The results can serve as OSRI or additional lines of evidence in WOE evaluations in combination with data from other toxicity information sources (e.g., exposure data, physical-chemical properties, toxicologically relevant studies in the published literature, (Q)SAR models and other data submitted to support chemical risk assessment). This aids in decision-making for chemical safety or supports intelligent test design by identifying the most biologically relevant species for further testing.

The IATA represents a scientifically transparent, regulatory decision-making method to improve chemical safety across species. It aligns with global shifts toward mechanistic and computational toxicology, enabling broader application of existing data, reducing animal testing, and enhancing environmental and human health protections under a One Health framework.

1 Introduction

Background

Chemical safety assessments are used to understand the toxicity of chemicals so that they can be regulated, labeled appropriately, and used in a manner that limits or eliminates adverse impacts on human health or the environment (Ball et al., 2022). To accommodate the increasing rate at which new chemicals are being developed and to align with global efforts to reduce, refine, and/or replace animal use, toxicity testing has begun shifting away from whole animal test methods toward the use of high throughput assays, typically cell- or omics-based, and computational approaches to inform such assessments (Krewski et al., 2020). A challenge that remains, regardless of the experimental methods used to understand the toxicity of a chemical, is the extrapolation of this knowledge from the organisms for which the assays are relevant, to the diversity of species required for decision-making regarding potential for environmental impact (LaLone et al., 2021). Such cross-species extrapolation can be defined as using information from one species to infer, predict, or estimate the impact or biological response of another species.

Globally, legislative mandates have been implemented to protect the assemblage of all species from unintended consequences of chemical use. Therefore, regulatory decisions rely on empirical evidence gathered from model or surrogate species to set limits and guidelines for chemicals. It is impractical and undesirable to use empirical tests to evaluate all species for toxicity, so decision-makers rely on extrapolation techniques to assess potential for chemical effects in unassessed organisms. Historically, tools such as safety factors and species sensitivity distributions (SSDs) have been used to extrapolate and estimate safe threshold concentrations across a broad range of environmentally relevant species from a few selected model organisms. These mathematical models, however, rely on assumptions of interspecies relatedness to infer biological responses to toxicants and, in the case of SSDs, largely rely on availability of at least some cross species empirical toxicity data.

There are difficulties in using surrogate species for chemical safety assessments, as the toxicity and response to chemicals can vary among organisms, making it challenging to extrapolate findings from one species to another (Khabib et al., 2022). Understanding the scientific principles determining species similarities or differences in chemical responses is essential for extrapolating toxicity knowledge/data across the diversity of species. Regulatory agencies, such as the US Environmental Protection Agency, Environment and Climate Change Canada, the United Kingdom Department for Environment, Food, and Rural Affairs, and the European Chemicals Agency, are increasingly promoting the use of new approach methodologies (NAMs) as part of their initiatives to improve chemical safety assessment and reduce the reliance on animal testing (Stucki et al., 2022). NAMs is used as an umbrella term to describe any *in vitro*, *in chemico*, or *in silico* (computational) method that evaluates chemical safety while minimizing the use of intact animals (Sewell et al., 2024). Many of these approaches inform the molecular basis to understanding chemical-induced toxicity, providing evidence to identify MIEs and other early KEs in an AOP framework. The AOP framework describes the causal linkages between the initial perturbation of a specific biological target, the MIE, and the resulting adverse outcome (AO) of regulatory concern at the individual or population level (Ankley et al., 2010). The pathways encompass multiple levels of biological organization. The toxicity knowledge used to describe the AOP comes from the published literature and AOPs are typically built with information from a single or few model organisms. There are also complementary

approaches that have been developed to inform assumptions of conserved biology across species that yield a scientific basis for the species extrapolation necessary in risk assessment (Rivetti et al., 2020). Overall, the transition to NAMs in toxicology represents a paradigm shift towards a “One Health” focus, with the intent to yield efficient and ethical approaches for assessing potential chemical safety, by leveraging advanced technologies and computational methods to provide a more comprehensive understanding of toxicity across species (Brooks et al., 2024). However, the shift to such methods requires changes in policy to accept NAM data for making regulatory decisions. Due to the nature of regulation, new approaches must be well-vetted, backed with strong science and transparency in their domain of applicability.

There are several key toxicokinetic (TK) and toxicodynamic (TD) considerations to evaluate conserved biology across species for extrapolation of toxicity knowledge, particularly if the intent is to use the information for regulatory decision-making. The likelihood for exposure and ADME of chemicals is major considerations. Further, the life stage or life history of organisms may play a role in sensitivity. When a chemical interacts with a biomolecule within an organism, the conservation of the biological pathway(s) leading to an adverse outcome becomes a major consideration for extrapolation. It is envisioned that approaches used to understand both TK and TD will ultimately provide the greatest context for extrapolating knowledge/data from one organism to others for chemical safety decisions (van den Berg et al., 2021). Advances have been made to consider conservation of biology for cross species extrapolation using bioinformatics.

Bioinformatics is a field of science that combines the use of mathematics, information science, and biology to answer complex biological questions using “big data” and allowing computers to process the information (Mitra et al., 2022). Therefore, the focus of this IATA is to demonstrate the use of bioinformatics tools that can be combined for the purpose of extrapolating toxicity knowledge across species. This IATA focuses on the application of two complimentary bioinformatics approaches, Genes-to-Pathways Species Conservation Analysis (G2P-SCAN) and the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool for the evaluation of conserved biological pathways across species. This approach is considered to help identify potentially susceptible species, facilitate the design of intelligent testing strategies which can be targeted at data generation in relevant (the most likely susceptible) species only and aimed at maximizing the use of all the available knowledge in decision making. While this IATA is focused on the demonstration of bioinformatics approaches for evaluating biological pathway conservation, the intent is to provide a framework for future integration of additional approaches that address toxicokinetic and (additional) toxicodynamic factors influencing species susceptibility. The combination of these factors will ultimately provide the greatest context for extrapolating knowledge/data across species for supporting chemical safety decisions (van den Berg et al., 2021).

Endpoints Addressed

The endpoints evaluated focus on generating lines of evidence for conservation of biology across species as a means to predict chemical susceptibility and therefore extrapolate toxicity knowledge. The two tools applied in this IATA for the conservation analysis include the G2P-SCAN and SeqAPASS. These approaches assess the level of conservation of molecular genes/targets across species and can be applied in the context of the AOP framework, where available, to aid in defining the biologically plausible tDOA (Jensen et al., 2023; Haigis et al., 2023). Given the different methodologies undertaken by the two approaches, there are clear benefits in their combined and integrated application. While SeqAPASS allows for the in depth evaluation of the conservation of proteins and further chemical-protein interactions across the diversity of species providing multiple lines of evidence customized to the degree of knowledge that exists for a chemical-protein interaction (sequence and structure), G2P-SCAN can complement SeqAPASS by allowing for a bi-dimensional analysis which looks at the conservation of entire pathways where the input gene/target has been well-annotated in human, despite being applicable to 6 model

species. Each approach has significant utility in-of-itself, as described in ANNEX I for SeqAPASS and G2P-SCAN, however, the combined approach adds both pathway breadth and species depth across potentially hundreds of species.

In the context of species extrapolation, the key question is: To what extent can the biology described in these empirical studies be reasonably extrapolated across non-tested species? Bioinformatics approaches tend to focus on comparative genomics/proteomics through evaluating biomolecules like DNA, RNA, and proteins. Thus, these approaches are particularly useful for evaluating the conservation of the MIE in an AOP, which represents the interaction between a chemical stressor and biomolecule, as well as early KEs at the molecular, cellular or tissue levels of biology that are dictated directly by genes and proteins. It is noteworthy that even if an AOP does not exist currently for the chemical-biomolecule interaction, that does not preclude the chemical from the tools presented in this IATA. Although in such instances the depth of knowledge of the causal linkage from the molecular interaction to an adverse effect may not be well-defined and therefore the application of the knowledge may differ in a regulatory context.

Type of Approach

The approach taken in this IATA is to combine available bioinformatics tools and databases strategically to address challenges in extrapolating toxicity knowledge across species, creating results that can be used in regulatory decision-making (Schumann et al., 2024). The focus on the use of combined bioinformatics approaches will be for the evaluation of conserved biology in the context of DNA, RNA, and protein. Specifically, the conservation of chemical-protein interactions and downstream molecular events (i.e., biological pathways) will be examined. The underlying assumption is that if the biology is conserved from the MIE through early KEs across species, where molecular level changes are essential, there is potential for the stressor to lead to an AO. This should be true even when the apical AOs may differ across taxa or species (Jensen et al., 2023).

2 PURPOSE

Regulatory Context

Globally, legislative mandates have been implemented to protect all species from the unintended consequences of chemical use. Regulatory decisions rely on empirical evidence gathered from model or surrogate species to set limits and guidelines for chemicals. In chemical safety evaluations for any risk assessment, decision-makers take advantage of existing toxicity knowledge from a single or a handful of species to inform protection goals for the diversity of organisms or for specific protected groups (e.g., threatened and endangered species) (Cegar et al., 2022). The species with toxicity data are, therefore, used as surrogates, representing other taxa through extrapolation. The challenge in a regulatory context is providing scientific evidence to support these cross-species extrapolations. The global regulatory landscape is also evolving to reduce reliance on animal testing. This shift requires greater use of mechanistic, cell-based, and computationally-derived information and the AOP conceptual framework is being readily considered or actively adopted in many regulatory, industry, and academic fora. In this respect, cross species extrapolation becomes a crucial aspect of regulatory science, especially in the context of providing confidence to drive environmental safety and chemical risk assessment decision, because it is impractical to test every chemical on all species directly, especially in a context of reducing animal testing. Through using data from one species to predict the impact or biological response in another species, it allows the definition of a plausible taxonomical domain of applicability of adverse effect(s) (Jensen et al., 2023).

Data Interpretation Process

The data interpretation process for species extrapolation will differ for each information source used although the overall intention is to examine the conservation of biology across species upon chemical perturbation. Each information source will have a well-defined, published data interpretation process. The case examples will demonstrate how the information sources described in Annex I can be brought together to strengthen the evidence for extrapolation in a regulatory context in Annex II. The intent is to continue to build upon this IATA with other bioinformatics approaches being added and new case studies for demonstrated application.

Regulatory Application

The significance of the methodology presented here lies in its potential to facilitate the application of mechanistically-based data for ERA and filling data-gaps. This approach aligns with global regulatory shifts towards the use of NAMs in safety assessments by enabling more substantiated species extrapolation. This approach is intended to support prioritization and cutoff criteria during the screening stage of the ERA framework and to enable the design of intelligent testing strategies which are 1) pragmatic and targeted at data generation in only species for which likely toxicity targets and pathways are relevant and 2) aimed at maximizing the use of all the available knowledge and including directed testing strategies and data

generated in other nonstandard species (i.e., humans) as support for environmental safety decision making. Results from the bioinformatics approaches are intended to inform hazard assessment and to be used as OSRI or additional lines of evidence in WOE evaluations in combination with data collected from other toxicity information sources (e.g., exposure data, physical-chemical properties, toxicologically relevant studies in the published literature, (Q)SAR models and other data submitted to support chemical risk assessment) to make decisions for chemical safety. In addition, it can be relevant in supporting, identifying, and leveraging existing data to help evaluate a chemical's potential hazard by characterizing the susceptibility of key model organisms of environmental relevance and building evidence for the conservation of biological pathways. Overall, this will support a reduction in the number of animals used for toxicity testing purposes. Moreover, this IATA provides an understanding within implementing agencies on how to interpret new data streams.

3 CHEMICAL DOMAIN

Limitations

The chemical applicability domain for this IATA includes any chemical with a known MIE or defined molecular target. To use bioinformatics approaches for extrapolation of toxicity knowledge across species by evaluating conservation of biology, it is necessary to thoroughly understand the biology in at least one species as a starting point for the extrapolation. For example, if the chemical-protein interaction is understood in one species, sequence, structural, and pathway conservation can be examined as lines of evidence to predict whether that chemical interaction is relevant in other species. Additionally, there must be empirical evidence to support a known chemical-biomolecule interaction. In the context of the AOP framework, MIEs and early KEs describe the biology from the progression of a chemical (stressor)-biomolecule perturbation of a pathway towards an apical outcome with regulatory significance. Typically, genes and proteins are described in the MIE or KE with supporting evidence for their essentiality (Villeneuve et al. 2014). Causal linkages from the MIE to AOs are also important and, therefore, the lack of AOPs can be a limitation for some applications of bioinformatics approaches in the context of evaluating the tDOA. A number of structural fragment / alert based schemes also exist to help provide lines of evidence in the prediction of the mode or mechanism of action of chemical-biomolecule interactions. Historic schemes such as the MOAtox database (Barron, Lilavois and Martin, 2015) and Toxicity Estimation Software Tool (TEST) can be run alongside more recent schemes based on AOP MIE classification such as MechoA (available as KREATiS MechoA scheme20) and the Sapounidou-Firman scheme (Sapounidou et al., 2021) (Firman et al., 2022). Whilst the more recent schemes provide increased chemical space coverage and enhanced MechoA identification over historic schemes, prediction discrepancies between tools continue to be a limitation in the wider chemical domain application of extrapolation approaches in some cases.

In addition, quality gene/protein sequence data and appropriate annotation must be available for the species of interest in the extrapolation analysis to inform whether there is sufficient evidence of conservation. If relevant sequence information does not exist or is of low quality, then the species extrapolation approaches are not appropriate/possible and will not be performed or of use in decision-making.

The results provided by bioinformatics approaches for extrapolating toxicity data/knowledge across species examine the conservation of biology at the molecular level. These results can be applied in hazard assessments and can be incorporated with other relevant information to inform risk assessment. Therefore, it is critical to acknowledge that toxicokinetic factors such ADME, as well as organism life-stage and life history should be applied in a WOE approach for decision-making.

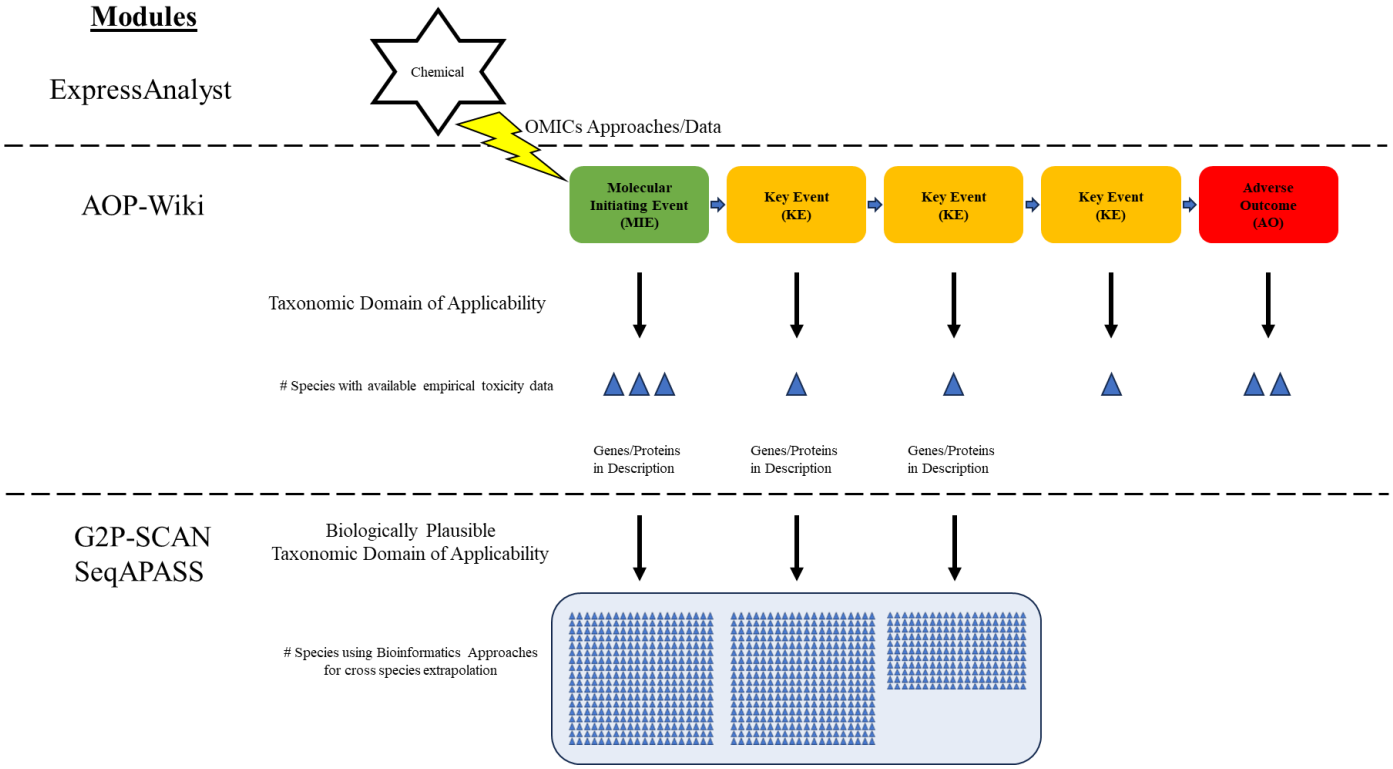
Applicability Domain

The applicability domain for the use of bioinformatics approaches to extrapolate toxicity knowledge from species to species requires knowledge of a chemical-biomolecule (i.e., gene/protein) interaction in at least one species. Depending on how well that chemical interaction has been characterized, additional lines of

evidence can be added to the extrapolation analysis. If the chemical target is unknown or the chemical acts non-specifically, for example via narcosis or cytotoxicity, approaches for evaluating conserved biology across species using bioinformatics are not always appropriate. Currently, such bioinformatics approaches are being applied to predict chemical susceptibility across species in the context of endocrine disrupting chemicals, and in pesticide and endangered species decision making frameworks.

4 DESCRIPTION OF WORKFLOW

Figure 4.1. Application of -OMICs and bioinformatics approaches to enhance species extrapolation in decision making.



Starting with a chemical of interest, tools like ExpressAnalyst (https://www.expressanalyst.ca/) allow for analysis of gene expression data that can provide insight into the chemical exposure and mechanisms for toxicity (i.e., molecular initiating event in an AOP). The AOP-Wiki (https://aopwiki.org/) captures AOP descriptions of causal linkages starting with the identified MIE and moving downstream from one KE to the next through the levels of biological organization leading to adverse outcomes relied upon in regulatory decision making (e.g., reproduction, growth, survival). The empirical evidence and biological plausibility are used to define the tDOA in the AOP-Wiki. However, through existing empirical studies, the tDOA is narrowly defined since few pathways have detailed toxicity information for more than a handful of species. With the biology, (such as key genes and proteins) and the tDOA defined in the AOP, bioinformatics approaches such as SeqAPASS (https://seqapass.epa.gov/seqapass/) and G2P-SCAN (https://github.com/seacunilever/G2P-SCAN) can be used to extrapolate toxicity knowledge beyond the model (surrogate) organism defined by the empirical studies describing the AOP to the biologically plausible tDOA, increasing to hundreds or thousands of species for decision-making purposes. The first iteration of this IATA focuses on demonstrating how G2P-SCAN and SeqAPASS can be combined for

gathering knowledge of conserved biology across species for extrapolation of toxicity knowledge to untested species, useful for numerous regulatory applications. Blue triangles represent individual species with scientific evidence to support the tDOA.

5 BASIS FOR IATA WORKFLOW

MODULES

Conservation Analysis

Modules included in this IATA are bioinformatics approaches/tools to evaluate the conservation of biology across species to inform extrapolation of toxicity data and knowledge for use in decision-making. Specifically, these modules evaluate DNA, RNA, and/or protein sequence and structural conservation across species.

INFORMATION SOURCES IN EACH MODULE

In silico

A. Genes-to-Pathways Species Conservation Analysis (G2P-SCAN) tool

Links: https://github.com/seacunilever/G2P-SCAN

G2P-SCAN Publication:

1. Genes-to-Pathways Species Conservation Analysis: Enabling the Exploration of Conservation of Biological Pathways and Processes Across Species (Rivetti et al., 2023)

B. United States Environmental Protection Agency's Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool

Links: https://seqapass.epa.gov/seqapass/

https://www.epa.gov/comptox-tools/sequence-alignment-predict-across-species-susceptibility-seqapass-resource-hub

SeqAPASS Publications:

1. From Protein Sequence to Structure: The Next Frontier in Cross Species Extrapolation for Chemical Safety Evaluations (LaLone et al., 2022)
2. Cross-species applicability of an adverse outcome pathway network for thyroid hormone system disruption (Haigis et al., 2023)
3. Weight of evidence for cross-species conservation of androgen receptor-based biological activity (Vliet et al., 2023a)

4. Demonstration of the Sequence Alignment to Predict Across Species Susceptibility Tool for Rapid Assessment of Protein Conservation (Vliet et al., 2023b)
5. Defining the Biologically Plausible Taxonomic Domain of Applicability of an Adverse Outcome Pathway: A Case Study Linking Nicotinic Acetylcholine Receptor Activation to Colony Death (Jensen, Blatz and LaLone, 2022)
6. Integrative Computational Approaches to Inform Relative Bioaccumulation Potential of Per-and Polyfluoroalkyl Substances Across Species (Cheng et al., 2021)
7. Evidence for Cross Species Extrapolation of Mammalian-Based High-Throughput Screening Assay Results (LaLone et al., 2018)
8. In Silico Site-Directed Mutagenesis Informs Species-Specific Predictions of Chemical Susceptibility (Doering et al., 2018)
9. Evaluation of the scientific underpinnings for identifying estrogenic chemicals in nonmammalian taxa (Ankley et al., 2016)
10. SeqAPASS: A Web-Based Tool for Addressing the Challenges of Cross-Species Extrapolation (LaLone et al., 2016)
11. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk (LaLone et al., 2013)

C. Cross-species extrapoaltion and combined SeqAPASS/G2P-SCAN approach

1. International Consortium to Advance Cross-Species Extrapolation of the Effects of Chemicals in Regulatory Toxicology (LaLone et al., 2021)
2. Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species (Schumann et al., 2024).
3. Bridging the Gap Between Human Toxicology and Ecotoxicology Under One Health Perspective by a Cross-Species Adverse Outcome Pathway Network for Reproductive Toxicity (Dufourcq Sekatcheff, Jeong and Choi, 2024)

D. Structural profilers

1. U.S. EPA *User Guide for T.E.S.T. (Version 4.2) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure* , 2016.
2. KREATiS, (2020) iSafeRat® (in silico Algorithms For Environmental Risk And Toxicity) Online version v2.1https://isaferat.kreatis.eu.

What is Required?

Employing knowledge of the chemical-biomolecule or protein-protein interactions that act as the MIE or connect upstream KEs to downstream responses in an AOP, G2P-SCAN and SeqAPASS have certain input requirements:

G2P-SCAN input – Extrapolation from human gene(s)

SeqAPASS input – Extrapolation from any protein from any species with sequence information

Level 1: Primary amino acid comparison requires the protein name or accession (National Center for Biotechnology Information (NCBI) Protein ID) for the target of interest and a known sensitive

species (e.g., the chemical was designed to target the species) or a model surrogate species (this can be the model species used in an assay or toxicity test).

Level 2: Functional domain comparison requires knowledge of the selection of functional domain(s) from NCBI's conserved domain database. Ideally, these are domains that are known to interact with the chemical of interest.

Level 3: Critical amino acid comparison requires knowledge of key amino acids that are directly involved in the chemical-protein interaction.

Level 4 (for experts in structural biology): Protein structural alignments require an available crystal structure for the protein of interest, ideally with the chemical of interest bound (e.g., derived from x-ray crystallography)

DATA SELECTION/GATHERING

How information is identified

Information for use in the bioinformatics approaches typically comes from gene/protein sequence and structural databases such as that curated in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/), SwissProt/UniProt (https://www.uniprot.org/), and the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB; https://www.rcsb.org/). Each information source described in ANNEX I will identify the source data for the approach.

Inclusion/Exclusion Criteria

Inclusion of an information source (ANNEX I) requires peer reviewed and published bioinformatics approaches that have been developed to aid in extrapolating knowledge across species. In addition, the approaches should have published case studies demonstrating their application in chemical safety evaluations. Inclusion/exclusion criteria for data sources are defined by the individual information sources as each have a specified domain of applicability.

DATA EVALUATION

Data Quality

Data quality should be defined by the individual information sources (ANNEX I) and how they are used in each case study (ANNEX II).

Uncertainty Considerations

Limitations or uncertainty considerations of approaches are described in ANNEX I and ANNEX II.

DATA INTEGRATION AND INTERPRETATION

How Data Are Combined

Data from different approaches are combined in a manner that is scientifically defensible as defined in ANNEX II. Typically, consensus of results demonstrating conservation of pathway knowledge across species will be highlighted.

How Data Are Weighted

If data are weighted the description will be included in ANNEX II. In ANNEX II, the case studies combining G2P-SCAN and SeqAPASS results do not have weighted data.

Data Interpretation Process

Data interpretation processes are unique to the information source used and will be defined in ANNEX I and in ANNEXII Guide to Combined Use of NAMs section.

Points for Expert Decisions and Rationale for Each

Points for expert decisions and elements that require scientific rationale for the decision-making process will be defined in ANNEX II and are dependent on the individual information sources combined.

EXPERT JUDGEMENT AND ALTERNATIVE INTERPRETATIONS

Necessary Expert Judgement at Each Point

Points throughout the extrapolation process using the information sources for data analysis that require expert judgement will be highlighted in ANNEX I and ANNEX II

DESCRIPTION OF UNCERTAINTIES

General Description

Uncertainties for bioinformatics approaches used in species extrapolation to define the biologically plausible tDOA for AOPs typically are relative to the strength in empirical support for the chemical-biomolecule interaction(s) in at least one known sensitive or targeted species. If the chemical-biomolecular interaction is not well characterized or if there is uncertainty regarding the specificity of the interaction(s), then uncertainty remains in the extrapolation of knowledge across species. In addition, bioinformatics approaches focus on presence/absence (conservation) of a particular biomolecule, which is the known target of the chemical, or further, conservation of the chemical-biomolecule interaction itself. Therefore, if there are unknown molecular targets or factors such as taxonomic toxicokinetic differences that drive sensitivities across species, the bioinformatics approaches alone will lead to uncertainties.

How Uncertainty is Addressed/Overcome

Uncertainties can be overcome/better characterized by bringing together all available information regarding a chemical, including toxicity knowledge from the available literature, toxicokinetic data, *in vitro* results, and other relevant *in silico* results. Evaluating conservation from multiple levels of biology addresses uncertainties (Ankley et al., 2016; Vliet et al., 2023).

6 REFERENCES

Scientific Publications:

Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, Serrrano JA. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry. 2010 Mar;29(3):730-41.

Ankley GT, LaLone CA, Gray LE, Villeneuve DL, Hornung MW. Evaluation of the scientific underpinnings for identifying estrogenic chemicals in nonmammalian taxa using mammalian test systems. Environmental toxicology and chemistry. 2016 Nov;35(11):2806-16.

Ball N, Bars R, Botham PA, Cuciureanu A, Cronin MT, Doe JE, Dudzina T, Gant TW, Leist M, van Ravenzwaay B. A framework for chemical safety assessment incorporating new approach methodologies within REACH. Archives of Toxicology. 2022 Mar;96(3):743-66.

Barron, M., C. Lilavois and T. Martin (2015), "MOAtox: A comprehensive mode of action and acute aquatic toxicity database for predictive model development", *Aquatic Toxicology*, Vol. 161, pp. 102-107, https://doi.org/10.1016/j.aquatox.2015.02.001.

Brooks BW, van den Berg S, Dreier DA, LaLone CA, Owen SF, Raimondo S, Zhang X. Towards precision ecotoxicology: Leveraging evolutionary conservation of pharmaceutical and personal care product targets to understand adverse outcomes across species and life stages. Environmental Toxicology and Chemistry. 2024 Mar;43(3):526-36.

Ceger P, Vinas NG, Allen D, Arnold E, Bloom R, Brennan JC, Clarke C, Eisenreich K, Fay K, Hamm J, Henry PF. Current ecotoxicity testing needs among selected US federal agencies. Regulatory Toxicology and Pharmacology. 2022 Aug 1;133:105195.

Cheng W, Doering JA, LaLone C, Ng C. Integrative computational approaches to inform relative bioaccumulation potential of per- and polyfluoroalkyl substances across species. *Toxicol Sci.* 2021;180(2):212-223.

Doering JA, Lee S, Kristiansen K, Evenseth L, Barron MG, Sylte I, LaLone CA. In silico site-directed mutagenesis informs species-specific predictions of chemical susceptibility derived from the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool. Toxicological Sciences. 2018 Nov 1;166(1):131-45.

Dufourcq Sekatcheff E, Jeong J, Choi J. Bridging the gap between human toxicology and ecotoxicology under one health perspective by a cross-species adverse outcome pathway network for reproductive toxicity. Environ Toxicol Chem. 2024;44(9):2511-2523. doi:10.1002/etc.5940

Firman J, Doering JA, Gust KA, et al. Construction of an in silico structural profiling tool facilitating mechanistically grounded classification of aquatic toxicants. Environ Sci Technol. 2022;56(24):17805-17814. doi:10.1021/acs.est.2c03736

Haigis AC, Vergauwen L, LaLone CA, Villeneuve DL, O'Brien JM, Knapen D. Cross-species applicability of an adverse outcome pathway network for thyroid hormone system disruption. Toxicological Sciences. 2023 Sep 1;195(1):1-27.

Jensen MA, Blatz DJ, LaLone CA. Defining the biologically plausible taxonomic domain of applicability of an adverse outcome pathway: A case study linking nicotinic acetylcholine receptor activation to colony death. Environmental Toxicology and Chemistry. 2023 Jan;42(1):71-87.

Khabib MN, Sivasanku Y, Lee HB, Kumar S, Kue CS. Alternative animal models in predictive toxicology. Toxicology. 2022 Jan 15;465:153053.

Krewski D, Andersen ME, Tyshenko MG, Krishnan K, Hartung T, Boekelheide K, Wambaugh JF, Jones D, Whelan M, Thomas R, Yauk C. Toxicity testing in the 21st century: progress in the past decade and future perspectives. Archives of toxicology. 2020 Jan;94:1-58.

LaLone CA, Basu N, Browne P, Edwards SW, Embry M, Sewell F, Hodges G. International consortium to advance cross-species extrapolation of the effects of chemicals in regulatory toxicology. Environmental toxicology and chemistry. 2021 Dec;40(12):3226

LaLone CA, Blatz DJ, Jensen MA, Vliet SM, Mayasich S, Mattingly KZ, Transue TR, Melendez W, Wilkinson A, Simmons CW, Ng C. From Protein Sequence to Structure: The Next Frontier in Cross‐Species Extrapolation for Chemical Safety Evaluations. Environmental Toxicology and Chemistry. 2023 Feb;42(2):463-74.

LaLone CA, Villeneuve DL, Burgoon LD, Russom CL, Helgen HW, Berninger JP, Tietge JE, Severson MN, Cavallin JE, Ankley GT. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk of chemicals with known modes of action. Aquatic toxicology. 2013 Nov 15;144:141-54.

LaLone CA, Villeneuve DL, Doering JA, Blackwell BR, Transue TR, Simmons CW, Swintek J, Degitz SJ, Williams AJ, Ankley GT. Evidence for cross species extrapolation of mammalian-based high-throughput screening assay results. *Environ Sci Technol.* 2018;52(23):13960-13971.

LaLone CA, Villeneuve DL, Lyons D, Helgen HW, Robinson SL, Swintek JA, Saari TW, Ankley GT. Editor's highlight: sequence alignment to predict across species susceptibility (SeqAPASS): a web-based tool for addressing the challenges of cross-species extrapolation of chemical toxicity. Toxicological Sciences. 2016 Oct 1;153(2):228-45.

Mitra D, Mitra D, Bensaad MS, Sinha S, Pant K, Pant M, Priyadarshini A, Singh P, Dassamiour S, Hambaba L, Panneerselvam P. Evolution of Bioinformatics and its impact on modern bio-science in the twenty-first century: Special attention to pharmacology, plant science and drug discovery. Computational Toxicology. 2022 Nov 1;24:100248.

Rivetti C, Allen TE, Brown JB, Butler E, Carmichael PL, Colbourne JK, Dent M, Falciani F, Gunnarsson L, Gutsell S, Harrill JA. Vision of a near future: Bridging the human health–environment divide. Toward an integrated strategy to understand mechanisms across species for chemical safety assessment. Toxicology In Vitro. 2020 Feb 1;62:104692.

Rivetti C, Houghton J, Basili D, Hodges G, Campos B. Genes‐to‐Pathways Species Conservation Analysis: Enabling the Exploration of Conservation of Biological Pathways and Processes Across Species. Environmental Toxicology and Chemistry. 2023 May;42(5):1152-66, https://doi.org/10.1002/etc.5600.

Sapounidou, M. et al. (2021), "Development of an Enhanced Mechanistically Driven Mode of Action Classification Scheme for Adverse Effects on Environmental Species", *Environmental Science & Technology*, Vol. 55/3, pp. 1897-1907, https://doi.org/10.1021/acs.est.0c06551.

Sewell F, Alexander-White C, Brescia S, Currie RA, Roberts R, Roper C, Vickers C, Westmoreland C, Kimber I. New approach methodologies (NAMs): identifying and overcoming hurdles to accelerated adoption. Toxicology Research. 2024 Apr 1;13(2):tfae044.

Schumann, P. et al. (2024), "Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species", *Science of The Total Environment*, Vol. 912, p. 168573, https://doi.org/10.1016/j.scitotenv.2023.168573.

Stucki AO, Barton-Maclaren TS, Bhuller Y, Henriquez JE, Henry TR, Hirn C, Miller-Holt J, Nagy EG, Perron MM, Ratzlaff DE, Stedeford TJ. Use of new approach methodologies (NAMs) to meet regulatory requirements for the assessment of industrial chemicals and pesticides for effects on human health. Frontiers in Toxicology. 2022 Sep 1;4:964553.

U.S. Environmental Protection Agency (EPA). *User Guide for T.E.S.T. (Version 4.2) – Toxicity Estimation Software Tool: A Program to Estimate Toxicity from Molecular Structure*. EPA/600/R-16/058. 2016. Retrieved from https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P100OFSG.TXT

Van den Berg SJ, Maltby L, Sinclair T, Liang R, van den Brink PJ. Cross-species extrapolation of chemical sensitivity. Science of the total environment. 2021 Jan 20;753:141800.

Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, Landesmann B, Lettieri T, Munn S, Nepelska M, Ottinger MA. Adverse outcome pathway (AOP) development I: strategies and principles. Toxicological Sciences. 2014 Dec 1;142(2):312-20.

Vliet SM, Markey KJ, Lynn SG, Adetona A, Fallacara D, Ceger P, Choksi N, Karmaus AL, Watson A, Ewans A, Daniel AB. Weight of evidence for cross-species conservation of androgen receptor-based biological activity. Toxicological Sciences. 2023 Jun 1;193(2):131-45. (Vliet et al., 2023a)

Vliet SM, Hazemi M, Blatz D, Jensen M, Mayasich S, Transue TR, Simmons C, Wilkinson A, LaLone CA. Demonstration of the sequence alignment to predict across species susceptibility tool for rapid assessment of protein conservation. JoVE (Journal of Visualized Experiments). 2023 Feb 10(192):e63970. (Vliet et al., 2023b)

Online Tools and Databases:

KREATiS. iSafeRat® (in silico Algorithms For Environmental Risk And Toxicity), Online version v2.1. Retrieved from https://isaferat.kreatis.eu [Accessed: October 2025]

National Center for Biotechnology Information (NCBI). *NCBI Database*. Retrieved from https://www.ncbi.nlm.nih.gov/ [Accessed: October 2025]

Research Collaboratory for Structural Bioinformatics (RCSB). *Protein Data Bank (PDB)*. Retrieved from https://www.rcsb.org/ [Accessed: October 2025]

Seacunilever. *Genes-to-Pathways Species Conservation Analysis (G2P-SCAN)* [GitHub repository]. Retrieved from https://github.com/seacunilever/G2P-SCAN [Accessed: October 2025]

UniProt Consortium. *UniProt: A worldwide hub of protein knowledge*. Retrieved from https://www.uniprot.org/ [Accessed: October 2025]

United States Environmental Protection Agency (US EPA). *SeqAPASS Resource Hub*. Retrieved from https://www.epa.gov/comptox-tools/sequence-alignment-predict-across-species-susceptibility-seqapass-resource-hub [Accessed: October 2025]

United States Environmental Protection Agency (US EPA). *Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)* [Online tool]. Retrieved from https://seqapass.epa.gov/seqapass/ [Accessed: October 2025]

Annex A. REPORTING INDIVIDUAL INFORMATION SOURCES for Genes to Pathways – Species Conservation Analysis

NAME OF THE INFORMATION SOURCE

Genes To Pathways – Species Conservation Analysis (G2P-SCAN), https://github.com/seacunilever/G2P-SCAN

- Screenshots directly from the G2P-SCAN tool and detailed descriptions of every feature, input, output, setting, and visualization are found in the GitHub page and S1- supplementary information for the publication (Rivetti et al., 2023)

RELEVANCE OF INFORMATION SOURCE

G2P-SCAN has been developed with two main objectives: (1) to support the analysis of conservation of orthology and functional families to substantiate the identification of conservation and susceptibility of toxicity pathways across species and (2) to provide an automated tool for structured extraction of information from multiple databases and facilitate their synthesis. The tool is designed to leverage existing information on human biological processes, thereby enhancing the application of human toxicological data to other species' susceptibility assessments.

G2P-SCAN has been developed as a functionality driven orthology analysis pipeline to link genes into pathways across species. Its intended use is to assess orthology and functional families for human MIEs in order to evaluate the conservation and susceptibility at the pathway level across selected model species. The pipeline has been built to be synergistic and manually interoperable with other computational tools to facilitate the use of species extrapolation to drive intelligent testing strategies for ERA purposes. It is intended to provide valuable biological insights and better enable the use of mechanistic-based data to inform potential species susceptibility for research and safety decisions purposes, supporting the characterization of the domain of applicability of NAMs, such as cell-based assays, or for AOP development.

DESCRIPTION

G2P-SCAN is a pipeline developed as an R package which extracts, synthesizes, and structures data on gene orthologs, proteins and protein families, entities, and reactions, across model species from open-source databases. The tool can collate and visualize conservation of human pathways across mouse, rat, zebrafish, nematode, fruit fly and budding yeast, chosen based on the available orthology knowledge provided by database sources. Given a human gene(s) input, the pipeline extracts orthologs, proteins, protein families, entities and reactions for each model species organized by human pathways related to the input. The gene input should represent a molecular target as defined by a toxicologically relevant assay

or MIE. **The selection of this input is therefore subject to expert judgement.** Output of the pipeline summarizes collated data of species conservation as counts and percentages relative to human pathway counts for each data type. G2P-SCAN does not generate new data or make predictions but instead utilizes available data to gather, consolidate and present data in a quick and systematic way. The results from G2P-SCAN give relative coverage/conservation of identified human pathways which can provide weight-of-evidence support for assessment of chemical susceptibility across species.

The main steps of G2P-SCAN include.

1) Mapping human gene(s) input to biological pathways
2) Extracting pathway gene orthology across species
3) Extracting protein conservation and functionality (protein families, entities, and reactions)
4) Data organization and output

Analysis requires the human gene input to be queried, the pathway hierarchy level, and whether to extract all orthologs or only the least divergent orthologs (LDOs) (defined by PANTHER). The pathway hierarchy level options are parental, terminal, intermediate, or a combination of any. The selection of the "LDOs" option results in a stricter search and more confidence in the orthology result. Furthermore, this option provides a smaller number of genes, and so the overall runtime is reduced. When the LDOs option is not selected, all identified orthologs (LDOs and not) are assessed.

RESPONSE(S) MEASURED

The G2P-SCAN pipeline socializes data from 5 databases to calculate human relative coverage of pathways given by gene orthologs, proteins and protein families, and pathway entities and reactions. Integrating these data through the various evaluations using G2P-SCAN increases evidence of conservation.

Mapping to pathways

Relevant human biological pathways are retrieved from Reactome (https://reactome.org/) via HumanMine (https://www.humanmine.org/) Application Programming Interface (API) queries. Relevant human biological pathways are defined as those in which any of the input genes are found (based on Reactome latest available knowledge). The user will have the choice to select which pathway level(s) to include in the analysis across Parental, Intermediate and Terminal pathways. Pathways which contain any of the queried input genes are filtered based on the user '*pathwayLevel*' parameter input. The Reactome pathway hierarchy is used to determine the hierarchy level of the selection of pathways. Parental level pathways are defined as those which have children in the hierarchy of selected pathways but do not have any parent themselves. Terminal level pathways are defined as those which have parent pathways, but no children, and intermediate pathways have both parent and child pathways. The level of a pathway is determined by its relative position in the hierarchy considering only the pathways identified by the input genes and not based on the full Reactome hierarchy of all pathways.

Conservation of Orthologue genes

The number of human genes involved in identified biological pathways is given by the pathway gene list from Reactome (queried via HumanMine). Each human gene in a given pathway is queried in HumanMine to receive PANTHER (https://www.pantherdb.org/pathway/) orthologue annotations. The number of unique orthologs found for the genes in each of the human pathways are counted for each species and a gene coverage percentage is calculated as the number of unique orthologs counted relative to the number of human genes in the given pathway. The user defines whether all orthologs or only LDOs, defined by

PANTHER, should be retrieved, and included in the analysis. Filtering for only LDOs will mean only the most nearly "equivalent" gene in another organism is selected (in case of gene duplications for instance).

Conservation of Proteins

Using Uniprot (https://www.uniprot.org/) API queries, the pipeline identifies proteins for each of the identified human genes and model species' orthologs. Using a systematic filtering process, based on protein evidence and confidence defined in Uniprot, a single protein accession is selected per gene to be used in subsequent steps. Unique protein accessions are counted for each identified human pathway and model species. Protein coverage is calculated as a percentage of the number of unique protein identifiers found relative to the number of human proteins in each pathway.

Conservation of Functional Families

Protein accessions are used to map to protein function families defined by InterPro (https://www.ebi.ac.uk/interpro/) using Interpro's API functionality. Beyond families, InterPro also allows for the identification of higher and lower levels of clustering. Superfamilies are the higher aggregation of families and are not considered by the tool as they are thought to be too broad to retain specific functionality across species. Domains, repeats, and actives represent the lower groupings and are also not considered by the tool because they are regarded as too specific, especially when extrapolated across species. When multiple families are returned for a single protein search, only level 2 (i.e., the first child of the family hierarchical structure) families are recorded. These are uniquely counted across identified pathways and species. The family coverage percentage is counted as the number of unique functional level 2 families identified relative to the human family count per pathway per species.

Conservation of Entities and Reactions

Entities and reactions are inferred across species as these biological descriptors might vary. The terminology and information come from the Reactome database. Entities are defined as biological pathway components such as nucleic acids, proteins, complexes, and small molecules, thus making it a broader concept than the gene unit. Reactions are described as molecular events which link entities. Reactions are the core unit of the structure of the pathway database, Reactome. Entities participating in reactions form a network of biological interactions and are grouped into biological pathways. Together these biological pathway features contribute to the weight of evidence to support inferences of pathway conservation. This information is extracted directly from the Reactome database using an API connection to the species comparison analysis tool. This allows for the comparison of human pathways with computationally predicted pathways in other model organisms, highlighting the elements of the pathway that are common to both species and those that may be absent in the model organism. This automated, computational process is based on orthology. The number of entities and reactions are presented in the G2P-SCAN output per species per pathway.

PREDICTION MODEL

As outlined in the description, G2P-SCAN is a pipeline developed as an R package which extracts, synthesizes, and structures data on gene orthologs, proteins and protein families, entities, and reactions, across model species from open-source databases. The tool collates and visualizes conservation of human pathways across mouse, rat, zebrafish, nematode, fruit fly and budding yeast. Given a human gene(s) input, the pipeline extracts orthologs, proteins, protein families, entities and reactions for each model species organized by human pathways related to the input. Output of the pipeline summarizes collated data of species conservation as counts and percentages relative to human pathway counts for each data type. Overall predictions of chemical susceptibility for each species in G2P-SCAN can be evaluated by assessing the multiple lines of evidence from conservation across species of protein sequences (based on alignment) and their presence in the respective identified pathways, considering

protein families entities, and reactions **(it requires expert judgment to consider the outputs from each endpoint for each species of interest to determine the level of confidence for the prediction of conservation).**

Sources used within G2P-SCAN for generating predictions include:

HumanMine (www.humanmine.org/) (Kalderimis et al., 2014; Smith et al., 2012)

UniProt (www.uniprot.org) (UniProt Consortium, 2019)

InterPro's (www.interpro.org) (Blum et al., 2021)

Reactome via Reactome Analysis and Content Services (www.reactome.org) (Jassal et al., 2020)

METABOLIC COMPETENCE

G2P-SCAN offers a comprehensive tool for analyzing the conservation of biological processes across various species by integrating data from multiple databases and focusing on the molecular information available. Its analysis covers all mapped pathways in the Reactome database (https://www.reactome.org/), where 'metabolism' (R-HSA-1430728), 'metabolism of proteins' (R-HSA-392499.12) and 'metabolism of RNA' (R-HSA-8953854) are included as one of the 29 parental pathways controlling the human hierarchy of biological events, covering all biological processes related to the metabolic competence of an organism. Hence, depending on the user input, considerations of metabolic processes and their conservation across the species can be assessed. In fact, analysis performed using G2P-SCAN would include the evaluation of the conservation of gene orthologs, proteins, families, entities, and reactions for all the pathways nested under the metabolism-related pathways in Reactome, if relevant to the user query. The underlying assumption is that if the other species show conservation of the necessary metabolizing enzyme(s) and/or process(es), there is higher likelihood that the species may be susceptible or not, based on known metabolism outcomes in the human species. Furthermore, G2P-SCAN could also be used to assess the conservation of human chemical metabolizing enzymes that play a key role in the metabolism of drugs, toxins, environmental pollutants, and other substances in the organisms (e.g. Cytochrome P450). Looking at the conservation of the gene/protein sequence(s) across the 6 model species enables the understanding of whether the critical enzymatic domain(s) are still present in the orthologs despite their overall sequence conservation.

STATUS OF DEVLEOPMENT/STANDARDIZATION/VALIDATION

The approach has been published in a relevant scientific journal (i.e., Environmental Toxicology and Chemistry, Rivetti et al., 2023) in which the validity of the developed pipeline and its effectiveness as species extrapolation support has been demonstrated using five case studies. The latest source code for the G2P-SCAN library is available on GitHub (https://github.com/seacunilever/G2P-SCAN). During the final code development stage, a proof-of-concept manual run was performed to demonstrate the validity and robustness of the script in generating results (data available in the same publication as above).

TECHNICAL LIMITATIONS/LIMITATIONS WITH REGARD TO APPLICABILITY

There are several technical considerations in the application of G2P-SCAN in the context of regulatory decision-making. These are specifically concerning the number of species included in the analysis as well as the biological domain of applicability of the effects which are covered. It is important to consider that G2P-SCAN does not cover (non-human) species-specific effects as pathways are re-constructed using orthologue data based on existing human data. This means that non-human genes (e.g., those not present

in the Class Mammalia), pathways and processes, such as photosynthesis in plants or molting in invertebrates, are not covered and these effects would have to be considered separately. Additionally, the species queried by the tool are limited to 7 as those are included in all databases utilized by the tool (HumanMine, Reactome, Panther, UniProt and InterPro), of which HumanMine tends to be the most restrictive. Another consideration for application relates to the differences in species sensitivity to chemicals, due to differences in important factors such as the exposure, ADME, organism life stage, etc. which are out of the scope of the tool and not evaluated as elements of a qualitative analysis of cross-species conservation.

From a computational point of view, the tool is only executable through the R programming language and therefore requires some computational technical knowledge and a software to run. One limitation of the implemented approach of the tool is the reliance on the external updates of the database queried in the analysis, which may at instances not correspond to the most up-to-date database (e.g., the integrated version of Reactome data within the HumanMine database is not always equivalent to the live Reactome database). It is important to note that when HumanMine and API Reactome versions are not the same there is a risk the pathway identifiers will not match, and information could be lost and not processed by the pipeline. Those running the tool must be aware of version discrepancies between InterMine and live APIs used by G2P-SCAN. These are highlighted in log messages and output files of the tool. Furthermore, the tool requires a static file of the Reactome pathway and InterPro family hierarchy, which is downloaded from the API, if not available from file.

WEAKNESSES AND STRENGTHS

Strengths: A major advantage of G2P-SCAN is that it brings all the latest information together on gene orthologs, proteins and protein families, and pathway entities and reactions across species and databases into a single, user-friendly output. Enhancing our understanding of the taxonomic domain of applicability of biological processes through existing scientific evidence supports inferences of biological pathway conservation and helps facilitate more reliable cross-species extrapolation. G2P-SCAN enables this by capitalizing on and integrating a large amount of existing molecular/orthology data from humans and six environmentally relevant species, which have hitherto not been extensively applied collectively to inform decision making. As an open-source, user friendly R package that has no data confidentiality restrictions, the outputs can be readily supplemented by other available tools (e.g., SeqAPASS) and there are no maintenance requirements by the user for the underlying data sources. Ultimately, the value is in the combined use of existing tools and approaches to create additional lines of evidence to inform next steps in hazard characterization and support conclusions. In this respect, G2P- SCAN can be used to support prioritization and cutoff criteria during the screening stage of the ERA framework and to enable the design of intelligent and pragmatic testing strategies.

Weaknesses: Some limitations of G2P-SCAN relate to instances of low-affinity binding, off-target effects, and species-specific or unknown biology. Also, although it is pragmatic to assume that a biological process/pathway occurs in another species when all proteins and/or functions involved in the human version of the reaction are represented in the species, this may not be the case in reality because of the potential functional diversification of the ortholog candidates or changes in the expression pattern in other species. Conversely, the proposed approach may miss a conserved equivalent process because proteins may have evolved in their sequence while exerting the same function. The tool is also limited to humans and six environmentally relevant species and dependent on both the maintenance of the sources from which data are extracted and on the upkeep of relevant URLs to data sources.

RELIABILITY

The pipeline does not generate new data but obtains data from several public databases. These databases were chosen based on their curation, reliability, stability, and upkeep, which underpin the trustworthiness of the results.

The availability and reliability of the G2P-SCAN tool relies on the availability of an API for the 5 databases: HumanMine, Reactome, PANTHER, UniProt and InterPro. Due to the direct connection to the database APIs to obtain and collate data, the reproducibility of the pipeline can only be assessed when databases remain unchanged. The tool has a live connection to the tools listed and therefore any updates will be reflected in the results.

Since its publication in 2023, it has been cited 7 times, and a few studies have now used the pipelines and published results using its output. Specifically, the pipeline has been applied in this context in combination with the US Environmental Protection Agency Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool to develop a framework for using biological pathway information to enhance chemical susceptibility predictions across species (Schumann et al., 2024). Furthermore, a recently published study (Dufourcq Sekatcheff; Jeong and Choi, 2024) used the same combined approached to extend the biologically plausible tDOA of an AOP to over 100 taxonomic groups.

PREDICTIVE CAPACITY

The pipeline has been applied using several case studies including estrogen receptor alpha, acetylcholinesterase and butyrylcholinesterase and tryptophan hydroxylase in the following publication:

Genes-to-Pathways Species Conservation Analysis: Enabling the Exploration of Conservation of Biological Pathways and Processes Across Species (Rivetti et al., 2023)

> Summary: We present the novel pipeline Genes-to-Pathways Species Conservation Analysis (G2P-SCAN) to support understanding on cross-species extrapolation of biological processes. This R package extracts, synthetizes, and structures the data available from different databases, that is, gene orthologs, protein families, entities, and reactions, linked to human genes and respective pathways across six relevant model species. The use of G2P-SCAN enables the overall analysis of orthology and functional families to substantiate the identification of conservation and susceptibility at the pathway level. Five case studies are discussed, demonstrating the validity of the developed pipeline and its potential use as species extrapolation support.

Example of application of the pipeline has also been shown in:

Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species (Schumann et al., 2024).

> Summary: This work investigated the complementary use of two computational new approach methodologies to support cross-species predictions of chemical susceptibility: the US Environmental Protection Agency Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool and Unilever's recently developed Genes to Pathways – Species Conservation Analysis (G2P-SCAN) tool. These stand-alone tools rely on existing biological knowledge to help understand chemical susceptibility and biological pathway conservation across species. The utility and challenges of these combined computational approaches were demonstrated using case examples focused on chemical interactions with peroxisome proliferator activated receptor alpha (PPARα), estrogen receptor 1 (ESR1), and gamma-aminobutyric acid type A receptor subunit alpha (GABRA1). Overall, the biological pathway information enhanced the weight of evidence to support cross-species susceptibility predictions. Through comparisons of relevant

molecular and functional data gleaned from adverse outcome pathways (AOPs) to mapped biological pathways, it was possible to gain a toxicological context for various chemical-protein interactions. The information gained through this computational approach could ultimately inform chemical safety assessments by enhancing cross-species predictions of chemical susceptibility. It could also help fulfill a core objective of the AOP framework by potentially expanding the biologically plausible taxonomic domain of applicability of relevant AOPs.

Bridging the Gap Between Human Toxicology and Ecotoxicology Under One Health Perspective by a Cross-Species Adverse Outcome Pathway Network for Reproductive Toxicity (Dufourcq Sekatcheff, Jeong and Choi, 2024)

Summary: This study aims to extend the biologically plausible taxonomic domain of applicability (tDOA) of AOP 207 (AOPwiki ID 207: NADPH oxidase and P38 MAPK activation leading to reproductive failure in Caenorhabditis elegans).. Various types of data, including in vitro human cells, in vivo, and molecular to individual, from previous studies have been collected and structured into a cross-species AOP network that can inform both human toxicology and ecotoxicology risk assessments. The first step included the collection and analysis of literature data to fit the AOP criteria and build a first AOP network. Then, key event relationships were assessed using a Bayesian network modeling approach, which added confidence in the overall AOP network. Finally, the biologically plausible tDOA was extended using in silico approaches (Genes-to-Pathways Species Conservation Analysis and Sequence Alignment to Predict Across Species Susceptibility), which led to the extrapolation of our AOP network across over 100 taxonomic groups. This approach showed that various types of data can be integrated into an AOP framework, thus facilitating access to knowledge and prediction of toxic mechanisms without the need for further animal testing.

PROPRIETARY ASPECTS

The tool is a freely downloadable R package to mine existing open-access databases.

PROPOSED REGULATORY USE

The development of the G2P-SCAN pipeline represents another meaningful leap forward towards integrating computational biology approaches into safety assessments. By synthetizing and integrating data from multiple databases, G2P-SCAN offers a comprehensive tool for analyzing the conservation of biological processes across various species. The methodology's significance lies in its potential to improve the accessibility and synthesis of genomic data, thus facilitating the application of mechanistically based data for ERA purposes and data-gaps filling. This approach aligns with global regulatory shifts towards new approach methodologies (NAMs), promoting the use of computational and cell-based approaches in safety assessments and supporting the use of alternatives to animal testing by enabling more substantiated species extrapolation.

It is foreseen that G2P-SCAN can be used to support prioritization and cutoff criteria during the screening stage of the ERA framework and to enable the design of intelligent testing strategies which are pragmatic, maximizing the use of all the available knowledge and including a better informed testing strategies and data generated in other nonstandard species (i.e., humans) as support for environmental safety decision making. In addition, it can be relevant in supporting, identifying, and leveraging existing data to help evaluate a chemical's potential hazard by characterizing the susceptibility of key model organisms of

environmental relevance and building evidence for the conservation of biological pathways. Overall, this will support a reduction in the number of animals used for toxicity testing purposes.

REFERENCES

Scientific Publications:

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., & Raj, S. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Research, 49(D1), D344–D354.

Dufourcq Sekatcheff, E., J. Jeong and J. Choi (2024), "Bridging the gap between human toxicology and ecotoxicology under one health perspective by a cross-species adverse outcome pathway network for reproductive toxicity", Environmental Toxicology and Chemistry, Vol. 44/9, pp. 2511–2523, https://doi.org/10.1002/etc.5940.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., & Haw, R. (2020). The Reactome pathway knowledgebase. Nucleic Acids Research, 48(D1), D498–D503.

Kalderimis, A., Lyne, R., Butano, D., Contrino, S., Lyne, M., Heimbach, J., Hu, F., Smith, R., Štěpán, R., & Sullivan, J. (2014). InterMine: Extensive web services for modern biology. Nucleic Acids Research, 42(W1), W468–W472.

Rivetti, C., Houghton, J., Basili, D., Hodges, G., & Campos, B. (2023). Genes ‐ to ‐ Pathways Species Conservation Analysis: Enabling the Exploration of Conservation of Biological Pathways and Processes Across Species. Environmental Toxicology and Chemistry, 42(5), 1152–1166.

Schumann P, Rivetti C, Houghton J, Campos B, Hodges G, LaLone C. Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species. Science of The Total Environment. 2024 Feb 20;912:168573.

Smith, R. N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., Lyne, M., Lyne, R., Kalderimis, A., & Rutherford, K. (2012). InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics, 28(23), 3163–3165.

UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506–D515.

Online Tools and Databases:

Seacunilever. Genes-to-Pathways Species Conservation Analysis (G2P-SCAN). Retrieved from https://github.com/seacunilever/G2P-SCAN [Accessed: October 2025]

HumanMine. HumanMine Database. Retrieved from https://www.humanmine.org/ [Accessed: October 2025]

InterPro. InterPro Database. Retrieved from https://www.ebi.ac.uk/interpro/ [Accessed: October 2025]

PANTHER. PANTHER Pathway Database. Retrieved from https://www.pantherdb.org/pathway/ [Accessed: October 2025]

Reactome. Reactome Pathway Database. Retrieved from https://reactome.org/ [Accessed: October 2025]

UniProt. UniProt API. Retrieved from https://www.uniprot.org/ [Accessed: October 2025]

Annex B. REPORTING INDIVIDUAL INFORMATION SOURCES for Sequence Alignment to Predict Across Species Susceptibility

NAME OF THE INFORMATION SOURCE

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS v.7.1 for case studies developed in ANNEX II. SeqAPASS v8.0 is the current version ; https://seqapass.epa.gov/seqapass/)

SeqAPASS User Guide: https://www.epa.gov/comptox-tools/seqapass-user-guide

- Screenshots directly from the SeqAPASS tool and detailed descriptions of every feature, input, output, setting, and visualization are found in the User Guide

RELEVANCE OF INFORMATION SOURCE

The web-based SeqAPASS tool is intended to generate lines of evidence of protein target conservation and chemical-protein interaction conservation across hundreds or thousands of species rapidly. These results are used to predict chemical susceptibility across species and to inform the definition of the biologically plausible tDOA for MIEs and early KEs in the AOP framework (LaLone et al., 2016; Jensen et al., 2023; Haigis et al., 2023). Importantly, the first three levels of analysis were developed so that anyone with a background in basic biology could complete a meaningful species extrapolation evaluation with limited training. Evidence is collected to evaluate protein conservation comparing protein sequence and structure for a known chemical-protein (in the context of the MIE) or protein-protein interaction (for KEs) in a sensitive species (e.g., the chemical was designed to target the species/model organism used in an assay and/or to describe the KE, empirical data have been used to demonstrate a chemical-induced response in the species) to all other species for which sequence information is available. The level of detail in the analysis is dependent upon how well the protein has been characterized and how much is known about the protein-chemical interaction. The tool can be applied for extrapolation of empirical knowledge (Ankley et al., 2016; LaLone et al., 2018). The assumption is that if the protein is present in another species, it is likely available for the chemical stressor to interact in the MIE or for the protein cascade to occur in the KE. There are four levels of evaluation in SeqAPASS described below. At each of these four levels additional lines of evidence supporting the likelihood of chemical-protein interactions are incorporated. Additional knowledge such as in vitro and in vivo evidence can be incorporated to support the computational evidence of conservation, though empirical toxicity data is typically limited across species.

DESCRIPTION

The SeqAPASS tool was developed to extrapolate chemical toxicity knowledge/data from one species to others. This extrapolation is termed 'cross species extrapolation' and is based on the current understanding of a chemical-protein or protein-protein interaction in one species and the collection of lines of evidence indicating a similar interaction in another species is likely to occur. The results from SeqAPASS provide evidence of protein conservation to be used for predictions of chemical susceptibility or pathway conservation across the diversity of species. This is important because most species will never be tested for toxicity in the laboratory, though researchers and regulators are interested in understanding the possible chemical impacts on all species that have potential for exposure in the environment.

Specifically, the tool takes advantage of existing knowledge of a chemical causing an effect through a particular protein in a given species (usually a species known to be sensitive to a chemical-induced response) or a protein involved in an assay where a test species is represented. The protein is queried for that species using the SeqAPASS tool to generate lines of evidence of conservation in other species. Therefore, the necessary input to SeqAPASS is the combination of a protein and an individual species. Importantly, SeqAPASS can be used to extrapolate from any species that has quality protein information available.

There are three levels to the SeqAPASS evaluation available to all users. Level 1 compares primary amino acid sequences, Level 2 compares functional domains, and Level 3 compares critical individual amino acids across species. Each level provides an additional line of evidence supporting potential conservation and yielding a prediction of susceptibility, which is equivalent to saying the protein is conserved in other species and the chemical is likely to interact similarly to the query species. Users can choose to run Level 2 or Level 3 comparisons if there is enough knowledge available relative to functional domains and critical amino acids. There are information buttons integrated throughout the tool to guide users through this efficient analysis as well as visualizations, summary reports, and strategic interoperability with other valuable resources for interpreting the results for cross species extrapolation.

The Level 4 evaluation is intended for advanced users and experts in protein structural biology. Level 4 generates protein structural models to perform structural alignments for an additional line of evidence of protein conservation (LaLone et al., 2023). Such results can also be exported for more advanced bioinformatics approaches like molecular docking, virtual screening, or molecular dynamic simulations (Schumann et al., 2024).

RESPONSE(S) MEASURED

Protein sequence and structural similarity metrics are used to determine conservation of proteins across species. Conservation includes capturing evidence of the presence or absence of a protein in another species and whether there is evidence that the specific chemical interaction is similar between species. Ortholog sequences are identified to allow for the detection of proteins with the greatest likelihood for functional similarity. The SeqAPASS algorithms mine, collect, and collate information from the National Center for Biotechnology Information (NCBI) protein database (http://www.ncbi.nlm.nih.gov/protein/), conserved domains database (http://www.ncbi.nlm.nih.gov/cdd/), taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy/), strategically utilizes the Stand-Alone Basic Local Alignment Search Tool for proteins (BLASTp) (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download) and the Constraint-based Multiple Alignment Tool (COBALT) (http://www.st-va.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi). Advanced users can request Level 4 access for protein structure generation and alignment to gather additional lines of evidence toward conservation. Tools used for structural evaluation include: Iterative Threading ASSEmbly Refinement (I-TASSER,

https://zhanggroup.org/I-TASSER/) for creation of protein structural models, AlphaFold (https://AlphaFold.ebi.ac.uk/) and other sources using PDB formatted protein structures for structural alignment using TM-align (https://zhanggroup.org/TM-align/).

Level 1

The SeqAPASS algorithms submit the query protein sequence selected by the user (**Requires expert judgment to select appropriate protein target**) to NCBI's standalone BLASTp (using default settings, including BLOSUM-62 matrix), which aligns the query protein with all proteins available in the NCBI protein database and provides a variety of metrics associated with each pairwise alignment between the query and hit sequences. SeqAPASS selectively captures output from BLASTp, including one sequence per species with the highest bit score. Detailed descriptions of metrics derived from BLASTp (e.g., BLASTp Bitscore, E-Value, Positives, Identity, Hit length) can be found in: The NCBI Handbook: (http://www.ncbi.nlm.nih.gov/books/NBK21106/); BLAST® Help: (http://www.ncbi.nlm.nih.gov/books/NBK62051/) and the NCBI Glossary Field Guide: (http://www.ncbi.nlm.nih.gov/Class/FieldGuide/glossary.html)

Common Domain Count

Reversed Position Specific BLAST (RPS BLAST) is used to compare each query and hit sequence to conserved domains defined in NCBIs Conserved Domain Database. A hit domain is considered in common with the query domain if it contains the same domain accession as the query and it aligns with the NCBI curated domain with the same or greater amino acid residue coverage than the query sequence.

Ortholog Candidate Identification

Ortholog sequences are those that have diverged from a speciation event and therefore are more likely to maintain similar function. SeqAPASS uses reciprocal best hit (RBH) BLAST for ortholog detection by automatically comparing each hit protein to all protein sequences available for the query species and if the original query protein or one of its identical protein matches is identified to be the best match to the hit or maintain the same bitscore, then the hit sequence would be considered an ortholog candidate. The sequence is indicated as an Ortholog Candidate or not with a yes (Y) or no (N) in the Level 1 data table.

Susceptibility cut-off

The susceptibility cut-off values listed on the "Level 1 (and Level 2) Susceptibility Cut-off" page are determined by plotting the % similarity data from the "Primary Report" or "Full Report" and identifying the local minimums in the data. The default cut-off is determined by taking the 1st local minimum and moving up in percent similarity until the next ortholog candidate is found. The susceptibility cut-off displayed in the list is the percent similarity of the identified ortholog candidate.

Criteria for Susceptibility Prediction (when "Primary Report Settings" is set to "Species Read Across:" Yes)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes". If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for "yes", then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes".

If the hit sequence is below the susceptibility cut-off but belongs to any organism class found above the susceptibility cut-off, the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for "yes". This criterion allows susceptibility predictions to be made across taxonomic groups based on the likelihood that the sequences above the cut-off are better matches to the query. If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for "no") and does not belong to any organism class found above the susceptibility cutoff, the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for "no".

Criteria for Susceptibility Prediction (when "Primary Report Settings" is set to "Species Read Across:" No)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for “yes”, then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for “no”), the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for “no”.

Level 2

Data obtained from the Level 1 RPS BLAST evaluation is used to assign sequence ranges that aligned with a user selected domain (from the NCBI CDD database) to each accession from the Level 1 Full report (**Requires expert judgment to select appropriate domain**). BLASTp is then used to align the query domain range to each hit domain range. The percent similarity is calculated based on the bit scores from the BLASTp alignment of the domain regions. For each sequence queried in Level 2, the top row query species is used to determine the maximum bitscore for the analysis, which is derived from aligning the query sequence to itself using BLASTp. To calculate percent similarity, the bitscore for each hit sequence is normalized to the maximum bit score and then multiplied by 100.

Level 2 Susceptibility cut-off (same method as used in Level 1)

The susceptibility cut-offs listed on the “Level 2 Susceptibility Cut-off” page are determined by plotting the % similarity data from the “Primary Report” or “Full Report” and identifying the local minimums in the data. The default cut-off is determined by taking the 1st local minimum and moving up in percent similarity until the next ortholog candidate is found. The susceptibility cut-off displayed in the list is the percent similarity of the identified ortholog candidate.

Level 2 Criteria for Susceptibility Prediction (when “Primary Report Settings” is set to “Species ReadAcross:” Yes)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for “yes,” then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. If the hit sequence is below the susceptibility cut-off but belongs to any organism class found above the susceptibility cut-off, the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. This criterion allows susceptibility predictions to be made across taxonomic groups based on the likelihood that the sequences above the cut-off are better matches to the query. If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for “no,”) and does not belong to any organism class found above the susceptibility cutoff, the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for “no” Note that the “Primary Report” may yield different Susceptibility Predictions than the “Full Report,” as the predictions are based on the data in the different reports. The Primary Report is filtered to only display E-value ≤ 0.01 and Common Domain Count ≥ 1.

Level 2 Criteria for Susceptibility Prediction (when “Primary Report Settings” is set to “Species ReadAcross:” No)

All sequences identified above the susceptibility cut-off are predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. If the hit sequence is below the susceptibility cut-off, but identified as an Ortholog Candidate = Y, for “yes,” then the hit is predicted to be susceptible; therefore, Susceptibility Prediction = Y for “yes”. If the hit sequence is below the susceptibility cut-off and not identified as an ortholog candidate (Ortholog Candidate = N, for “no,”), the hit is predicted to not be susceptible; therefore, Susceptibility Prediction = N for “no”

Level 3

COBALT is used to align all user selected sequences (from Level 1 hits) with a user defined template sequence (Requires expert judgment to select appropriate critical amino acids, typically those known to interact with the chemical of interest). Because COBALT algorithms align all sequences, it is recommended that the user align the template sequence with sequences that are most similar to one another. As a means to capture the most similar sequences from the SeqAPASS data it is recommended that the user filter the Level 1 data by taxonomic group and step through the Level 1 data pages one by one while selecting sequences. It is recommended that the user look at the name of the sequence and exclude 'partial' sequences when possible (Requires expert judgment to remove low quality, partial sequences, etc., however there is a feature in SeqAPASS that prioritizes high quality sequences for the user to consider). Requesting a query from one taxonomic group at a time breaks the data down in manageable alignments.

Level 3 Selecting Amino Acid Residues to Align

The user may select up to 50 amino acid residues to compare across selected species in Level 3 (**Requires expert judgment to select and submit the critical amino acids for cross species comparisons**). The SeqAPASS tool was developed to automatically compare the identity of amino acids for each selected species at selected protein sequence positions against the template species sequence for common side chain classification and molar mass. Therefore, species that have key amino acid residues that share the same side chain classification and/or have a molar mass within an absolute value of 30 g/mol are identified as similar and predicted to share susceptibility to chemicals with the template species. Species that have one or more key amino acid residues that do not share the same side chain classification and have a difference in molar mass of 30 g/mol or greater relative to the template sequence are predicted as less likely of sharing susceptibility to chemicals with the template species. Requiring one or more key amino acids to share either the same side chain classification or molar mass as the template species was used for determination of chemical susceptibility in order to produce conservative predictions, as dramatic differences among amino acid residues are more likely to change the protein-chemical interaction relative to minor differences.

Level 4 Structural Alignments

Level 4 is intended for use by experts in protein structural biology, only. The I-TASSER standalone program (Ver 5.1; https://zhanggroup.org/I-TASSER/) was integrated in SeqAPASS to generate protein structural models by submitting the aligned FASTA sequences from the SeqAPASS using the respective hit proteins and setting a restraint (i.e., a specific structure suggested by the user as a template to be considered in the threading process) from those identified in the Protein Data Bank (PDB; https://www.rcsb.org/).

Specifically, the PDB, which archives available protein structural data, is queried to identify any available crystal structures that would be suitable for representing the query proteins from the SeqAPASS analysis, as well as assist in identifying ligand-binding coordinates on the proteins of interest (**Requires expert judgment to select appropriate protein structure as a restraint**). Priority protein sequences from SeqAPASS output are then submitted as FASTA to I-TASSER with PDB restraints to generate protein structural models and collect metrics relative to the quality of the protein structure generated (**Requires expert judgment to select quality and accurately annotated sequences to generate protein structures**).

Tables are automatically generated by importing a list of proteins and results from a SeqAPASS Level 1 report with assigned priorities, and the selected proteins can then be submitted to I-TASSER. The SeqAPASS Level 4 features then capture metrics associated with each protein structure model from I-TASSER output (including confidence score, TM-score, RMSD, and the number of decoys). Outputs include confidence score (C-score) for estimating the quality of predicted models based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations. The C-score is commonly between −5 and 2, where the greater the value, the higher the confidence in the model. The TM-score is a metric for assessing the topological similarity of protein structures and has a value between 0 and 1, where 1 indicates a perfect match between two structures. A TM-score >0.5

indicates a model of correct topology, whereas a TM-score <0.17 indicates a random similarity. The root mean square deviation (RMSD) of atomic positions is the measure of the average distance between the atoms of superimposed proteins. Therefore, the lower the RMSD, the closer the model is to the target structure. The number of protein structure decoys, which are the artificial structural conformations of proteins used to guide the design, testing, and training of the protein folding force fields, is reported for each model. The cluster density is also reported and used to define the number of structure decoys at a unit of space in the cluster. A higher cluster density means the structure occurs more often in the simulation trajectory and is therefore likely a higher-quality model. Models are generated and saved as PDB files for further analysis.

Level 4 TM-align

The protein structural models from I-TASSER can then be compared using the integrated TM-align (https://zhanggroup.org/TM-align/) feature in SeqAPASS Level 4 (**Requires expert judgment to evaluate the metrics associated with the predicted structures from I-TASSER to decide which should be aligned in TM-align**). TM-align compares two protein structures generating optimized amino acid residue alignment based on structural similarity using heuristic dynamic programming iterations. Output from TM-align includes optimal superpositions of the two compared structures as PDB files and the TM-score for each structural alignment. Those TM-scores <0.2 correspond to randomly chosen unrelated proteins, whereas those >0.5 assume generally the same fold. The TM-align susceptibility prediction is "Y", indicating yes, when the Average Percent Similarity is greater than 20% and a "N", indicating no, when less than 20%. (**Requires expert judgment to examine the structures thoroughly and determine whether the predictions of yes or no are appropriate**).

PREDICTION MODEL

Predictions for chemical susceptibility or protein conservation are derived from several sources integrated in the SeqAPASS workflow. The Decision Summary Report brings together predictions of susceptibility based on conservation from Level 1-3 analyses (Level 4 will be integrated during FY25). Overall predictions of chemical susceptibility for each species in SeqAPASS are generated through multiple lines of evidence from protein sequence and structural alignments as described above (**Requires expert judgment to consider the SeqAPASS from each Level for each species of interest to determine the level of confidence for the prediction of susceptible yes or no**).

Sources used within SeqAPASS for generating predictions include:

NCBI Protein Database(http://www.ncbi.nlm.nih.gov/protein/) (Pruitt et al., 2007)

NCBI Taxonomy Database(http://www.ncbi.nlm.nih.gov/taxonomy/) (Federhen, 2012)

NCBI Basic Local Alignment Search Tool for Proteins (BLASTp) (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download) (Altschul et al., 1990)

NCBI Conserved Domain Database (http://www.ncbi.nlm.nih.gov/cdd/) (Marchler-Bauer et al., 2012).

NCBI Constraint Based Multiple Alignment Tool (COBALT) (http://www.st-va.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi) (Papadopoulos & Agarwala, 2007).

Environmental Conservation Online System (ECOS) (https://ecos.fws.gov/ecp/)

ECOTOX Knowledgebase (Olker et al., 2022)

CompTox Chemicals Dashboard (Williams et al., 2017)

Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (Rose et al., 2016)

Iterative Threading ASSEmbly Refinement (I-TASSER) (https://zhanggroup.org/I-TASSER/) (Yang & Zhang, 2015)

AlphaFold (https://AlphaFold.ebi.ac.uk/) (Varadi et al., 2022)

TM-align (https://zhanggroup.org/TM-align/) (Zhang & Skolnick, 2005)

METABOLIC COMPETENCE

SeqAPASS is an online tool for evaluating protein conservation and predicting subsequent conservation of chemical-induced interactions across a range of species. In addition to interactions between parent chemical and a molecular target, the effects of chemical metabolites could be considered using this software by evaluating the degree of protein conservation of metabolizing enzymes across species, if those data are available.

Metabolic considerations in the context of SeqAPASS include evaluation of protein conservation of primary metabolizing enzymes known to metabolize a particular chemical. The assumption is that if other species have the metabolizing enzyme(s) necessary they are likely to be similarly susceptible or not as susceptible based on known metabolism outcomes in a sensitive or insensitive species, respectively. It is noteworthy that analyses of metabolizing enzymes using SeqAPASS are difficult for a few reasons. First, it is rarely the case where one specific enzyme (with a specific isoform) is known to be solely responsible for the enzymatic activity. Second, enzyme sequence data is lacking for the majority of species. Finally, where enzyme sequence data exists it may not be well annotated. Regardless, focusing on the current state of the science for the primary metabolizing enzyme(s) for a particular chemical in a particular species provides the starting information to consider conservation in predicting chemical susceptibility.

In addition, the SeqAPASS tool can be used for hypothesis generation in cases where differences in sensitivity are initially assumed to be sequence related, but analyses indicate complete protein conservation. In such instances the driver of sensitivity could be due to metabolism, gene expression, or a variety of other factors. Such information from SeqAPASS results can guide users to look for other explanations for differences in chemical sensitivity outside of protein target conservation.

STATUS OF DEVELOPMENT/STANDARDIZATION/VALIDATION

The initial release of SeqAPASS version 1.0 by the US EPA was a web-based tool that allowed for Level 1 and Level 2 analyses. Each year since, the SeqAPASS development team releases a new version of the interface with new features requested by stakeholders. The current version as of January 2025 is SeqAPASS v8.0. The SeqAPASS tool has been described as a tool for cross species evaluation that is ready for use in a nonanimal testing regulatory environment. In 2018, the OECD published the Revised Guidance Document 150 on Standardized Test Guidelines for Evaluating Chemicals for Endocrine Disruption, which recommended the SeqAPASS tool for evaluation of protein conservation for cross species extrapolation (https://www.oecd.org/publications/guidance-document-on-standardised-test-guidelines-for-evaluating-chemicals-for-endocrine-disruption-2nd-edition-9789264304741-en.htm). In 2023 the United States Environmental Protection Agency (US EPA) Office of Chemical Safety and Pollution Prevention (OCSPP) announced the release of a draft White Paper titled “Availability of New Approach Methodologies (NAMs) in the Endocrine Disruptor Screening Program (EDSP)” for public comment via the Federal Register (https://www.epa.gov/sciencematters/epa-research-contributes-using-alternatives-screen-chemicals-endocrine-disruption). The paper describes validated NAMs that could be used as alternatives for certain EDSP tests. Specifically, it highlights the SeqAPASS approach for use by the US EPA in weight of evidence

evaluations. The validation of SeqAPASS outputs have been described in multiple publications and analyses have evolved over time. Data comparisons initially were between sequence similarity and the geometric mean of the LC50 and has incorporated comparisons to *in vitro* AC50 data and even site-directed mutagenesis studies. Consistently the results have indicated high correlation between the SeqAPASS predictions and empirical toxicity data (Russom et al., 2014; Hauck et al., 2021; Mayasich et al., 2023). The publications listed in the PREDICTIVE CAPACITY section below highlight multiple case studies demonstrating the development/standardization/validation of the SeqAPASS tool, methods, output, and comparisons to and combination with empirical toxicity data.

TECHNICAL LIMITATIONS/LIMITATIONS WITH REGARD TO APPLICABILITY

The potential negative impacts of chemicals on the environment are evaluated from a handful of model species. In general, these species were selected because of ease of maintenance in laboratory settings, rapid and consistent reproduction, and there is an ability to readily control diet and surroundings, characteristics which may not make them ideal surrogates for the tens of thousands of naive species that may be exposed to chemicals. SeqAPASS provides a computational tool to assesses whether there are adequate lines of scientific evidence to support extrapolation of effects measured in the few model animals to a broad range of taxa, and alternatively, identify groups of animals that may not be adequately predicted due to lack of sequence conservation of critical protein targets.

There are technical considerations in the application of the SeqAPASS tool, specifically centered around the domain of applicability. The SeqAPASS tool requires an understanding of a chemical's molecular target in a known sensitive or model species. If the protein target has not been defined and is unknown for a given chemical, or the chemical does not interact with a specific protein, the SeqAPASS tool cannot be applied for cross species extrapolation in the context of regulatory decision-making. In addition, each level of SeqAPASS analysis requires greater detail about the chemical-protein or protein-protein interaction. Therefore, if a protein has not been well characterized, the user may only be able to perform a Level 1, primary amino acid sequence alignment where predictive capabilities are at the level of comparing between vertebrates, invertebrates, and plants. Level 2, which aligns functional domains adds greater taxonomic resolution to the predictions of chemical susceptibility at the defined taxonomic lineage (e.g., Class, Order). As the user moves to Level 3, requiring information on specific amino acid interactions with the chemical or Level 4, requiring empirically derived crystal structures with the chemical of interest bound to the receptor, the predictions become species specific, though data and knowledge requirements become greater as well. Ultimately, the SeqAPASS method is limited by the degree of characterization of the protein, the availability and quality of the sequence information, and the annotation quality for the protein(s) of interest. For Level 4, the approach is limited by the available empirically derived crystal structures with chemicals of interest bound.

WEAKNESSES AND STRENGTHS

Application of this tool is predicated on the recognition that conservation of a molecular target is but one key component for predicting susceptibility across species; there are multiple other contributing factors related to the exposure route, intensity (dose) and timing, as well as ADME that also play key roles in potential susceptibility. The SeqAPASS application is not intended to account for these additional factors in making predictions of susceptibility; rather, SeqAPASS results are envisioned to be integrated with other data sources to inform risk assessment or study design.

The SeqAPASS tool provides a platform for computational predictions of intrinsic susceptibility that are supported by concepts in evolutionary biology where multiple case examples have been published that compare predictions to available empirical results. In addition, SeqAPASS is free and publicly available on

a well-supported web-based platform that is widely accessible (https://seqapass.epa.gov/seqapass/). Because SeqAPASS leverages sequence data and protein information from existing databases, the ability of SeqAPASS to predict chemical susceptibility to a broader diversity of species is constantly improving as sequencing technology advances and the genomes of new species are sequenced and annotated. Although this offers distinct advantages regarding data availability, publicly available sequence information can also be challenging to assess due to inconsistent quality, poor annotation, and incompleteness of protein sequences for some species. However, it is promising that omics technologies and methods in bioinformatics are advancing rapidly, allowing sequence curation and quality to continue to improve over time.

One major goal of the SeqAPASS tool is transparency, providing links to all data sources and tools that are integrated in the back end of SeqAPASS. Such transparency allows the user rapid access to original sources of the sequence or taxonomic information from NCBI. The domain of applicability for the SeqAPASS tool is defined by the information needed to conduct a meaningful SeqAPASS analysis. Since knowledge of a chemical-protein or protein-protein interaction in a known sensitive or targeted species are key elements to begin a SeqAPASS query, it must be acknowledged that queries conducted without this information are not meaningful. Additionally, chemicals that have multiple biological targets, or interact with different targets with differing degrees of potency, also present a challenge in combining SeqAPASS results from the various targets. However, those with multiple targets can be evaluated in SeqAPASS by submitting each protein target as a query and then combining the results to understand where there is overlap in conservation.

Summary of SeqAPASS strengths:

- Rapid prediction of chemical susceptibility to hundreds or thousands of untested species
- Publicly accessible and US EPA maintained web-based tool
- Published User Guide and numerous published case studies to demonstrate application
- SeqAPASS results have been integrated in regulatory decision-making processes
- Takes advantage of rapidly expanding protein sequence and structure data that does not require the destruction of animals
- Transparent and consistent methods based on state-of-the-science bioinformatics approaches
- Levels 1-3 of SeqAPASS analysis is intended for users with basic biology background
- Experts in protein structural biology have the option to include structural similarity in their evaluation of protein conservation using Level 4 analysis
- Integrated features guide users to necessary data and inputs as well as prioritize and highlight sequences that are likely of highest quality
- Training is available both online and in-person with developers of the tool

Summary of SeqAPASS limitations:

- Limited by available protein sequence and structural information
- Limited to evaluation of chemicals where the molecular target is known
- Considers only conservation of protein target and chemical-protein interaction to make predictions of chemical susceptibility
- Currently cannot predict the degree of susceptibility

RELIABILITY

The SeqAPASS approach to evaluating protein conservation is reliable since the methods for evaluating proteins are streamlined, transparent, and consistent. Additionally, the source data are continuously documented and links to the source data are provided throughout the tool. As new versions of the SeqAPASS tool are released to the public, the data and interface are rigorously and consistently tested using a well-defined and published SeqAPASS Testing Standard Operating Procedure. Efforts are also underway to automate interface testing through the development process using the Cypress tool. Command line has been generated for testing Level 1-3 using Cypress and R code has been developed for consistently comparing data used in the SeqAPASS backend from minor release to minor release. Consistently, 30 protein targets from a variety of species are submitted to SeqAPASS to compare new data versions to old data versions and ensure any changes in the data are due to updates in the NCBI protein, taxonomy, or CDD databases or due to updates in the executables used for comparing sequences and structures. Further, the SeqAPASS development team continues to publish case examples to demonstrate proper use and interpretation of results in aligning with current regulatory challenges. These publications include decision points where scientific justification must be demonstrated for changing any of the default settings. (LaLone et al., 2022)

PREDICTIVE CAPACITY

Published examples of the predictive capacity of SeqAPASS are listed below:

1. Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species (Schumann et al., 2024)

- Summary: This publication is described in detail in ANNEX II

2. From Protein Sequence to Structure: The Next Frontier in Cross Species Extrapolation for Chemical Safety Evaluations (LaLone et al., 2022)

- Summary: This publication lays the foundation for expanding the capabilities of the web-based SeqAPASS tool to include structural comparisons for species extrapolation, facilitating more rapid and efficient toxicological assessments among species with limited or no existing toxicity data.For the past decade there have been enhanced efforts to use knowledge of protein conservation to predict chemical susceptibility across species using bioinformatics approaches. These approaches have primarily focused on protein sequence comparisons yielding predictions of susceptible “yes” or “no” across the diversity of species in considering the likelihood for similar chemical-protein interactions. The intent of the predictive methods are to enable extrapolation from toxicity knowledge in one species to many species based on molecular similarities in the context of regulatory decision-making. Considering the desire to use predictive computational approaches for species extrapolation in chemical safety evaluations it is essential to take steps to incorporate functional prediction and to consider how the science may advance toward quantitative metrics to predict the degree of susceptibility among species aligning with needs in risk assessment. Fortunately, a new era is dawning in protein structural prediction with the incorporation of advanced threading approaches and the use of artificial intelligence. From the Critical Assessment of protein Structure Prediction (CASP) experiments, there have been consistent improvements in approaches and tools for identifying protein structure from sequence. Tools such as Iterative Threading ASSEmbly Refinement (I-TASSER) and AlphaFold have been of particular interest in the field, consistently making advances to achieve accuracy in their predictions of protein structure. With such improvements in structural prediction methods the US EPA has begun exploration of protein structural conservation to enhance current sequence-based predictive methods for species extrapolation in the context of chemical safety. Release of SeqAPASS v7.0, which includes the

ability to now compare protein sequence and generate protein structures across species, enables users to add lines of evidence toward protein conservation for predicting chemical susceptibility across species. Further, with these new capabilities the SeqAPASS team has moved toward the use of molecular docking to further understand binding and capture multiple metrics that advance the work in a direction toward more quantitative predictions. ***Two case examples illustrating this pipeline from SeqAPASS sequences to I-TASSER-generated protein structures were created for human liver fatty acid–binding protein (LFABP) and androgen receptor (AR).*** Ninety-nine LFABP and 268 AR protein models representing diverse species were generated and analyzed for conservation using template modeling (TM)-align. The results from the structural comparisons were in line with the sequence-based SeqAPASS workflow, adding further evidence of LFABL and AR conservation across vertebrate species.

3. Cross-species applicability of an adverse outcome pathway network for thyroid hormone system disruption (Haigis et al., 2023)

    - Summary: This review advanced the description of the taxonomic domain of applicability (tDOA) in the adverse outcome pathwy network describing thyroid hormomne system discuption (THSD) to improve its utility for cross-species extrapolation. The paper focused on 32 AOPs and focused on 13 MIEs and adverse outcomes (AOs). It evaluated both their plausible domain of applicability using SeqAPASS (taxa they are likely applicable to) and empirical domain of applicability (where evidence for applicability to various taxa exists) in a THSD context. The evaluation showed that all MIEs in the AOP network are applicable to mammals. With some exceptions, there was evidence of structural conservation across vertebrate taxa and especially for fish and amphibians, and to a lesser extent for birds, empirical evidence was found. Current evidence supports the applicability of impaired neurodevelopment, neurosensory development (eg, vision) and reproduction across vertebrate taxa. The results of this tDOA evaluation are summarized in a conceptual AOP network that helps prioritize (parts of) AOPs for a more detailed evaluation. In conclusion, this review advances the tDOA description of an existing THSD AOP network and serves as a catalog summarizing plausible and empirical evidence on which future cross-species AOP development and tDOA assessment could build.

4. Weight of evidence for cross-species conservation of androgen receptor-based biological activity (Vliet et al., 2023a)

    - Summary: This paper describes how SeqAPASS results can be combined with systematic methods for thorough literature review to understand biological pathway conservation brining together in silico, in vitro, and in vivo knowledge. Among the endocrine targets Endocrine Disruptor Screening Program is ***focused on, is the androgen receptor (AR).*** Many environmental and industrial chemicals have been shown to have androgenic activity and can disrupt the endocrine system by mimicking or antagonizing natural hormones. Using a combination of in-silico structural comparisons and systematic analysis of available literature, this study conducted a comprehensive evaluation of the cross-species comparability of chemical interactions at the AR. The results of this work suggest that structural conservation of chemical-AR interactions is conserved across vertebrate species and that chemicals binding to the AR ligand binding domain should behave similarly in non-mammalian vertebrates.

5. Demonstration of the Sequence Alignment to Predict Across Species Susceptibility Tool for Rapid Assessment of Protein Conservation (Vliet et al., 2023b)

    - Summary: This paper describes the protocol to guide users through submitting jobs, navigating the various levels of protein sequence comparisons, and interpreting and displaying the resulting data. New features of SeqAPASS v2.0-6.0 are highlighted. Furthermore, ***two use-cases focused on transthyretin and opioid receptor protein conservation*** using this tool

are described. Finally, SeqAPASS' strengths and limitations are discussed to define the domain of applicability for the tool and highlight different applications for cross-species extrapolation.

6. Defining the Biologically Plausible Taxonomic Domain of Applicability of an Adverse Outcome Pathway: A Case Study Linking Nicotinic Acetylcholine Receptor Activation to Colony Death (Jensen, Blatz and LaLone, 2022)

   - Summary: The work described here lays out the initial guidance on how to incorporate predictive bioinformatics approaches to define the biologically plausible taxonomic domain of applicability for the AOP framework. Specifically, the AOP framework provides a means to organize existing knowledge and data from the literature to understand causal linkages connecting a MIE to an adverse outcome (AO) at a biological level of organization relevant to risk assessment. Typically, AOPs are developed considering one or a handful of species for which empirical data describing the biological pathway exist. Although developers tend to assume broader species coverage, based on biological plausibility, in AOPs actual species-specific evidence supporting the tDOA remains relatively narrow, limiting confidence in application across species. The tDOA of an AOP is an important consideration in development and use of an AOP in regulatory decision-making, particularly when an understanding of the appropriateness of surrogate species, relative to broader species representation, may be important. To date, the tDOA is defined by the specific species used in the studies describing the KEs; and in some cases, text descriptions assume broader taxonomic coverage with limited documented evidence to do so. The US Environmental Protection Agency's (USEPA's) Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool can be used in a hierarchical framework alongside a collection of in vitro and in vivo toxicity test data to determine if structural conservation and functional conservation exist across the AOP in other species. Here, the objective was to demonstrate the utility of this bioinformatics approach in a manner consistent with the available and developing features within the AOP-Wiki (https://aopwiki.org/). Further application of SeqAPASS using a case example with an AOP extracted from an existing AOP network highlights the utility of these approaches and the challenges. ***The AOP network describing nAChR activation leading to colony death/failure (focused on Apis mellifera) was selected with specific attention on one of the described AOPs (AOP 89) as a case example for demonstration of how SeqAPASS could aid in defining the tDOA.***

7. International Consortium to Advance Cross‐Species Extrapolation of the Effects of Chemicals in Regulatory Toxicology (LaLone et al., 2021)

   - Summary: The primary impact of this article is that it lays the foundation for the International Consortium to Advance Cross Species Extrapolation in Regulation (ICACSER; https://www.setac.org/page/scixspecies). This is a consortium that I co-founded and have served as the Chair since 2020. The ICACSER was founded to address an increasing need to make greater use of existing knowledge and mechanistic information relative to chemical safety, particularly to protect the diversity of species in the environment. Recognizing the challenges of utilizing empirical toxicity data generated from one or a handful of model organisms to represent potential toxicity in other species and the advances that have been made in gene and protein sequencing technologies and bioinformatics, a number of new approach methods in bioinformatics have been developing for purposes of species extrapolation. Because these bioinformatics approaches are evolving rapidly and becoming more recognized by decision-makers as providing useful lines of evidence for hazard and risk assessment, it was considered ideal timing by the ICACSER co-founders to bring global tool/database developers together to advance the science on cross species extrapolation. In addition, such a shift to the use and integration of data from bioinformatics approaches in the

risk assessment paradigm requires a continuous global discussion with decision makers on typical workflows and criteria for use of new methods in chemical safety evaluations. Therefore, this article is a call to action for those working on the science of species extrapolation for human and environmental health and to begin a dialog centered initially on the use of bioinformatics, with the recognition that additional insights relative to toxicokinetic and toxicodynamic influences on chemical sensitivity across species will provide the most thorough understanding of the potential adverse effects across species. This paper describes the next steps in breaking out of typical tool/database development silos and bringing together the top scientists and decision-makers in this field to make the leap to implementation of bioinformatics in decision-making.

8. Integrative Computational Approaches to Inform Relative Bioaccumulation Potential of Per-and Polyfluoroalkyl Substances Across Species (Cheng et al., 2021)

   - Summary: This is the first publication demonstrating the combined utility of the US Environmental Protection Agency Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool and advanced molecular modeling, molecular docking, and molecular dynamic simulation for computational cross species extrapolation to predict PFAS bioaccumulation potential. This paper lays the groundwork for the development of a pipeline to move from protein sequence to structure to functional predictions of chemical effects. This work is important because in demonstrating what the strengths were in combining these methods, it has led to the first iterations of the development of a command line code to generate protein structural models for any protein using SeqAPASS output in preparation for molecular docking and virtual screening across species.

   - Specifically, the growing necessity of predictive toxicology in chemical safety evaluation has resulted in innovative application and expansion of existing computational methods and the creation of new tools that can address some of the most pressing questions in emerging contaminants research. Of importance is the evaluation of differences in contaminant impacts across species, which can inform both ecosystem protection and the identification of appropriate model species for human toxicity studies. ***Here we evaluated tools to predict cross-species differences in binding affinity between per- and polyfluoroalkyl substances (PFAS) and the liver fatty acid binding protein (LFABP) which is considered a key determinant of bioaccumulation potential for these ubiquitous environmental contaminants.*** We focused on two complementary tools: the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool and molecular dynamics (MD). Using human LFABP as the query sequence in SeqAPASS it was determined that the protein was conserved in the majority of vertebrate species, therefore indicating these species would have similar bioaccumulation potential to humans. Level 3 SeqAPASS evaluation was also used to identify potentially destabilizing amino acid differences across species, which were generally supported by DUET, which is a web-based server that predicts the effects of mutations on protein stability. Key differences predicted by SeqAPASS and DUET were selected for further evaluation by molecular dynamics in two ways: by investigating nine single-residue mutations and seven whole-sequence species differences. For the single-point mutations, predicted binding affinities were compared for PFOA and PFNA, and a comparison to a recently published in vitro study was also included as validation. For single residue differences flagged as potentially critical by SeqAPASS, one single point mutation (F50V for PFNA) showed a statistically significant difference with higher affinity than wild-type human LFABP. Molecular dynamic simulations were then expanded to evaluate binding affinities for 9 different PFAS across 7 species (human, rat, chicken, rainbow trout, zebrafish, Japanese medaka and fathead minnow). Human, rat, chicken, zebrafish, and rainbow trout had similar binding affinities to one another for each PFAS, whereas Japanese medaka and fathead minnow had

significantly weaker LFABP binding affinity for some PFAS. In all cases and for all species a strong negative correlation was found between LFABP binding affinity and PFAS chain length. Based on these analyses, the combined use of SeqAPASS and molecular dynamics provides for rapid screening for potential species differences with deeper structural insight. This approach can be easily extended to investigate other important biological receptors and PFAS as potential ligands.

9. Evidence for Cross Species Extrapolation of Mammalian-Based High-Throughput Screening Assay Results (LaLone et al., 2018)

   - Summary: The publication was the initial broad application of the US Environmental Protection Agency Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool for extrapolation of high-throughput screening data across taxa with a focused evaluation of endocrine targets.

     Cell-based high throughput screening (HTS) technologies are becoming mainstream in chemical safety evaluations. The US Environmental Protection Agency (EPA) Toxicity Forecaster (ToxCastTM) and the multi-agency Tox21 Programs have been at the forefront in advancing this science, making screening results publicly accessible for both research and regulatory decision-making. Initially these programs were developed with the intent of screening chemicals for hazard potential as a means to prioritize further in vivo toxicity testing for the protection of human health. Therefore, primarily mammalian based HTS assays were developed. Subsequently, it was recognized that ToxCast screening data may be more broadly applicable for the protection of wildlife. Therefore, a challenge emerged for understanding whether these predominantly mammalian-based prioritization approaches reasonably reflect potential impacts on other species. To address this species-extrapolation challenge, the US EPA Sequence Alignment to Predict Across Species Susceptibility tool (SeqAPASS; https://seqapass.epa.gov/seqapass/) was employed. SeqAPASS evaluations comparing primary amino acid sequences and functional domains were performed for all 450 molecular targets associated with the ToxCast assays and data are being made publicly available through both the SeqAPASS tool and the Comptox Chemistry Dashboard (https://comptox.epa.gov/dashboard). ***To demonstrate the practical application of the SeqAPASS approach to a current regulatory challenge, case studies relevant to the US Endocrine Disruptor Screening Program were developed evaluating androgen receptor, enzymes involved in sex steroid synthesis, and proteins found in the thyroid axis.*** Overall, conservation of these mammalian protein targets suggests that screening results identifying chemical disruption via HTS assays could be reasonably extrapolated across vertebrate species.

   - This application of the SeqAPASS tool led to its recognition among international partners and the Organisation for Economic Co-operation and Development (OECD) that published the Revised Guidance Document 150 on Standardized Test Guidelines for Evaluating Chemicals for Endocrine Disruption in 2018 which recommended the SeqAPASS tool for evaluations of protein conservation for cross species extrapolation.

10. In Silico Site-Directed Mutagenesis Informs Species-Specific Predictions of Chemical Susceptibility (Doering et al., 2018)

    - Summary: This study established the mthods and default settings for comparing critical amino acids across species for the integration of the Level 3 analysis in SeqAPASS. The evaluation used in silico site-directed mutagenesis coupled with docking simulations of computational models for ***acetylcholinesterase (AChE) and ecdysone receptor (EcR) to investigate how specific amino acid substitutions impact protein-chemical interaction***. This study found that computationally derived substitutions in identities of key amino acids caused no change

in chemical interaction with a protein if residues share the same side chain functional properties and have comparable molecular dimensions, while differences in these characteristics can reduce protein-chemical interaction. These findings were considered in the development of automated Level 3 susceptibility predictions, which were incorporated in SeqAPASS v.3.0, and enable automatically generated species-specific predictions of chemical susceptibility. These predictions were shown to agree with Level 1 and 2 predictions of AChE and EcR for more than 90 % of investigated species, but also identified dramatic species-specific differences in chemical susceptibility that align with results from standard toxicity tests. The results provide a compelling line-of-evidence for use of SeqAPASS v.3.0 in deriving screening level, species-specific, susceptibility predictions across broad taxonomic groups.

11. Evaluation of the scientific underpinnings for identifying estrogenic chemicals in nonmammalian taxa (Ankley et al., 2016)

    - Summary: The publication lays out a hierarchical framework for collecting evidence of biological pathway conservation from available comparative analyses ranging from in silico approaches (primarily SeqAPASS), to in vitro assays (i.e., cell-based) to in vivo (i.e., whole organism) test results which can then be used to define the taxonomic domain of applicability for many applications, including data derived from mammalian-based high-throughput screening assays. In the context of the US Environmental Protection Agency Endocrine Disruptor Screening Program (EDSP), mammalian-based in vitro high-throughput screening assays (e.g., ToxCast assays) are being used to identify chemicals likely to act as endocrine disruptors, which then move on to more costly toxicity testing protocols. ***The focus of this paper is on demonstrating how the hierarchical framework can be applied to comprehensively analyze cross-species conservation using a case study to evaluate chemical-estrogen receptor-α (ERα) interactions.*** There is a suite of high-throughput screening assays designed to detect chemicals that act through mammalian ERα, but the objective of screening is to identify chemicals that could act as endocrine disruptors in both mammalian and non-mammalian species. So, a major uncertainty is whether the mammalian-based prioritization approach reasonably reflects potential interactions with non-mammalian species. SeqAPASS as a key component for defining how taxonomic relevance can be considered for other pathways and targets relevant to the EDSP program.

12. SeqAPASS: A Web-Based Tool for Addressing the Challenges of Cross-Species Extrapolation (LaLone et al., 2016)

    - Summary: An underutilized data source for purposes of species extrapolation is protein sequence and structural information (e.g., the National Center for Biotechnology Information protein database currently contains 81,027,309 protein sequences representing 68,165 organisms). With sequencing and annotation techniques becoming more cost-effective and streamlined, this growing source of data served as the motivation for developing the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool, that utilizes protein sequence/structural data to predict chemical susceptibility across species based on concepts derived from evolutionary biology. The publication describes the public release of the web-based tool, Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS), developed to rapidly predict chemical susceptibility across thousands of species (e.g., vertebrates, invertebrates, plants, viruses) simultaneously through automated comparisons of protein sequence/structural data. In the manuscript, ***presentation of SeqAPASS methodology is accompanied by demonstration, through case studies, of how the tool and data can be applied toward current challenges in species extrapolation, including predicting pollinator susceptibility to neonicotinoid and molt-accelerating insecticides.*** The SeqAPASS tool was developed for both the scientific and regulatory communities, by

scientists in the Office of Research and Development with input from US EPA regulators (e.g., Office of Pesticide Programs, Office of Science Coordination and Policy). This collaboration led to flexibility in the SeqAPASS analysis for comparison of chemical molecular targets across species at varying levels of complexity, depending on available data and knowledge about a chemical protein interaction and the degree of protein characterization. A key element of the SeqAPASS tool is that species comparison metrics based on protein sequence are translated to a final susceptibility prediction, with each species being assigned as "susceptible" or "not susceptible" to a given chemical, providing a risk assessor with information they can readily interpret for purposes of species extrapolation of chemical effects. The SeqAPASS tool has been applied to key projects in the Program Offices, including understanding potential chemical susceptibility to endangered or threatened species, identification of the most likely susceptible species for revisions to the 1985 Aquatic Life Criteria Guidelines, and for initial considerations in extrapolating high-throughput screening data from mammalian-based assays to other vertebrates in the Endocrine Disruptor Screening Program.

13. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk (LaLone et al., 2013)

    - Summary: A combination of primary amino acid sequence alignments and detailed analyses of conserved functional domains were used to provide quantitative assessments of sequence similarity. Computational approaches for calculating and collating these quantitative metrics of protein similarity were automated to make such analyses efficient enough to be performed routinely. ***Case examples focused on the actions of (a) 17α-ethinyl estradiol on the human (Homo sapiens) estrogen receptor; (b) permethrin on the mosquito (Aedes aegypti) voltage-gated para-like sodium channel; and (c) 17β-trenbolone on the bovine (Bos taurus) androgen receptor are presented to demonstrate the potential predictive utility of this species extrapolation strategy.*** Using three case studies, susceptibility predictions based on the approach were then compared with available empirical toxicity data as a means to evaluate underlying assumptions and define appropriate domains of applicability. The computational studies described represent several critical, initial steps toward the development and routine application of quantitative molecular target similarity-based approaches to predictive species extrapolation.

PROPRIETARY ASPECTS

The SeqAPASS tool is publicly accessible upon request of a personal login. All requests are granted for SeqAPASS Level 1-3 which were developed for all to use. For Level 4 access, the requester is asked to provide proof of their expertise in protein structural biology prior to being granted access. Due to how computationally expensive the Level 4 evaluation is, it is intended for experts only with specific questions relative to species extrapolation in line with the intended use of SeqAPASS. There are no proprietary aspects to the SeqAPASS tool.

PROPOSED REGULATORY USE

The intended regulatory use for the SeqAPASS tool is to provide a transparent scientific method for cross species extrapolation of toxicity knowledge and data. The overall aim is to utilize protein conservation as a line of evidence to support understanding the taxonomic domain of applicability of AOP knowledge and for predicting chemical susceptibility for untested species. Results from the SeqAPASS tool have use as a screening method in the context the US Endocrine Disruptor Screening Program, to be considered as OSRI in weight of evidence evaluations (https://www.regulations.gov/document/EPA-HQ-OPP-2021-

0756-0002). This example specifically points to the use of the data as additional lines of evidence in any weight of evidence evaluation. The use of SeqAPASS was also discussed by Cegar et al., 2022, outlining current ecotoxicity testing needs for the US Federal Agencies. (Ceger et al., 2022, Current ecotoxicity testing needs among selected US federal agencies)

REFERENCES

Scientific Publications:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ankley GT, LaLone CA, Gray LE, Villeneuve DL, Hornung MW. Evaluation of the scientific underpinnings for identifying estrogenic chemicals in nonmammalian taxa using mammalian test systems. Environmental Toxicology and Chemistry. 2016 Nov;35(11):2806-16.

Ceger P, Vinas NG, Allen D, Arnold E, Bloom R, Brennan JC, Clarke C, Eisenreich K, Fay K, Hamm J, Henry PF. Current ecotoxicity testing needs among selected US federal agencies. Regulatory Toxicology and Pharmacology. 2022 Aug 1;133:105195.

Doering JA, Lee S, Kristiansen K, Evenseth L, Barron MG, Sylte I, LaLone CA. In silico site-directed mutagenesis informs species-specific predictions of chemical susceptibility derived from the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool. Toxicological Sciences. 2018 Nov 1;166(1):131-45.

Federhen, S. (2012). The NCBI taxonomy database. Nucleic Acids Research, 40(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178

Haigis AC, Vergauwen L, LaLone CA, Villeneuve DL, O'Brien JM, Knapen D. Cross-species applicability of an adverse outcome pathway network for thyroid hormone system disruption. Toxicological Sciences. 2023 Sep 1;195(1):1-27.

Jensen MA, Blatz DJ, LaLone CA. Defining the biologically plausible taxonomic domain of applicability of an adverse outcome pathway: A case study linking nicotinic acetylcholine receptor activation to colony death. Environmental Toxicology and Chemistry. 2023 Jan;42(1):71-87.

LaLone CA, Basu N, Browne P, Edwards SW, Embry M, Sewell F, Hodges G. International consortium to advance cross‐species extrapolation of the effects of chemicals in regulatory toxicology. Environmental toxicology and chemistry. 2021 Dec;40(12):3226.

LaLone CA, Blatz DJ, Jensen MA, Vliet SM, Mayasich S, Mattingly KZ, Transue TR, Melendez W, Wilkinson A, Simmons CW, Ng C. From protein sequence to structure: The next frontier in cross‐species extrapolation for chemical safety evaluations. Environmental Toxicology and Chemistry. 2023 Feb;42(2):463-74.

LaLone CA, Villeneuve DL, Burgoon LD, Russom CL, Helgen HW, Berninger JP, Tietge JE, Severson MN, Cavallin JE, Ankley GT. Molecular target sequence similarity as a basis for species extrapolation to assess the ecological risk of chemicals with known modes of action. Aquatic toxicology. 2013 Nov 15;144:141-54. (LaLone et al., 2013)

LaLone CA, Villeneuve DL, Doering JA, Blackwell BR, Transue TR, Simmons CW, Swintek J, Degitz SJ, Williams AJ, Ankley GT. Evidence for cross species extrapolation of mammalian-based high-throughput screening assay results. Environmental science & technology. 2018 Oct 10;52(23):13960-71.

LaLone CA, Villeneuve DL, Lyons D, Helgen HW, Robinson SL, Swintek JA, Saari TW, Ankley GT. Editor's highlight: sequence alignment to predict across species susceptibility (SeqAPASS): a web-based tool for addressing the challenges of cross-species extrapolation of chemical toxicity. Toxicological Sciences. 2016 Oct 1;153(2):228-45.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., ... & Bryant, S. H. (2012). CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Research, 41(D1), D348–D352. https://doi.org/10.1093/nar/gks1243

Olker, J. H., Elonen, C. M., Pilli, A., Anderson, A., Kinziger, B., Erickson, S., ... & Hoff, D. (2022). The ECOTOXicology knowledgebase: A curated database of ecologically relevant toxicity tests to support environmental research and risk assessment. Environmental Toxicology and Chemistry, 41(6), 1520–1539. https://doi.org/10.1002/etc.5324

Papadopoulos, J. S., & Agarwala, R. (2007). COBALT: constraint-based alignment tool for multiple protein

sequences. Bioinformatics, 23(9), 1073–1079. https://doi.org/10.1093/bioinformatics/btm076
Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 35(suppl_1), D61–D65. https://doi.org/10.1093/nar/gkl842
Rose, P. W., Prlić, A., Altunkaya, A., Bi, C., Bradley, A. R., Christie, C. H., ... & Green, R. K. (2016). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Research, gkw1000. https://doi.org/10.1093/nar/gkw1000
OECD. (2018). *Guidance Document on Standardised Test Guidelines for Evaluating Chemicals for Endocrine Disruption (2nd edition)*. OECD Publishing. https://www.oecd.org/en/publications/guidance-document-on-standardised-test-guidelines-for-evaluating-chemicals-for-endocrine-disruption-2nd-edition_9789264304741-en.html
Schumann PG, Chang DT, Mayasich SA, Vliet SM, Brown TN, LaLone CA. Cross-species molecular docking method to support predictions of species susceptibility to chemical effects. Computational Toxicology. 2024 Jun 1;30:100319.
Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Žídek, A. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061
Vliet SM, Hazemi M, Blatz D, Jensen M, Mayasich S, Transue TR, Simmons C, Wilkinson A, LaLone CA. Demonstration of the sequence alignment to predict across species susceptibility tool for rapid assessment of protein conservation. JoVE (Journal of Visualized Experiments). 2023 Feb 10(192):e63970. (Vliet et al., 2023b)
Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., ... & Richard, A. M. (2017). The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. Journal of Cheminformatics, 9, 1–27. https://doi.org/10.1186/s13321-017-0247-6
Yang, J., & Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Research, 43(W1), W174–W181. https://doi.org/10.1093/nar/gkv342
Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, 33(7), 2302–2309. https://doi.org/10.1093/nar/gki524

Online Tools and Databases

European Bioinformatics Institute. AlphaFold.
Retrieved from https://AlphaFold.ebi.ac.uk/ [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). BLAST® Help. [Accessed: October 2025]
Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK62051/ [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). BLASTp Tool. Retrieved from
http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
[Accessed: October 2025]
National Center for Biotechnology Information (NCBI). COBALT Tool. [Accessed: October 2025]
Retrieved from http://www.st-va.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). Conserved Domains Database.
Retrieved from http://www.ncbi.nlm.nih.gov/cdd/ [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). NCBI Glossary Field Guide.
Retrieved from http://www.ncbi.nlm.nih.gov/Class/FieldGuide/glossary.html [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). NCBI Protein Database.
Retrieved from http://www.ncbi.nlm.nih.gov/protein/ [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). Taxonomy Database.
Retrieved from http://www.ncbi.nlm.nih.gov/taxonomy/ [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). The NCBI Handbook.
Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21106/ [Accessed: October 2025]
Research Collaboratory for Structural Bioinformatics (RCSB). Protein Data Bank (PDB).
Retrieved from https://www.rcsb.org/ [Accessed: October 2025]
Seacunilever. Genes-to-Pathways Species Conservation Analysis (G2P-SCAN) [GitHub repository].
Retrieved from https://github.com/seacunilever/G2P-SCAN [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Draft White Paper: Availability of New Approach Methodologies (NAMs) in the Endocrine Disruptor Screening Program (EDSP).

Retrieved from https://www.epa.gov/sciencematters/epa-research-contributes-using-alternatives-screen-chemicals-endocrine-disruption [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Federal Register Document. Retrieved from https://www.regulations.gov/document/EPA-HQ-OPP-2021-0756-0002 [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). SeqAPASS Tool. Retrieved from https://seqapass.epa.gov/seqapass/ [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). SeqAPASS User Guide. Retrieved from https://www.epa.gov/comptox-tools/seqapass-user-guide [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). SeqAPASS Resource Hub. Retrieved from https://www.epa.gov/comptox-tools/sequence-alignment-predict-across-species-susceptibility-seqapass-resource-hub [Accessed: October 2025]
US Fish and Wildlife Service. (2018). ECOS Environmental Conservation Online System. Website https://ecos.fws.gov/ecp/ [Accessed: 26 June 2023].
Zhang Lab. Iterative Threading ASSEmbly Refinement (I-TASSER). Retrieved from https://zhanggroup.org/I-TASSER/ [Accessed: October 2025]
Zhang Lab. TM-align. Retrieved from https://zhanggroup.org/TM-align/ [Accessed: October 2025]

Annex C. REPORTING IATA OUTCOMES

Published Case Study: (Schumann et al., 2024)

NAME OF CHEMICAL

Case Study 1:

Chemical:

2-ethylhexanoic acid: Used in organic and industrial chemical synthesis to prepare lipophilic metal derivatives that are soluble in nonpolar organics solvents. For full chemical details see:

https://comptox.epa.gov/dashboard/chemical/details/DTXSID9025293

Gene/Protein target: Peroxisome proliferator activated receptor alpha (PPARα)

Evidence for target selection: 2-ethylhexanoic acid was shown to trans-activate PPARα (EC2x = 500 µM). This is further supported by the ToxCast active hit below Cytotox lower bound. (See Schumann et al., 2024, Supplementary material 4)

Case Study 2:

Chemical:

Diethylstilbestrol: a nonsteroidal estrogen medication, which is presently rarely used. In the past, it was widely used for a variety of indications, including pregnancy support for those with a history of recurrent miscarriage, hormone therapy for menopausal symptoms and estrogen deficiency, treatment of prostate cancer and breast cancer, and other uses. For full chemical details see:

https://comptox.epa.gov/dashboard/chemical/details/DTXSID3020465

Gene/Protein target: Estrogen receptor 1 (ESR1)

Evidence for target selection: protein-ligand crystallization. (See Schumann et al., 2024, Supplementary material 4)

Chemical:

Butylparaben: antimicrobial preservative in cosmetics. It is also used in medication suspensions, and as a flavoring additive in food. For full chemical details see:

https://comptox.epa.gov/dashboard/chemical/details/DTXSID3020209

Gene/Protein target: Estrogen receptor 1 (ESR1)

Evidence for target selection: protein-ligand crystallization. (See Schumann et al., 2024, Supplementary material 4)

Chemical:

Oxybenzone: used in sunscreen formulations, plastics, toys, furniture finishes, and other products to limit UV degradation. For full chemical details see:

https://comptox.epa.gov/dashboard/chemical/details/DTXSID3022405

Gene/Protein target: Estrogen receptor 1 (ESR1)

Evidence for target selection: Oxybenzone was shown to competitively bind to ESR1 against E2 *in vitro* at relatively high concentrations (IC50 = 70 µM). This is further supported by the ToxCast active hit call below the Cytotox lower bound. (See Schumann et al. 2024, Supplementary material 4)

Chemical:

Dibutyl phthalate: used as a plasticizer. For full chemical details see:

https://comptox.epa.gov/dashboard/chemical/details/DTXSID2021781

Gene/Protein target: Estrogen receptor 1 (ESR1)

Evidence for target selection: DBP demonstrated binding affinity to ESR1. This is further supported by the ToxCast assay results, despite having an AC50 greater than the Cytotox lower bound. (See Schumann et al., 2024, Supplementary material 4)

Case Study 3:

Chemical:

Topiramate: a medication used to treat epilepsy and prevent migraines. It has also been used in alcohol dependence and essential tremor. For full chemical details see:

https://comptox.epa.gov/dashboard/chemical/details/DTXSID8023688

Gene/Protein target: Gamma-aminobutyric acid type A receptor subunit alpha (GABRA1)

Evidence for target selection: Topiramate is known to enhance GABA-mediated chloride flux, although the exact mechanism remains unclear. Ligand-binding studies show that topiramate does not block GABA or benzodiazepine binding sites and therefore mediates its effects through an uncharacterized mechanism. (See Schumann et al., 2024, Supplementary material 4)

PURPOSE

The purpose of the IATA "was to demonstrate the value of combining two computational NAMs, the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool (LaLone et al., 2016) and the Genes to Pathways – Species Conservation Analysis (G2P-SCAN) tool (Rivetti et al., 2023), which make use of available omics data across species to support predictions of chemical susceptibility. This is achieved through the contribution of multiple related lines of evidence associated with chemical effects on biological pathways and taxonomic relevance. The potential improvements arising by the combination of these computational NAMs in predicting chemical susceptibility across species could ultimately help inform chemical safety assessment." (From Schumann et al., 2024)

"An assumption of common species extrapolation approaches used in toxicology is that taxonomic relatedness confers similar susceptibility to chemicals. This assumption underlies the use of information on surrogate species to predict potential chemical hazards for other species (Perkins et al., 2013). A key approach to ecological toxicity testing relies on information from representative species as surrogates to represent all other species, typically within a related ecological taxonomic group, namely producers, primary and apical consumers (Colbourne et al., 2022; Spurgeon et al., 2020). It would not be possible, permissible, or desirable to perform toxicity tests on each species that may be exposed to an environmental contaminant. Therefore, NAMs must be thoroughly evaluated in accordance with essential elements of the scientific confidence framework (Van Der Zalm et al., 2022) and define the domain of applicability to

support prediction and, ultimately, safety evaluations for a wide range of chemicals." (From Schumann et al., 2024)

DESCRIPTION

The detailed description of the information sources is available in the Methods section of Schumann et al., 2024 Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species. In addition, the US EPA Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool and Unilever's Genes to Pathways – Species Conservation Analysis (G2P-SCAN) tool have been described in ANNEX I as individual information sources that can also be used separately to provide insights for species extrapolation. These tools are the primary information sources for the IATA described here. It is noteworthy that there are a number of publications demonstrating application of the SeqAPASS tool in the context of species extrapolation and integration of results for defining the biologically plausible taxonomic domain of applicability of adverse outcome pathways (Jensen et al., 2023; Haigis et al., 2023; ANNEX I Information Source SeqAPASS).

DATA REVIEW

All Data sources presented below are described in detail in Schumann et al., 2024.

Chemical target identification: Chemicals were identified by leveraging: 1) US Environmental Protection Agency high-throughput *in vitro* data, 2) ToxCast bioactivity data, 3) structural data available through the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), and 4) existing chemical activity data available in the literature through manual searches. The EPA high throughput *in vitro* data was accessed through version 2 of the RefChemDB, and transcriptional points of departure (tPODs) for a subset of chemicals of interest were extracted. ToxCast data was obtained by searching the CompTox Chemicals Dashboard (available online at https://comptox.epa.gov/dashboard/) using DSSTox substance ID (DTSXID) chemical identifiers. Additionally, the RCSB PDB (https://www.rcsb.org/) was used to search for protein-ligand crystallization data by using a combination of protein target gene symbols and relevant chemical identifiers (SMILES or InChI String). Lastly, a literature search was performed manually by using Google Scholar and Boolean strings containing keywords (e.g., gene names, chemical names, "mechanism of action", "binding", "activation", "inhibition", "molecular docking", "site-directed mutagenesis", etc.) to find experimental data on the specific chemical-target interactions. All databases used in this study were accessed in September 2022.

Genes to Pathways – Species Conservation Analysis (G2P-SCAN) tool: See ANNEX I.

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool: See ANNEX I

Pathway prioritization: Besides G2P-SCAN and SeqAPASS, the latest version (v11.5) of a web-based tool called STRING (https://string-db.org/) was used to construct protein-protein interaction (PPI) networks. Networks were then visualized using Cytoscape software (v3.9.1) where only reciprocal edges (i.e., edges that mutually linked two nodes: A-B, B-A) were used. Molecular Complex Detection (MCODE), available as an automated Cytoscape plugin, was used to perform a connectivity-based cluster analysis on the PPI networks.

Linking pathway data from G2P-SCAN with Adverse Outcome Pathway data: Much of this molecular information is made readily available through the AOP-Database (AOP-DB) (https://aopdb.epa.gov/). A batch search was performed within the AOP-DB using the ID numbers of all the available AOPs at the time of this analysis (September 2022) as an input and specifying genes as the output.

DATA INTEGRATION AND INTERPRETATION

One way in which G2P-SCAN helps infer pathway conservation across 7 species is by the identification of protein orthologs. Orthologous proteins identified by G2P-SCAN across 7 species were compared with the susceptibility predictions obtained from SeqAPASS, which provides output for hundreds of species. Any consensus in the selection of an ortholog and a susceptibility call provides additional support for the conservation of those proteins across species. When considering the consensus across multiple proteins within a mapped biological pathway, this inference of conservation can be extrapolated to the pathway itself.

Case Study 1: 2-ethylhexanoic acid - Peroxisome proliferator-activated receptor alpha (PPARα). See Schumann et al., 2024 section 3.3.1

Case Study 2: Diethylstilbestrol, butylparaben, oxybenzone, and dibutyl phthalate - Estrogen receptor 1 (ESR1 or ERα). See Schumann et al., 2024 section 3.3.2

Case Study 3: topiramate - Gamma-aminobutyric acid type A receptor subunit alpha1 (GABRA1). See Schumann et al., 2024 section 3.3.3

EXPERT JUDGEMENT

Selections of analysis that require Expert Judgement include:

Target identification. See Schumann et al., 2024 section 2.1 – "All potential targets were assigned to chemicals if there was **strong literature evidence to support a direct interaction** with the compound of interest. In this way, high-throughput screening data from either ToxCast or RefChemDB was used as additional support for selecting targets but was not used to assign targets in the absence of literature evidence. **Direct interactions were considered as having evidence of binding or alteration of protein activity** (e.g., catabolic, anabolic, transport, macromolecule binding, etc.) in a concentration-dependent manner. **Interactions that were solely based on impacts on mRNA or protein expression were considered as indirect, and therefore in these cases it was determined that there was not enough evidence to consider the gene or protein a molecular target.**" All references used for supporting target selection are found in the supplementary material (Supplementary material 4) of Schumann et al., 2024.

Evaluating molecular targets using G2P-SCAN. See Schumann et al., 2024 section 2.2 – "The analysis parameters used for each target evaluation were as follows: **use of all 7 model species** (*Homo sapiens, Rattus norvegicus, Mus musculus, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans,* and *Saccharomyces cerevisiae*), **analysis over Least Divergent Orthologs** (LDOs; which represent nearly 'equivalent' gene pairs between different organisms based on a phylogenetic tree analysis) and **limiting the analysis to terminal pathways** (defined as those which have parent pathways, but no children in Reactome's pathway hierarchy). **To determine the likelihood of pathway conservation across species** using the information gathered through G2P-SCAN **several factors were considered**. First, whether **the molecular target had orthologs identified for any of the 6 query species**." "Second, whether **the identified protein families for the molecular target are identical to the human protein families across any of the other 6 species**. Lastly, **by determining how significant the differences in overall count values of the pathway elements** (proteins, protein families, entities, and reactions) are across each species. This was done by performing a one-way analysis of variance (ANOVA) followed by Tukey's honestly significant difference (HSD) post-hoc test using the combined count data (i.e., the count value for each pathway element was considered as a sample within each species group) from each pathway element with respect to each species (using humans as the reference group). Using this test procedure, pathways were considered significantly different at p-values < 0.10."

Pathway prioritization. See Schumann et al., 2024 section 2.3 – "It was important to identify mapped pathways where the molecular target is most likely to play an essential role in the overall pathway function. Two methods were relied upon for prioritizing mapped Reactome pathways from G2P-SCAN for further analysis. The first method was to simply select pathways comprised of a low number of genes with the assumption that a chemical perturbation of the protein target is likely to be more consequential to the function of a pathway with fewer genes. It was decided that pathways with ≤10 genes would be considered priority where the molecular target would represent at least 1 out of 10 (or 10 %) of the pathway coverage. For mapped pathways that were comprised of >10 genes, a second approach was used that relied on the use of protein-protein interaction (PPI) networks to better determine the biological significance of the molecular target within the pathway. This was done using an analysis method external to the G2P-SCAN tool. A critical feature of biological networks is that essentiality correlates positively with centrality. Genes or proteins found at the core of interaction networks with high interconnectivity have been shown to be significant in determining phenotypic states." "The latest version (v11.5) of a web-based tool called STRING (https://string-db.org/) was used to construct PPI networks from the gene lists of all mapped Reactome pathways that were obtained via G2P-SCAN and comprised of >10 genes. The network edges represented the type of interaction evidence, and the minimum required interaction score was set to 0.4 on a scale of approximate confidence from zero to one with one representing approximately 100 % confidence of a true association. These networks were then visualized using Cytoscape software (v3.9.1) where only reciprocal edges (i.e., edges that mutually linked two nodes: A-B, B-A) were used." "MCODE calculates the degree of interconnectivity between each node." "The highest scoring clusters from each network were used to infer the essentiality of the relevant molecular target. If a molecular target was found in the highest scoring cluster, then the proteins within the cluster were considered for SeqAPASS evaluation." "To balance analysis efficiency with prediction confidence, if a molecular complex contained >10 proteins, then only the top ten scoring proteins (i.e., ten most interconnected nodes within the PPI network) that were also directly connected to the target protein were used in SeqAPASS evaluations. This was a pragmatic decision made based on the lines of evidence indicating these proteins as being the most essential to the pathway."

Isoform selection for protein target evaluation in SeqAPASS. See Schumann et al., 2024 Section 2.4 – "If a **target had multiple known isoforms**, they were **prioritized by considering whether the isoform is** 1) **a known reference sequence**, 2) is **the longest sequence of the available isoforms within the reference species** (which was humans, in this case), 3) is **the most recently modified sequence**, and 4) if it is the **first version of that sequence** (i.e., if it is isoform "a" or isoform 1)." "**If the selected isoform yielded a Level 1 susceptibility cutoff at 100 % similarity**, this would suggest that the cross-species protein alignments were not ideal, and **a different isoform should be evaluated**. **Other factors could supersede** this weighted approach if, for example, there is **known evidence of a chemical interaction with a specific protein isoform**."

SeqAPASS Level 1, 2, and 3 evaluations. See Schumann et al., 2024 Section 2.4 – "The **susceptibility cut-off was adjusted where the default ortholog used for driving the percent similarity cutoff was a partial sequence of or not similar to the query accession in terms of protein annotation**." " Level 2 evaluations were performed by first identifying conserved functional domains." "**Conserved domains were selected for evaluation that were supported by evidence of direct interaction with the chemical stressor of interest or** by showing **evidence of interactions with similar compounds**." "To perform a Level 3 evaluation, information on the chemical protein interaction with respect to individual amino acid residues was needed. This information was obtained by relying on published studies that utilized techniques such as site-directed mutagenesis, molecular docking, crystallography, (Q)SAR, or gave evidence of mutational resistance." "In cases **where there was no clear experimental evidence of key amino acids mediating a chemical-protein interaction, residues were selected for Level 3 evaluations where there were chemical interactions such as hydrogen bonding and stacking shown within the PDB crystal structure**."

Pathway-level species susceptibility. See Schumann et al., 2024 Section 2.6 – "Mapped Reactome pathways were **prioritized for further evaluation if the molecular target was a member of a pathway consisting of 10 or fewer proteins or** if the **target was identified within the highest scoring molecular complex** (based on network connectivity) of the pathway's PPI network. The full list of genes that comprised the prioritized pathways with 10 or fewer proteins were used for SeqAPASS evaluations or in the case of prioritized pathways with **>10 proteins, only the top ten highest scoring proteins** (i.e., highest connectivity scores) **found within the highest scoring molecular complexes were used.**" "Once each protein within the prioritized pathway or within **the highest scoring molecular complex** (that contained the molecular target) **were evaluated by SeqAPASS**, the **common susceptible species across each of these protein lists were merged**." This merged list of species represents those in which the pathway is most likely to be conserved.

Linking pathway data from G2P-SCAN with AOP data. See Schumann et al., 2024 Section 2.7 – "Regarding functional comparisons, **the biological outcomes of a MIE, KE, or KER was compared against those of the mapped Reactome pathways**. **If a mapped pathway resulted in the activation of a specific protein, for instance, and this protein activation was also used to describe a particular KE within an AOP, then this was considered as an overlap**. In this way, relatedness in either gene identity or biological function provided a means of deriving a toxicological context for the mapped Reactome pathway information."

Table C.1. Expert Judgment Summary

Section of Workflow	Judgement
Target identification	• Strength of literature evidence to support chemical-protein interaction. • Is there direct evidence of binding or alteration of protein activity? • Impacts on mRNA or protein expression are indirect evidence therefore cannot be used for target identification. • Based on the problem formulation and the known strengths and weaknesses of G2P-SCAN and SeqAPASS (See ANNEX I), is it appropriate to combine the approaches for species extrapolation or use one of the approaches individually.
Evaluating molecular targets using G2P-SCAN	• Decision to use all 7 model species in the analysis. • Analysis is conducted over Least Divergent Orthologs. • Determine likelihood of pathway conservation. • Are orthologs identified for any of the 6 query species? • Are the protein families identical to the human protein families for any of the other 6 species? • How significant are the differences in overall count values of the pathway elements across the 6 species?
Pathway prioritization	• Identify mapped pathways where the molecular target is most likely to play an essential role in the overall pathway function. • Select pathways comprised of a low number of genes with the assumption that a chemical perturbation of the protein target is likely to be more consequential to the function of a pathway with fewer genes. • Pathways with ≤10 genes would be considered priority. • Genes or proteins found at the core of interaction networks with high interconnectivity are deemed significant in determining phenotypic states. • STRING was used to construct Protein-Protein Interaction (PPI) networks from the gene lists of all mapped Reactome pathways that were obtained via G2P-SCAN and comprised of >10 genes. • The highest scoring clusters from each network were used to infer the essentiality of the relevant molecular target. If a molecular target was found in

	the highest scoring cluster, then the proteins within the cluster were considered for SeqAPASS evaluation • If a molecular complex contained >10 proteins, then only the top ten scoring proteins (i.e., ten most interconnected nodes within the PPI network) that were also directly connected to the target protein were used in SeqAPASS evaluations.
Isoform selection for protein target evaluation in SeqAPASS	• Which protein accession will be used to represent the query species/protein in SeqAPASS? • Prioritized which isoform of the protein to query in SeqAPASS. • Is the isoform a known reference sequence? • Is the isoform the longest sequence of the available isoforms with in the reference species? • Is the isoform the most recently modified sequence? • If the selected isoform yielded a Level 1 susceptibility cutoff at 100 % similarity a different isoform should be evaluated. • Is there existing evidence of a chemical interaction with a specific protein isoform?
SeqAPASS Level 1, 2, and 3 evaluations	• Does the susceptibility cut-off need to be adjusted because the default ortholog used for deriving the percent similarity cutoff was a partial sequence of or not similar to the query accession in terms of protein annotation. • Are conserved domains for Level 2 query supported by evidence of direct interaction with the chemical stressor of interest or by showing evidence of interactions with similar compounds? • If there is no clear experimental evidence of key amino acids mediating a chemical-protein interaction, residues were selected for Level 3 evaluations where there were chemical interactions such as hydrogen bonding and stacking shown within the PDB crystal structure.
Pathway-level species susceptibility	• Mapped Reactome pathways were prioritized for further evaluation if the molecular target was a member of a pathway consisting of 10 or fewer proteins or if the target was identified within the highest scoring molecular complex of the pathway's PPI network. • In prioritized pathways with >10 proteins, only the top ten highest scoring proteins (i.e., highest connectivity scores) found within the highest scoring molecular complexes were used for SeqAPASS analysis. • The prioritized pathway or within the highest scoring molecular complex (that contained the molecular target) were evaluated by SeqAPASS and the common susceptible species across each of these protein lists were merged
Linking pathway data from G2P-SCAN with AOP data from AOP-DB	• Compare the biological outcomes of a molecular initiating event (MIE), key event (KE), or key event relationship (KER) against those of the mapped Reactome pathways. • If a mapped pathway resulted in the activation of a specific protein, for instance, and this protein activation was also used to describe a particular KE within an AOP, then this was considered as an overlap.

UNCERTAINTIES IN WORKFLOW

PPARA. See Schumann et al., 2024 Section 3.3.1 – "By using the isoform selection approach described NP_005027.2 was used as the query accession for the SeqAPASS Level 1evaluation representing the human PPARα. As there was (at the time of this study) no evidence of 2-ethylhexanoic acid binding directly to PPARα, **it was assumed that the interaction would likely occur at the ligand binding site of the protein since previous studies had demonstrated an activating effect on PPARα similar to known ligands**. Therefore, the conserved domain cd06932 was used for the SeqAPASS Level 2 evaluation. No

information on critical amino acid specific information was available to support a Level 3 evaluation. Using PPARA as an input for G2P-SCAN yielded 10 Reactome pathways. **None of these mapped pathways had a percent coverage by PPARA of over 10 %, so separate PPI networks were constructed using STRING for each of these pathways and an MCODE analysis was performed on each for the connectivity-based cluster analysis.** The pathways in which PPARα was found within the highest scoring molecular complex were selected for further SeqAPASS evaluations. In some cases, the molecular complexes containing PPARα were too large (>10 proteins) to efficiently evaluate each protein of the complex within SeqAPASS. Therefore, **the top ten scoring proteins contained within the complex that were also directly connected to PPARα were used in Level 1 evaluations**. This approach to filtering key pathway proteins was intended to balance analysis efficiency with prediction confidence. **The resulting "susceptible" species lists obtained from each of the Level 1 evaluations using the high scoring molecular complex proteins and the Level 2 evaluation of PPARα were merged to identify the identical "susceptible" species across each list."** See Schumann et al., 2024 Section 3.4 – "At present, SeqAPASS and G2P-SCAN rely on gathering evidence regarding the presence or absence of proteins for similarity across species at the target and pathway level, respectively, as a predictor of chemical susceptibility. **By combining approaches that consider additional factors such as life stage, life history, biological sex and toxicokinetic factors like absorption, distribution, metabolism, excretion (ADME) – these predictions will yield a more complete understanding of the chemical exposure and resulting biological impacts**. Using biological pathway information and assessing protein interaction networks offers another promising way to enhance the predictive capabilities of computational approaches towards a systems-based view.

Currently, the limitations of this approach make it **most suited for hazard identification** within the context of environmental safety evaluation. To broaden the application of computational NAMs there **will need to be an inclusion of exposure data to adequately link a chemical protein interaction to an adverse outcome, improved ontological annotation of many AOP KEs and KERs, reliable *in silico* predictions of chemical-protein interaction information, and empirical evidence linking the modulatory effects of chemicals on target activity to downstream pathway elements.**"

Table C.2. Uncertainties in Workflow Summary

Section of Workflow	Uncertainty
PPARα Case Study	• It was assumed that the interaction of 2-ethylhexanoic acid likely occurs at the ligand binding site of the protein since previous studies had demonstrated an activating effect on PPARα similar to other known ligands. Therefore, the conserved domain representing the Ligand binding domain, cd06932, was used for the SeqAPASS Level 2 evaluation. • None of the 10 Reactome pathways mapped by G2P-SCAN had a percent coverage by PPARα of over 10 %, so separate PPI networks were constructed using STRING for each of these pathways and a MCODE analysis was performed on each for the connectivity-based cluster analysis. The pathways in which PPARα was found within the highest scoring molecular complex were selected for further SeqAPASS evaluations. • Some molecular complexes containing PPARα contained >10 proteins. Therefore, the top ten scoring proteins contained within the complex that were also directly connected to PPARα were used in Level 1 SeqAPASS evaluations.

- Consideration of additional factors such as life stage, life history, biological sex and toxicokinetic factors like absorption, distribution, metabolism, excretion (ADME) outside of these approaches would undoubtably enhance these predictions and yield a more complete understanding of the chemical exposure and resulting biological impacts.
- To adequately link a chemical-protein interaction to an adverse outcome, improved ontological annotation of many AOP KEs and KERs and empirical evidence linking the modulatory effects of chemicals on target activity to downstream pathway elements are needed.

UNCERTAINTIES SPECIFIC TO THE CHEMICAL ASSESSED

2-Ethylhexanoic acid. See Schumann et al., 2024 Section 3.3.3 – "2-Ethylhexanoic acid, which is an industrial compound used to make paints and plasticizers, has been shown to target PPARα in previous studies. For example, 2-Ethylhexanoic acid was shown to activate PPARα with an EC2X (two-fold effect concentration) of 500 µM, and this interaction with PPARα was shown to be more selective compared to other." "PPARs. Although there is **currently no clear evidence of 2-ethylhexanoic acid directly binding to PPARα, these studies demonstrated a dose-dependent activation**. Therefore, PPARα was considered as a molecular target of 2-ethylhexanoic acid for this study."

Topiramate. See Schumann et al., 2024 Section 3.3.3 – "The weight of evidence to support GABRA1 as a direct target of topiramate is less substantial. Rather, **studies have shown that topiramate is able to modulate GABA-evoked currents in neurons via $GABA_A$ receptors and this activity is dependent on the specific subunit combinations. Thus, the exact mechanism underlying this activity is not fully understood.**"

The evidence supporting PPARα as a molecular target of 2-Ethylhexanoic acid is stronger than the evidence for GABRA1 as a target for topiramate. For 2-Ethylhexanoic acid, there is an empirically derived dose-dependent effect, a clear directional response, and a well-characterized mechanism of action through PPARα. In contrast, topiramate's activity on $GABA_A$ receptors lacks these elements. Therefore, the species susceptibility predictions based on evaluations of PPARα-mapped pathways are better supported by scientific evidence than those derived from GABRA1-mapped pathways.

Table C.3. Uncertainties Specific to the Chemical Assessed Summary

Section of Workflow	Uncertainty
2-Ethylhexanoic acid	• Currently no clear evidence of 2-ethylhexanoic acid directly binding to PPARα, However it was demonstrated to lead to a dose-dependent activation of PPARα. Therefore, PPARα was considered as a molecular target of 2-ethylhexanoic acid.
Topiramate	• Studies have shown that topiramate is able to modulate GABA-evoked currents in neurons via $GABA_A$ receptors and this activity is dependent on the specific subunit combinations. Thus, the exact mechanism underlying this activity is not fully understood.

OUTCOME OF THE ASSESSMENT

Results from the combined bioinformatics approaches (See ANNEXIIData README.DOC and ANNEXIIData.XLSx for details)

Case Study 1:

2-ethylhexanoic acid: In examining overlapping species among PPARα and the top 10 scoring proteins connected to PPARα, **SeqAPASS predicted 187 mammalian species as likely susceptible to 2-ethylhexanoic acid**. Of these 187 species, *Rattus norvegicus* and *Mus musculus* (out of the 6 G2P-SCAN query species) were found. Furthermore, the analysis of the pathway protein, protein family, reaction and entity counts supported the conservation of this pathway within rats and mice as the count values for these species did not differ significantly (p-values > 0.10) from the human counts. In this way, **both SeqAPASS and G2P-SCAN results provided separate lines of evidence to support the inference of this pathway as being conserved within these two species and through principles of species read-across the results supported the hypothesis that this pathway is likely to be conserved within the mammalian taxon**.

See Schumann et al., 2024 Section 3.3.1 “Ten AOPs that involved PPARα were identified from the AOP-DB. One of these AOPs was “NR1I3 (CAR) suppression leading to hepatic steatosis” (AOP 58; https://aopwiki.org/aops/58), which had the most gene-level information out of any other existing AOP (i.e., 33 genes in total). This high amount of gene-level information allowed for comparison of the genes contained within this AOP to those identified via G2P-SCAN from mapped Reactome pathways. Of the 33 genes that were associated with AOP 58, 11 were unique, human genes. Ten of these 11 human genes were found within at least one PPARα mapped Reactome pathway derived from G2P-SCAN. The fact that most (10 out of 11) of the human genes known to be associated with AOP 58 were found within PPARα -mapped Reactome pathways indicated that these pathways may be useful in broadening the biological information used to describe the MIE, KEs, or KERs of this AOP. Overlaps in the function of the Reactome pathways and the AOP KEs offers support that the interaction between 2-ethylhexanoic acid and PPARα may lead to an adverse outcome. However, AOP 58 defines PPARα inhibition as a MIE, whereas studies showed that 2-ethylhexanoic acid activated PPARα (Lampen et al., 2003; Maloney and Waxman, 1999). Although the role of PPARα in the development of liver steatosis differs in AOP 61 “NFE2L2/FXR activation leading to hepatic steatosis,” which lists PPARα activation as an early KE. Therefore, it is still possible that 2-ethylhexanoic acid could lead to liver steatosis by activating PPARα, but this apparent discrepancy in the modulation of PPARα highlights a challenge in making predictions of apical outcomes from chemical perturbation through this computational approach alone and underscores the importance of considering chemical-protein interactions in the context of biological pathways.”

Case Study 2:

Diethylstilbestrol, Butylparaben, Oxybenzone, Dibutyl phthalate: The top 10 scoring proteins that were directly connected to ESR1 in the molecular complexes of these PPI networks produced by STRING were used to query SeqAPASS for evaluation of pathway conservation. As examples, the SeqAPASS evaluations on the key pathway proteins for “Estrogen-dependent gene expression” (R-HSA-9018519) **resulted in 139 “susceptible” mammalian species**, and the evaluations for “RUNX1 regulates transcription of genes involved in WNT signaling” (R-HSA- 8939256) resulted in 272 “susceptible” species across 8 taxa (Actinopteri, Amphibia, Aves, Chondrichthyes, Crocodylia, Lepidosauria, Mammalia, and Testudinata). The G2P-SCAN approach identified ESR1 orthologs for *Rattus norvegicus*, *Mus musculus*, and *Danio rerio*. Additionally, the protein, protein family, reaction, and entity counts for each of the 6 priority pathways within these three species are not significantly different from the counts in humans (p-values > 0.10), and the protein functional families across these species were identical to the human protein family for ESR1. When considered in combination with the SeqAPASS results, multiple lines of evidence are

derived to support the inference of pathway conservation for these specific species and for the hypothesis that the mapped pathways are conserved across species within identical taxonomic groups.

See Schumann et al., 2024 Section 3.3.2 “When searching for AOPs that involved ESR1, 9 AOPs were identified. However, the total amount of gene-level information contained within these AOPs was limited. The AOP with the greatest number of genes associated with it (AOP 67; https://aopwiki.org/aops/67) had 9 genes in total (which accounted for 6 vertebrate species associated with the AOP). Only 2 of these 9 genes were unique human genes – ESR1 and NR2F2. Moreover, NR2F2 was not found to be associated with any of the ESR1-mapped Reactome pathways. This lack of mapping reduced confidence in assigning a toxicological context to these pathways by simply assessing the amount of gene-level overlap there is between relevant AOPs. Instead, there was an increased reliance on the functional aspects of the pathways and the KEs of the relevant AOPs to make such comparisons.”

See Schumann et al., 2024 Section 3.3.2 “Diethylstilbestrol is considered a prototypical stressor of AOP 167 (“Early-life estrogen receptor activity leading to endometrial carcinoma in the mouse”) and is known to activate ESR1 (Shiau et al., 1998). ESR1 activation is the first KE of AOP 167 (KE ID: 1065). The second KE is “Promotion, SIX-1 positive basal-type progenitor cells.” The KER between these two events is considered “non-adjacent” and the steps by which the activation of ESR1 leads to the promotion of SIX-1 positive cells is not well described. However, one of the prioritized ESR1-mapped Reactome pathways, “RUNX1 regulates transcription of genes involved in WNT signaling” (R-HSA-8939256), contains detailed molecular information on how ESR1 activity may impact the promotion of SIX1 via the activation of the pathway “degradation of beta-catenin by the destruction complex” (R-HSA-195253) (Fig. 7), which is also indirectly related to certain forms of cancer (Kimelman and Xu, 2006) and therefore is in line with AOP 167 (development of endometrial carcinoma in the mouse).”

See Schumann et al., 2024 Section 3.3.2 “While AOP 167 described ESR1 activation leading to SIX1 upregulation in mice, another study found that ESR1 activation via diethylstilbestrol led to SIX1 downregulation in mice (Terakawa et al., 2020). The effect of ESR1 activity on SIX1 expression could be dependent on factors like developmental stage (Suen et al., 2016). Despite these discrepancies in the directional impact on SIX1 expression, both sources agree that ESR1 activation via diethylstilbestrol indeed alters SIX1 expression (in some way) and ultimately leads to the promotion of endometrial carcinoma (Suen et al., 2016; Terakawa et al., 2020). Therefore, the opportunity for using pathway information derived from these bioinformatics approaches to fill knowledge gaps in existing AOPs remains. This added molecular information derived from the combined computational approaches described here could then be used to enhance cross-species extrapolation of chemical susceptibility.”

Case Study 3:

Topiramate: The SeqAPASS evaluation of the key pathway proteins resulted in a list of **199 “susceptible” mammalian species** that included *Rattus norvegicus* or *Mus musculus* (according to SeqAPASS full reports). Therefore, both tools gave support for the hypothesis that this pathway is likely to be conserved across mammalian species.

When searching for GABRA1 within the AOP-DB, only one AOP was found – “Binding to the picrotoxin site of ionotropic GABA receptors leading to epileptic seizures in adult brain” (AOP 10). This AOP had 12 genes total associated with it. Three of these 12 genes were unique human genes: GABRA1, GABRA5, and GABRA6. When comparing overlapping genes in the mapped Reactome pathways derived from G2P-SCAN, all three of these genes were found in at least one of the two mapped pathways, but only GABRA1 was found in the prioritized pathway. Like the analysis described using ESR1, this lack of overlap in molecular information led to a greater dependency on functional overlaps to understand potential toxicity.

Linking pathway Data to AOPs:

See Schumann et al., 2024 Section 2.7 “The AOPs and the mapped G2P-SCAN identified Reactome pathways of these three targets, ESR1, PPARα, and GABRA1, were compared to identify overlapping biology. These comparisons were performed through a qualitative assessment of the molecular information and functional outcomes of the AOP MIEs, KEs, and KERs and the mapped Reactome pathways. More specifically, the number of genes that overlapped between the mapped Reactome pathways and the AOP was determined. Regarding functional comparisons, the biological outcomes of an MIE, KE, or KER was compared against those of the mapped Reactome pathways. If a mapped pathway resulted in the activation of a specific protein, for instance, and this protein activation was also used to describe a particular KE within an AOP, then this was considered as an overlap. In this way, relatedness in either gene identity or biological function provided a means of deriving a toxicological context for the mapped Reactome pathway information. Therefore, in cases of AOP and Reactome pathway relatedness, the susceptible species lists obtained via SeqAPASS evaluations performed on those pathways could be used as a line of evidence for expanding the biologically plausible tDOA of the related AOP (as described by Jensen et al., 2023).”

In the instance that overlap does not exist between the genes in the mapped Reactome pathways and the AOP, extrapolation of the protein target in the MIE (ESR1, PPARα, and GABRA1) can be completed using SeqAPASS results.

Overall, this work establishes an initial framework for using biological pathway information to enhance chemical susceptibility predictions across species and demonstrates the synergy that can be obtained through the combined use of existing methods. SeqAPASS uses protein sequence information to help make predictions of chemical susceptibility, while G2P-SCAN accesses information from various databases to help infer biological pathway conservation across select species. In combination, it was demonstrated that these tools can be used to expand the prediction of biological pathway conservation across all species with relevant protein data, aid in the prediction of cross-species susceptibility, extend the biologically plausible tDOA of relevant AOPs, and provide additional biological information to help better characterize KEs and KERs.

DISCUSSION OF OTHER INTERPRETATIONS

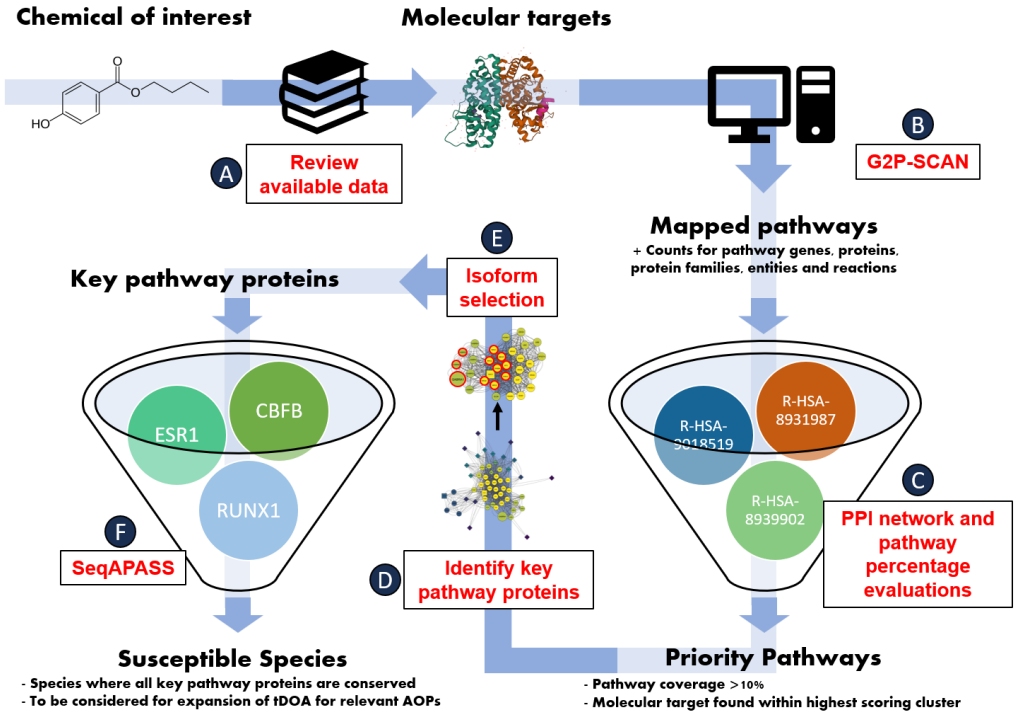
The ability to make predictions of biological pathway conservation across species using combined approaches is limited to the amount of **publicly available biological data**. The lines of evidence presented in this IATA for species extrapolation are intended to build upon one another where consensus in conservation enhances the reliability of the predictions. **Ultimately, the weight of this evidence must be judged by the end user**. To extend these predictions of pathway conservation to chemical induced adversity, this depends on either **known sensitive species** or is judged by the **degree of overlap in the function of biological pathways and key events of relevant adverse outcome pathways**.

GUIDE TO THE COMBINED USE OF NAMs: SeqAPASS AND G2P-SCAN

Overview

The purpose of this protocol is to provide a practical guide for researchers and regulators to utilize the Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool and Genes to Pathways – Species Conservation Analysis (G2P-SCAN) tool in combination for supporting chemical risk assessment. These tools require that the user has knowledge of a protein target responsible for or related to the biological effects of the chemical of interest. Additionally, each of these tools has extensive existing documentation on their use and application. Therefore, this guide will assume that the user is familiar with the basic theory and operations of the individual tools. The application of this approach was demonstrated using case examples as described in the published case study (Schumann et al., 2024).

Figure C.1. Diagram summarizing the steps in the approach described in this guide for combining the use of Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) and Genes to Pathways – Species Conservation Analysis (G2P-SCAN) tools.



Acronyms in this figure: taxonomic domain of applicability (tDOA), adverse outcome pathways (AOPs), molecular initiating events (MIEs), key events (KEs), key event relationships (KERs), and protein-protein interaction (PPI).

1. Required Software/Tools

1.1 Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)

SeqAPASS is a web-based tool that utilizes protein sequence information to predict species susceptibility to chemical effects. Briefly, by knowing the protein target of a chemical in one species, SeqAPASS compares that protein sequence to all other known protein sequences on a per species basis to determine in which species that protein is likely conserved. By understanding protein conservation, SeqAPASS can

make predictions about cross-species susceptibility to chemical effects. In this way, any quality protein sequence can be used from any relevant species to extrapolate toxicity knowledge to all other species.

SeqAPASS can be accessed at https://seqapass.epa.gov/seqapass/. All new users to SeqAPASS must register for an account following the instructions at https://www.epa.gov/comptox-tools/guide-login-seqapass-users.

1.2 Genes to Pathways – Species Conservation Analysis (G2P-SCAN)

G2P-SCAN leverages multiple existing databases through various application programming interfaces (APIs) to both map genes to known biological pathways and predict pathway conservation across several common model organism species. The tool only accepts a human gene symbol as input. By querying the various databases, orthologs are identified for each protein in the mapped pathways for each of the six model organisms that are part of the analysis i.e. (rat (*Rattus norvegicus*), mouse *(Mus musculus*), zebrafish (*Danio rerio*), fruit fly (*Drosophila melanogaster*), nematode (*Caenorhabditis elegans*), and budding yeast (*Saccharomyces cerevisiae*), to help assess the overall pathway conservation in those species.

G2P-SCAN is implemented as an R package. Therefore, if the user does not already have R and RStudio installed, they can do so at https://posit.co/download/rstudio-desktop/.

Within R, the G2P-SCAN package can be installed with the following commands:

```
install.packages("devtools")
library(devtools)
devtools::install_github("seacunilever/G2P-SCAN")
```

and then loaded with:

```
library(Genes2Pathways)
```

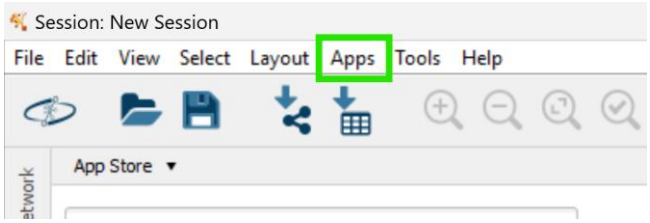
1.3 Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)

The STRING database is a comprehensive collection of curated biological information that can be integrated and scored to generate protein-protein interaction (PPI) networks. This tool is available in a web format and is accessible at https://string-db.org/.

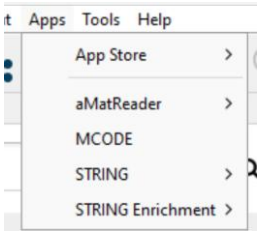
1.4 Cytoscape

Cytoscape is an open-source software used for visualizing and analyzing networks (e.g., PPI networks). It can be downloaded at https://cytoscape.org/.

Within Cytoscape, two apps are needed. To install them, open Cytoscape and go to the “Apps” tab:



Next, search the App Store for “stringApp” and “MCODE” and install both. After installation, these apps should be available within the “Apps” drop-down menu:



2. Target identification

2.1 Primary Literature

Arguably, the strongest evidence for identifying molecular targets is by searching through primary literature. Direct, experimental data that shows a dose-dependent change in activity or function of the protein target in response to exposure to the chemical of interest is ideal. Ultimately, the ability to identify a protein target for any given chemical will be limited by the available data as well as the extent of the expert judgement. Methods for literature review can include scientific search engines such as Google Scholar or Web of Science. Using these common search engines, Boolean strings can be used to help refine the search results. For example, to search for relevant data on the chemical diethylstilbestrol, the following Boolean string could be used as a starting point:

(diethylstilbestrol OR "diethyl stilbestrol" OR DES) AND ("molecular target*" OR "binding site*" OR receptor* OR "gene expression" OR "protein interaction*" OR "signaling pathway*" OR "cellular mechanism*") AND (estrogen OR "endocrine disrupt*" OR "hormone-like")

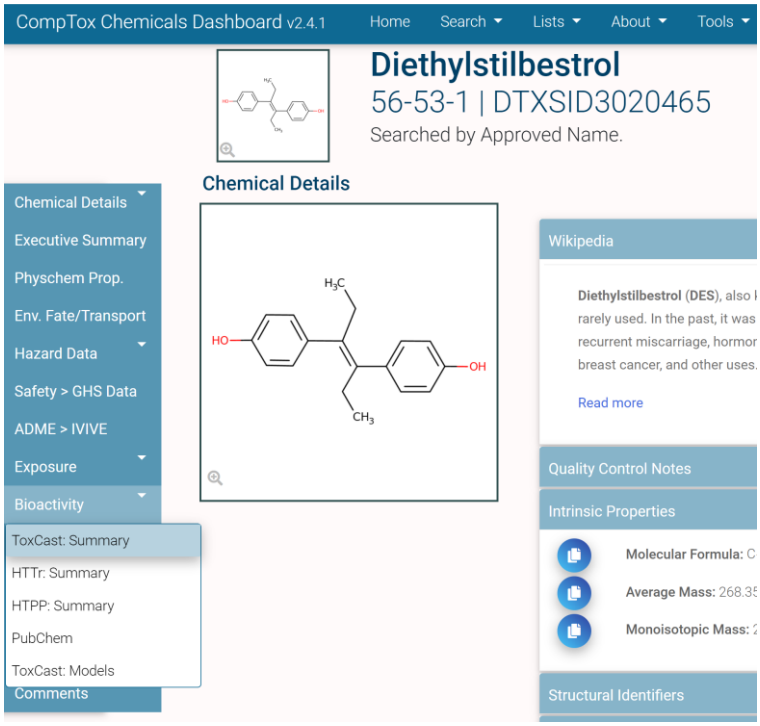
Depending on the chemical of interest specific databases can be used. For instance, if the chemical of interest is a pharmaceutical, the DrugBank database could be used, which provides detailed information on the known mechanisms of action of drugs. If the chemical is a pesticide, then the Toxin and Toxin Target Database (T3DB) could be used.

Efforts are underway to develop and improve upon systematic review methods (Vliet et al., 2023). These methods could help researchers and decision-makers more rapidly generate a collection of relevant literature for extracting molecular target data from existing sources.

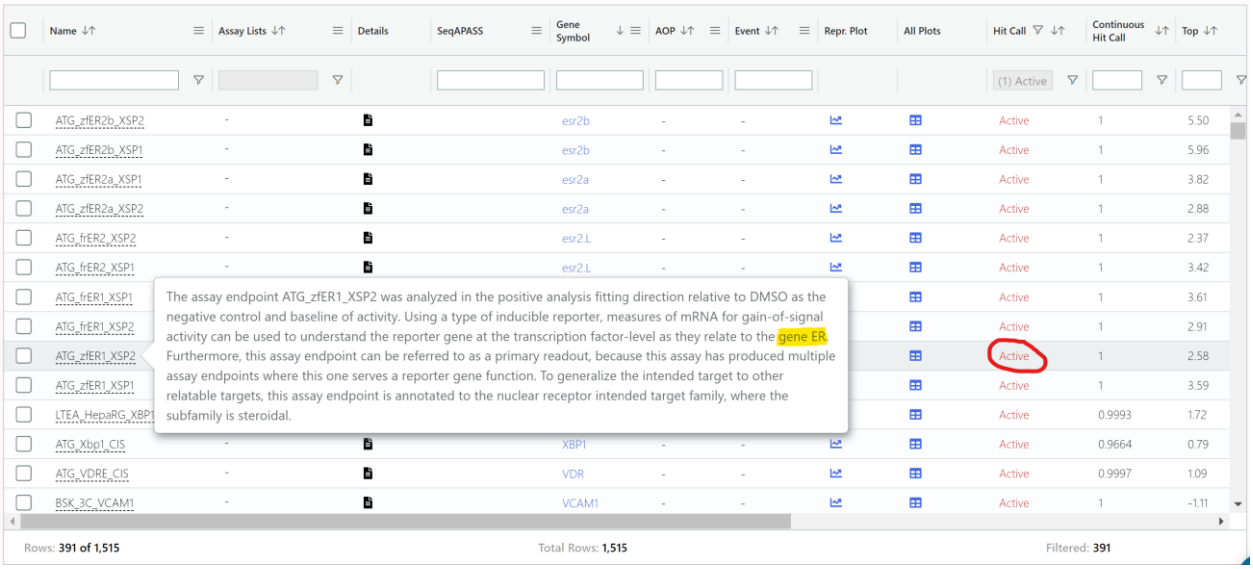
2.2 ToxCast

The United States Environmental Protection Agency's Toxicity Forecasting (ToxCast) program generates and stores biological data from various high-throughput screening assays within a database. This database can be easily queried using a web interface called the CompTox Chemicals Dashboard, which is available at https://comptox.epa.gov/dashboard/.

For example, searching for diethylstilbestrol in the Chemical Dashboard will lead to the page shown in the screenshot below. The ToxCast data can be accessed under the "Bioactivity" tab.



Within the ToxCast results, there is a summary table listing all the assays that were performed using this chemical. The assay highlighted in the following screen shot shows that diethylstilbestrol resulted in an "Active" hit call for multiple assays that measured impacts on mRNA levels for the gene "esr1". Therefore, the results provide evidence that esr1, estrogen receptor isoform1, is likely a target of diethylstilbestrol.



2.3 RCSB PDB

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) is a central repository for all protein structural data. It includes both empirically derived structures through traditional techniques like x-ray crystallography as well as, more recently, computed structural models.

There are instances where researchers will co-crystallize (or cryogenically freeze, in the case of cryo-EM) a chemical ligand in complex with a protein target and upload this data to the PDB. For example, the

structure of the estrogen receptor alpha (ESR1) in complex with diethylstilbestrol was solved (albeit with a mutated structure for “conformation trapping” purposes) as shown in the screenshot below.

This type of data provides strong evidence that the chemical of interest is indeed binding to the protein it was found in complex with. Presumably, this binding interaction has a biological significance. Therefore, RCSB can be used in this manner to identify evidence for chemical-protein interactions.

2.4 Transcriptomics data

More recently, there has been a growing interest in using transcriptomic data to help identify molecular targets. While this data type is less direct than binding assays or crystallography, it can still be useful in gaining some confidence in target identification. However, there is currently no standardized or straightforward method of using this data type to do this. Therefore, the use of this data type for target identification will be left entirely to expert judgement, at present.

3. Use of Genes to Pathways Species Conservation Analysis

3.1 Inputting to G2P-SCAN

To both map the molecular targets to known biological pathways and to generate initial evidence of species conservation across various model organisms, the official human gene symbols of the targets must be input to G2P-SCAN.

Detailed instructions on using the R package can be found at https://github.com/seacunilever/G2P-SCAN.

As an example, within R the following lines of code can be run to query the gene ESR1:

```
library(Genes2Pathways)
library(parallel) # for detectCores()
ACHE_BCHE_T_ALL <- runGenes2Pathways(inputGenes = "ESR1",
                        pathwayLevels = c("terminal"),
                        species = NULL,
                        orthologueFilter = "ALL",
                        orthologueOutput = "both",
                        cores = (detectCores() - 1),
                        g2pScanData = ".",
```

```
                    outputDir = "example_dir",
                    outputPrefix = "example",
                    pathwaysTabs = TRUE,
                    countSummary = TRUE,
                    versions = TRUE,
                    inputSummary = TRUE,
                    reactomeOnMissingUpdate = "update",
                    useSynonyms = FALSE)
```

3.2 Interpreting outputs

A major benefit to the use of G2P-SCAN is its ability to organize a large array of complex biological pathway information into an easily interpretable and usable format. The outputs of this tool are formatted as Excel files. The example below shows the results for the ACHE and BCHE gene queries.

This table (found in the outputted “counts.xlsx” file) shows a list of all the pathways that were mapped by the inputted genes:

	A	B	C	D
1	Pathway ID	Pathway Name	Human_pathway_coverage_percent	Human_Input Found
2	R-HSA-2219530	Constitutive Signaling by Aberrant PI3K in Cancer	1.41	ESR1
3	R-HSA-9018519	Estrogen-dependent gene expression	0.68	ESR1
4	R-HSA-9009391	Extra-nuclear estrogen signaling	1.39	ESR1
5	R-HSA-383280	Nuclear Receptor transcription pathway	2	ESR1
6	R-HSA-1251985	Nuclear signaling by ERBB4	3.23	ESR1

This table (found in the outputted “data.xlsx” file) shows a list of all the genes in each of the mapped pathways:

	A	B	C
1	Gene.Identifier	Gene.Symbol	Pathway.Identifier
2	10718	NRG3	R-HSA-2219530
3	10818	FRS2	R-HSA-2219530
4	118788	PIK3AP1	R-HSA-2219530
5	145957	NRG4	R-HSA-2219530
6	152831	KLB	R-HSA-2219530
7	1839	HBEGF	R-HSA-2219530
8	1950	EGF	R-HSA-2219530
9	1956	EGFR	R-HSA-2219530
10	2064	ERBB2	R-HSA-2219530
11	2066	ERBB4	R-HSA-2219530
12	2069	EREG	R-HSA-2219530
13	2099	ESR1	R-HSA-2219530
14	2100	ESR2	R-HSA-2219530
15	2246	FGF1	R-HSA-2219530
16	2247	FGF2	R-HSA-2219530

To determine which of the 6 model organism species are most likely to have conserved the mapped pathways, the user should consider the results holistically. This evaluation should include examining: reaction, entities, identified orthologs, protein families, pathway sizes, and pathway gene percentages relative to human pathways.

Evaluating pathway conservation across these model organisms will be useful for comparison with SeqAPASS results for species extrapolation.

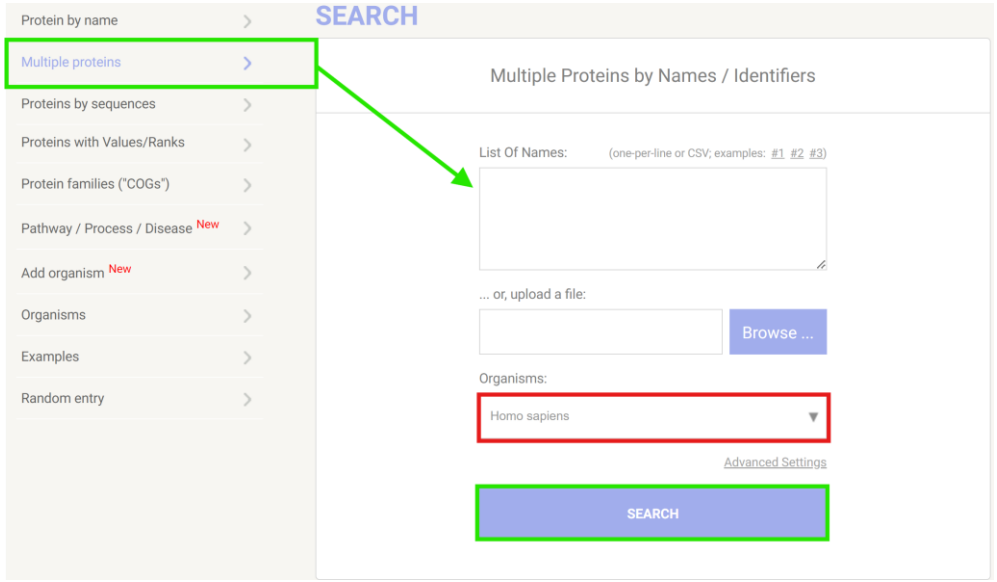
4. Prioritizing pathways

The pathway prioritization process involves evaluation of pathway protein-protein interaction (PPI) networks. This approach is less useful for pathways comprised of a low number of proteins. Therefore, it is recommended to only apply this network approach to pathways comprised of 10 or more proteins.

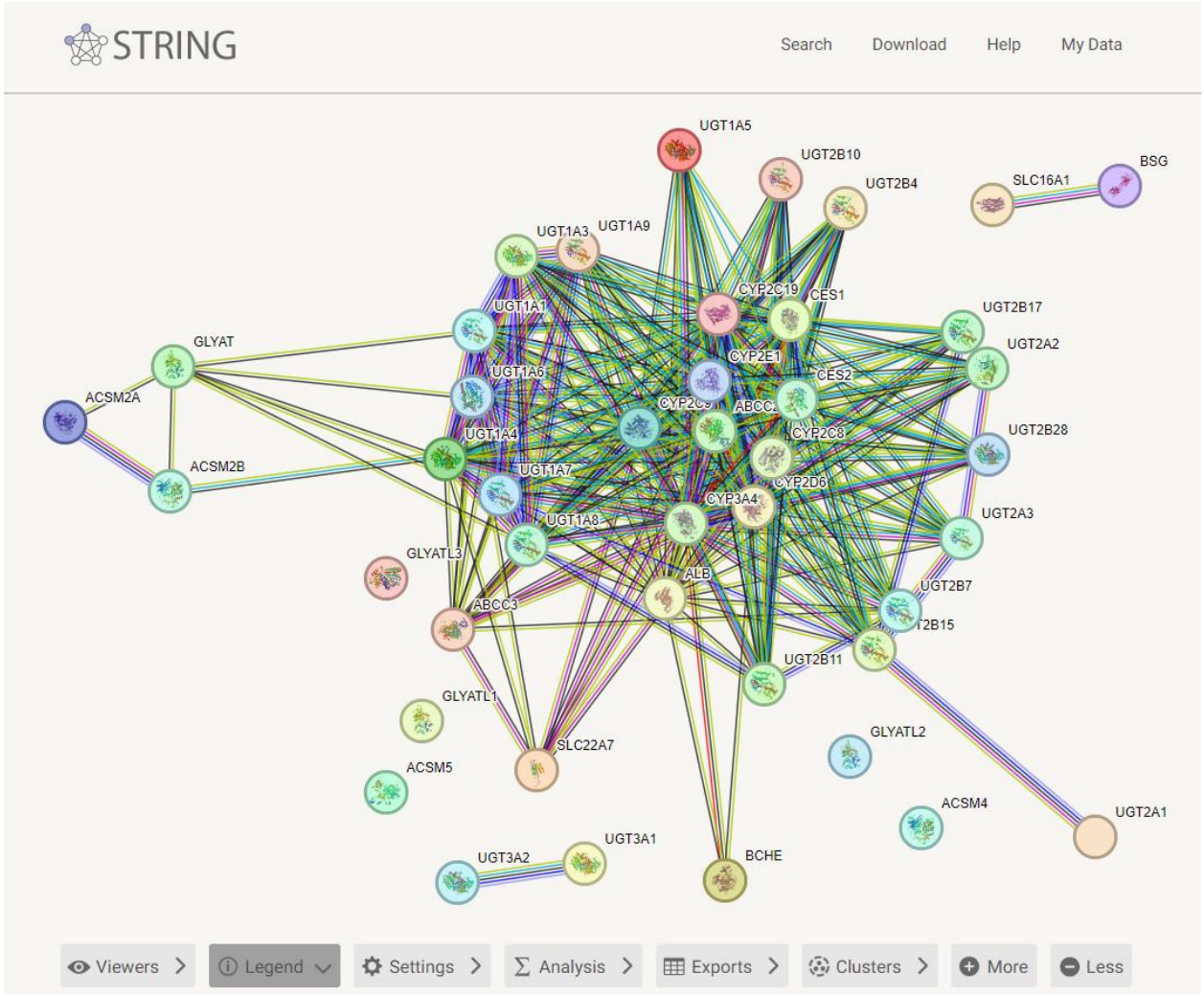
4.1 Generating PPI networks

To generate the PPI networks, the gene lists for each of the mapped pathways is inputted to STRING. Open the "data.xlsx" file from the G2P-SCAN output. Copy the list of gene symbols for one of the pathways. On the STRING website, click the "search" button on the home page:

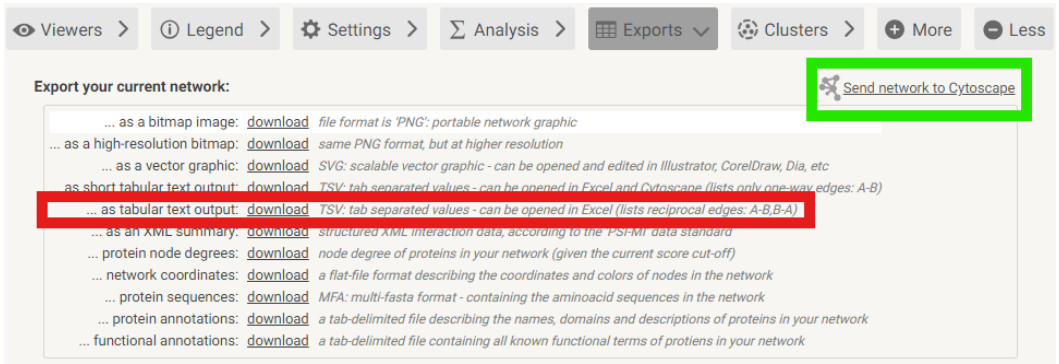
Next, on the STRING search menu, go to "Multiple proteins", paste the list into the "List of Names" box, set the organism as *Homo sapiens*, and then click the "search" button:



After reviewing the protein identifiers, a PPI network is generated using the default parameters:

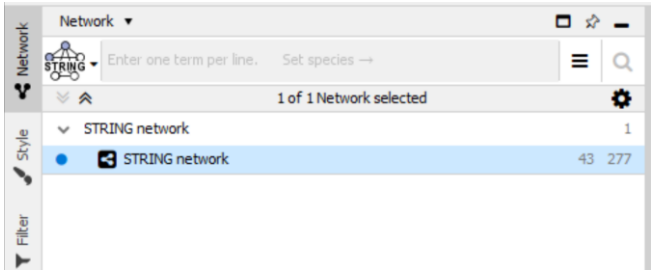


Continuing with the default settings of the PPI network, this can be exported directly to Cytoscape by first opening the Cytoscape software, and then in STRING clicking “Send Network to Cytoscape”. Download the “tabular text output” describing the reciprocal edges of the network (the file will be called “string_interactions.tsv” by default). This file can also be used to load the network into Cytoscape for further analysis:

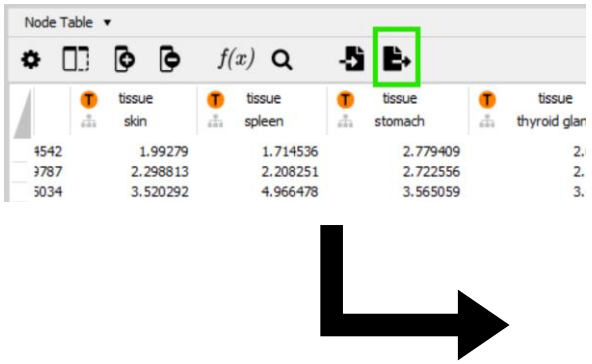


4.2 Assessing degree of interconnectedness

After loading the network into Cytoscape, the degree of protein (i.e., “node”) interconnectedness can be analyzed via MCODE. To do this, click on the loaded network:



Then, go to “Apps” à “MCODE” à “Analyze Current Network”. This will load a new sidebar tab called “MCODE” that lists the detected molecular complexes. By clicking on the complex, the proteins will be highlighted in the network figure. Most importantly, the goal of this is to determine if the molecular target is also found in the molecular complex. This can be done by downloading and searching the MCODE node results table:



	M	N	O
n	display name	MCODE::Clusters (1)	MCODE::N
9	UGT1A6	Cluster 1	Clustered
7	UGT1A8	Cluster 1	Clustered
3	UGT1A7	Cluster 1	Clustered
2	UGT1A3	Cluster 1	Clustered
7	CYP2E1	Cluster 1	Clustered
9	UGT1A4	Cluster 1	Clustered
5	CYP2D6	Cluster 1	Clustered
5	ABCC3	Cluster 1	Clustered
2	CES2	Cluster 1	Clustered
1	CES1	Cluster 1	Clustered
6	CYP2C8	Cluster 1	Clustered
5	ALB	Cluster 1	Clustered
1	CYP3A4	Cluster 1	Clustered
8	CYP2C19	Cluster 1	Clustered
5	ABCC2	Cluster 1	Clustered
9	UGT1A9	Cluster 1	Clustered
5	CYP2C9	Cluster 1	Clustered
3	UGT1A1	Cluster 1	Seed

If the molecular target protein is not found in the highest scoring MCODE cluster, then the pathway represented by this network is considered *not priority* and will not be analyzed further. Otherwise, the network will be used for SeqAPASS evaluation.

5. Identifying key pathway proteins

The purpose of identifying key pathway proteins is mainly to reduce the number of inputs required for SeqAPASS evaluation. The combination of SeqAPASS results for multiple query proteins is currently not an automated process. Therefore, performing evaluations on tens of proteins for many pathways is inefficient in most cases. To best represent the entire pathway with key proteins based on their known interactions with other proteins, it is suggested to prioritize the top ten highest scoring proteins within the molecular complex determined by MCODE analysis.

To determine what these key proteins are, open the exported PPI network node table (as shown at the end of section 4, above). Then, sort the data in descending order by “MCODE:Score”. Use the top ten proteins **that also have a direct connection to the molecular target** in the subsequent SeqAPASS evaluations.

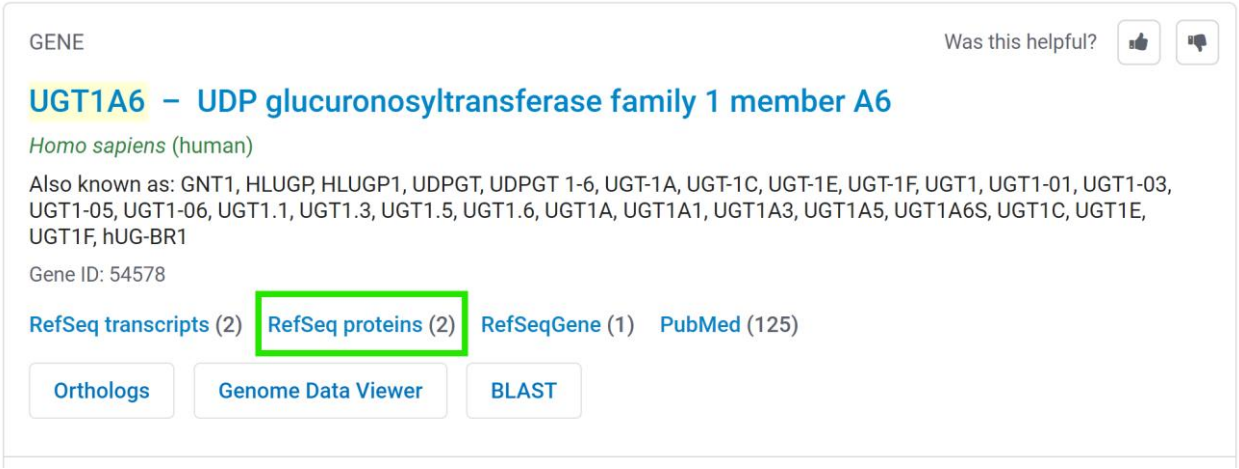
To check if the protein is connected to the molecular target, open the “string_interactions.tsv” file using Excel. Filter the data by either the “#node1” or “node2” columns and refer to the proteins that are connected to the molecular target. For example, if the molecular target is ESR1, then the output might look like the following:

	A	B	C	
1	#node1	node2	node1_s	no
2	AR	NR3C1	9606.ENSP	96
3	AR	PIAS1	9606.ENSP	96
4	AR	PIAS4	9606.ENSP	96
5	AR	HDAC4	9606.ENSP	96
6	AR	PGR	9606.ENSP	96
7	AR	UBE2I	9606.ENSP	96
8	AR	NR3C2	9606.ENSP	96
9	AR	ESR1	9606.ENSP	96
10	AR	PIAS3	9606.ENSP	96
11	AR	PIAS2	9606.ENSP	96
12	AR	SUMO1	9606.ENSP	96
13	ESR1	NR3C1	9606.ENSP	96
14	ESR1	PIAS1	9606.ENSP	96
15	ESR1	RARA	9606.ENSP	96
16	ESR1	HDAC4	9606.ENSP	96
17	ESR1	PGR	9606.ENSP	96
18	ESR1	NR2C1	9606.ENSP	96
19	ESR1	UBE2I	9606.ENSP	96
20	ESR1	NR5A2	9606.ENSP	96
21	ESR1	AR	9606.ENSP	96
22	ESR1	SUMO1	9606.ENSP	96
23	ESR1	PIAS3	9606.ENSP	96
24	ESR1	SUMO2	9606.ENSP	96
25	ESR1	NR1H4	9606.ENSP	96
26	HDAC4	NR1H2	6.ENSP	96

In this case, there are 13 other proteins in this PPI network that have a reciprocal connection to ESR1. If all 13 of these proteins are in the highest scoring molecular complex, it is recommended to use only the top 10 with the highest MCODE scores for SeqAPASS evaluation.

6. Isoform selection

Before performing SeqAPASS evaluations with the key pathway proteins, proper isoforms must be identified for each. Search for the protein in the NCBI database, then go to “RefSeq proteins”:

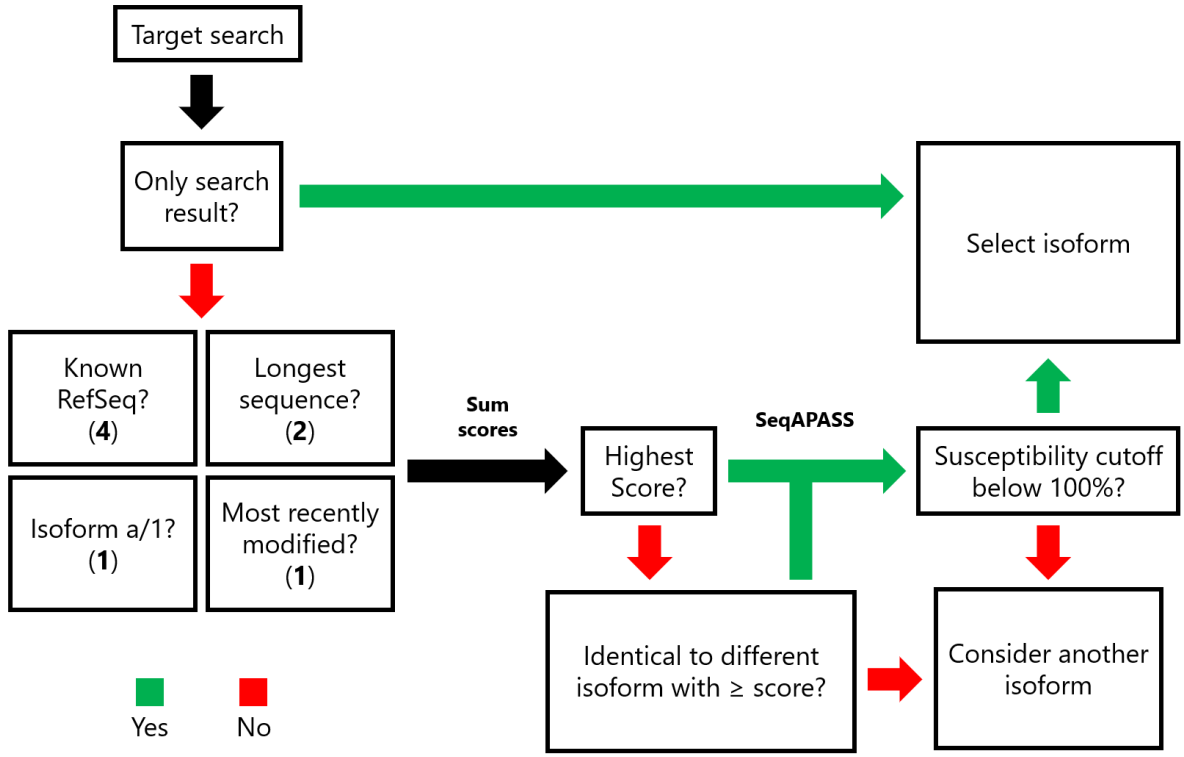


If there is only 1 RefSeq protein, this is the isoform that will be used in SeqAPASS. Otherwise, score each result using the following criteria:

Feature/Score	Description
Known reference sequence? Score = 4	Reference sequences should be empirically derived and ideally would have an "NP_" prefix. The prefix "XP_" means that this is a predicted model and is therefore less reliable for SeqAPASS evaluations.
Longest sequence? Score = 2	The length of the sequences is shown by "### aa protein".
Isoform "a" or "1"? Score = 1	The first isoform described is generally more broadly representative of that protein, which is important to ensure that the selected isoform matches the isoform from the primary literature search.
Most recently modified? Score = 1	Using the most up-to-date sequence helps ensure that the quality of the sequence is higher. This is not always the case.

This scoring scheme is designed to help guide a user in selecting appropriate isoforms. If a specific accession is mentioned in primary literature, the user should just use that accession. Otherwise, the exact scores are meant to be interpreted more qualitatively rather than as a strict categorization.

Once an isoform is selected, it should be submitted as the query to SeqAPASS for Level 1 evaluation. It is important to check that the default susceptibility cutoff presented in the SeqAPASS Level 1 results is not set to 100% since this is not going to yield useful results. If the cutoff is set to 100%, consider evaluating a different isoform, if available. Otherwise, consider selecting a different protein within the molecular complex. The flow chart below summarizes this process:



If the user is still uncertain on which isoform to select as the query protein, the user can query each isoform in SeqAPASS. In the Level 1 results, the user can view the number of ortholog candidates identified for each isoform in the "View SeqAPASS Reports" tab in the 3rd column of the table with header "Ortholog Count" and select the isoform that has the greatest number of ortholog candidates.

7. Use of SeqAPASS

Once appropriate isoforms are identified for each key pathway protein, the SeqAPASS evaluations can proceed, and all resulting data can be interpreted concurrently to generate lists of species that are most likely to conserve the associated pathway.

A Level 1 evaluation must be performed for the molecular target in addition to any key pathway proteins. That is, if 10 key pathway proteins were selected, 11 total queries will need to be submitted to SeqAPASS for each prioritized pathway. The user should refer to the SeqAPASS user guide and relevant publications for a detailed discussion on using the SeqAPASS tool.

7.1 SeqAPASS evaluation of molecular target

The following table is a summary of the SeqAPASS information that was used to perform Level 1 – 3 evaluations, where applicable, for each of the case examples described in Schumann et al., 2024:

Table C.4. SeqAPASS query information for the 3 case examples (adapted from Schumann et al., 2024)

Chemical Name	Target Gene Symbol	Level 1 Query Species (Protein Accession)	Level 2 Domain (Domain Accession)	Level 3 template species (Protein Accession)	Level 3 Amino Acids
Oxybenzone; Butylparaben; Dibutyl phthalate; Diethylstilbestrol	ESR1	*Homo sapiens* (NP_0000116.2)	Ligand binding domain of Estrogen receptor (cd06949)	Homo sapiens (NP_000116.2)	Butylparaben (353E, 394R, 404F); Diethylstilbestrol (343 M, 353E, 394R, 404F, 524H)
2-Ethylhexanoic acid	PPARA	*Homo sapiens* (NP_005027.2)	Ligand binding domain of peroxisome proliferator-activated receptors (cd06932)	NA	NA
Topiramate	GABRA1	*Homo sapiens* (NP_001121116.1)	Neurotransmitter-gated ion-channel ligand binding domain (pfam02931)	NA	NA

The selection of a query species and sequence accession (**Table 1**, column header "Level 1 Query Species (Protein Accession)") is determined by the literature evidence that was used to determine that the protein was a molecular target of the chemicals of interest. If the chemical of interest is known to interact with or bind to a particular functional domain of the protein target (i.e., ligand binding domain), then the user should identify the domain using the NCBI conserved domain database (https://www.ncbi.nlm.nih.gov/cdd/) and select a "specific hit" representing the domain. The accession of this domain can then be used for performing a Level 2 evaluation (**Table 1**, column header "Level 2 Domain (Domain Accession)"). Lastly, if there is information regarding individual amino acid residues that are critical for mediating the interaction of the protein target with the chemical of interest, then these residues should be used to perform a Level 3 evaluation (**Table 1**, column header "Level 3 Amino Acids").

Level 3 information can be obtained from published studies that utilize techniques such as site-directed mutagenesis, x-ray crystallography, (Q)SAR, or provide evidence of mutational resistance (e.g., pesticide resistance). Searching the RCSB Protein Data Bank (PDB) (https://www.rcsb.org/) is a useful first step by searching the protein target name and filtering for the chemical of interest to find structural data on chemical-bound crystal structures. In cases where there was no clear experimental evidence of key amino acids mediating a chemical-protein interaction, residues may be selected where chemical interactions such as hydrogen bonding and pi-stacking occur according to the solved PDB structure.

Data from Level 1, 2, and 3 can be used to predict chemical susceptibility across species with each level of the analysis adding taxonomic resolution to the analysis. Level 1 results can make predictions for broad taxonomic groups (e.g., phylum, class, order) whereas species specific predictions based on chemical interaction are generated from Level 3 results. It is recommended to make use of the data from the level of evaluation with the highest taxonomic resolution when integrating these results with other pathway proteins. The primary results tables for the SeqAPASS evaluations should be saved using the "Download Table" function:

7.2 SeqAPASS evaluation of key pathway proteins

When performing SeqAPASS evaluations on key pathway proteins, for simplification and efficiency, it is recommended to use the Level 1 results for these proteins without the consideration of conserved domains

(Level 2) and critical amino acid residue information (Level 3). Level 2 and Level 3 analyses could be applied if knowledge is available to guide selection of appropriate domains and amino acids, however, collecting this knowledge and combining the data for each key protein of every mapped pathway would be a tedious endeavor and should be considered only if that level of detail is necessary for the application of the data.

The primary Level 1 evaluation results should be saved as was done with the molecular target evaluation using the “Download Table” function. Generally, CSV formatted files are easier to manipulate later.

7.3 Generating a list of species that are likely to conserve the pathway

SeqAPASS results from key pathway proteins and the molecular target can be used to generate a list of species where the pathway is likely conserved. There are a variety of ways of doing this. However, making use of R software is recommended. Basic filtering and merging functions are provided below to assist with this process:

```
library(tidyverse)

# Function to remove hyperlinks in CSV files
cleanCSV <- function(colName, inputData) {
  for(i in 1:nrow(inputData)){
    inputData[i,colName] <- lapply(inputData[i,colName], FUN = function(x) {gsub("[\\\")]", "", strsplit(x,split=",")[[1]][2])})
  }
  inputData
}

columns <- c("NCBI.Accession", "Scientific.Name", "Species.Tax.ID")

# Function to extract susceptible species from SeqAPASS reports
pull_species <- function(x) {
  cleaned <- cleanCSV(columns, x)
  if ("Susceptibility.Prediction" %in% names(cleaned)) {
    cleaned <- cleaned %>%
      filter(Susceptibility.Prediction == "Y")
  } else if ("Similar.Susceptibility.as.Template" %in% names(cleaned)) {
    cleaned <- cleaned %>%
      filter(Similar.Susceptibility.as.Template == "Y")
  } else {
    warning("Neither 'Susceptibility.Prediction' nor 'Similar.Susceptibility.as.Template' column found.")
    return(NULL)
  }
  cleaned <- cleaned %>%
    select(Species.Tax.ID, Scientific.Name, Taxonomic.Group)
  return(cleaned)
}

# Function to process all CSV files in a folder
process_folder <- function(folder_path) {
  csv_files <- list.files(folder_path, pattern = "\\.csv$", full.names = TRUE, recursive = TRUE)
  seqfiles <- lapply(csv_files, read.csv)
  names(seqfiles) <- sapply(basename(csv_files), function(x) strsplit(x, "_")[[1]][1])
  sus_species <- lapply(seqfiles, pull_species)
  shared_sus_species <- Reduce(intersect, sus_species)
  output_file <- file.path(folder_path, "shared_susceptible_species.csv")
```

```
  write.csv(shared_sus_species, output_file, row.names = FALSE)
  cat("Processing complete. Results saved to:", output_file, "\n")
  return(list(seqfiles = seqfiles, sus_species = sus_species, shared_sus_species = shared_sus_species))
}

# Replace with the actual path to your folder containing CSV files
folder_path <- "path/to/SeqAPASS/reports/folder"
setwd(folder_path)
results <- process_folder(folder_path)
```

To run this code, copy and paste it into an R document within RStudio. Then prepare the SeqAPASS reports by moving all 11 primary reports (1 for the molecular target and 10 for the key pathway proteins) into a single folder. It is a good practice to name the folder after the biological pathway that the proteins are meant to represent. Then, update the "folder_path" object by replacing "path/to/SeqAPASS/reports/folder" with the actual path to the folder. The code can then be run either line by line using "Crtl + Enter" (on Windows OS) or all at once with "Ctrl + a" then "Ctrl + Enter" or by clicking on the "Run" button in the script window in RStudio: 

This code will output a file called "shared_susceptible_species.csv" formatted like so:

	A	B	C
1	Species.Tax.ID	Scientific.Name	Taxonomic.Group
2	9606	Homo sapiens	Mammalia
3	9598	Pan troglodytes	Mammalia
4	9597	Pan paniscus	Mammalia
5	9595	Gorilla gorilla gorilla	Mammalia
6	9601	Pongo abelii	Mammalia
7	9544	Macaca mulatta	Mammalia
8	9565	Theropithecus gelada	Mammalia
9	9555	Papio anubis	Mammalia
10	9541	Macaca fascicularis	Mammalia
11	9545	Macaca nemestrina	Mammalia

It is a list of the species that had a susceptibility call of "yes" for each of the proteins. Since these proteins are representative of a biological pathway, this list of "susceptible" species can be extended to conclude that the biological pathway is likely conserved in this list of species, according to SeqAPASS. This list of species can be supported by making comparisons to the G2P-SCAN results.

In combination the SeqAPASS and G2P-SCAN can add consensus to ortholog candidate identification among the species evaluated in both tools. Ortholog candidate identification is critical to both tools for making cross-species predictions of conservation. The Level 1 Full Report results containing hundreds of species can be examined to identify the 6 species evaluated by G2P-SCAN and compared for consensus with ortholog candidates identified in SeqAPASS adding greater confidence to the prediction. Concepts of cross-species extrapolation can be applied in cases where there is consensus to broaden the predictions of pathway conservation across broader taxonomic groups.

8. Interpretation of Data for Regulatory Use

8.1 Surrogate model organisms in toxicity testing

Model organisms in ecotoxicology are generally species that are extensively studied to understand the effects of chemicals on biological systems. Usually, they have been chosen because they are easy to maintain and manipulate in a laboratory setting, and they provide valuable insights into the interactions of chemicals with biological processes. Commonly used model organisms in ecotoxicology include various species of algae, invertebrates, and fish. This approach helps predict the toxicity of chemicals for other organisms and to avoid more extensive animal trials, which can be expensive, time-consuming, and raise ethical concerns. They are often used in standardized toxicity tests, which have high reproducibility and acceptance. However, there are often concerns about their ecological relevance and representation of other organisms in the environment. Cross-species extrapolation approaches, such as the methodology provided by SeqAPASS and G2P-SCAN, can significantly enhance the confidence in using model organisms in ecotoxicology by providing a more comprehensive understanding of how different species respond to chemical exposures. These methods involve using data from well-studied model organisms to predict the effects of chemicals on other species, thereby bridging the gap between laboratory findings and real-world environmental impacts. For instance, techniques such as protein sequence alignment and molecular docking can help predict how chemicals interact with biological targets across different species. This helps in understanding the structural conservation of chemical targets and provides evidence for the utility of relying on specific model species tests to predict toxicity in other species. Overall, it does not only help in identifying potential risks to a broader range of non-model organisms, but also supports regulatory decision-making by providing robust empirical evidence.

8.2 Endangered Species Act

The Endangered Species Act (ESA) is a critical piece of legislation in the USA aimed at protecting and conserving species that are at risk of extinction. It provides a framework for the conservation and recovery of endangered and threatened species and the ecosystems upon which they depend. Through the SeqAPASS output listed threatened and endangered species are identified where sequence information exists yielding predictions of chemical susceptibility (and hence at risk following the specific chemical exposure) based on lines of evidence for conservation of the molecular target responsible for the toxic effect. This knowledge provides the scientific basis to guide informed conservation strategies and implementing recovery plans for designating critical habitats and listed species.

8.3 Endocrine Disruption

Endocrine disruption occurs when chemicals interfere with the endocrine system of an organism, and may lead to various adverse effects, including developmental and reproduction abnormalities. It is usually driven by chemicals that can mimic, block, or interfere with organisms' hormones. The European Commission has recently adopted a strategy to protect EU citizens and the environment from hazardous chemicals, including endocrine disruptors, which involves their identification and regulation. The US EPA and the National Institute of Environmental Health Sciences (NIEHS) have also been actively researching and regulating endocrine disruptors. Overall, there is the need to understand the concentrations of chemicals that may cause such adverse effects and develop standard test methods to identify potential endocrine disruptor to allow for chemical identification. In this respect cross-species extrapolation approaches and the integrated methodology presented here based on a mechanistic understanding of toxicity effects of a chemical and their domain of applicability across species, can greatly support the prioritization of chemicals for potential endocrine disruption and conducting directed research to better understand the real-life impacts of exposures to such potential endocrine-disrupting chemicals. Using the example described above in the assessment of diethylstilbestrol, G2P-SCAN mapped 12 Reactome pathways using ESR1

resulting in two of the mapped pathways, RUNX1 (regulating transcription of genes involved in WNT signaling (R-HSA-8939256)) and RUNX1 (regulating estrogen receptor mediated transcription(R-HSA-8931987)) being prioritized for having ESR1 coverage of over 10 %. Subsequent SeqAPASS evaluations on the key pathway proteins for R-HSA-8939256 resulted in 272 “susceptible” species across 8 taxa (Actinopteri, Amphibia, Aves, Chondrichthyes, Crocodylia, Lepidosauria, Mammalia, and Testudinata). As such it is possible to understand how such approaches could be applied in the identification of endocrine disruption effects and their evaluation of susceptibility across species (i.e. ESR1 mostly conserved across the vertebrates only). This may also guide subsequent intelligent testing strategies by focusing on the most relevant species at risk and minimizing unnecessary testing.

8.4 Pesticide Registration

Before an active substance can be used in a Plant Protection Product (PPP), it must be approved by the relevant regulatory body, which evaluates the safety of the chemical and lays out the rules for its authorization and use. While the registration requirements and procedures differ from country to country, they usually abide to strict scientific and legal protocols to ensure that the substance does not pose unacceptable risks to human health or the environment. In this respect, cross-species extrapolation approaches become essential for enhancing the predictive accuracy and ecological relevance of pesticide registration. By using data from well-studied model organisms to predict the effects of chemicals on other species, the evaluation of the specificity of the effect can be facilitated. The identification of the molecular targets and the follow-up genes/proteins responsible for the cascade of events leading to an adverse outcome through recognized pathways is of paramount importance in the assessment of the specificity of the effect across species (e.g. while the effect of chlorpyrifos is directed to insects, the molecular targets, i.e. ACHE/BCHE , can be identified across a wider range of species including vertebrates and human, justifying the side effects of the use of such pesticides and providing additional information over the restriction measurements of its use that must be put in place).

REFERENCES

Scientific Publications:

Colbourne JK, Shaw JR, Sostare E, Rivetti C, Derelle R, Barnett R, Campos B, LaLone C, Viant MR, Hodges G. Toxicity by descent: A comparative approach for chemical hazard assessment. Environmental Advances. 2022 Oct 1;9:100287.

Haigis AC, Vergauwen L, LaLone CA, Villeneuve DL, O’Brien JM, Knapen D. Cross-species applicability of an adverse outcome pathway network for thyroid hormone system disruption. Toxicological Sciences. 2023 Sep 1;195(1):1-27.

Jensen MA, Blatz DJ, LaLone CA. Defining the biologically plausible taxonomic domain of applicability of an adverse outcome pathway: A case study linking nicotinic acetylcholine receptor activation to colony death. Environmental Toxicology and Chemistry. 2023 Jan;42(1):71-87.

LaLone CA, Villeneuve DL, Lyons D, Helgen HW, Robinson SL, Swintek JA, Saari TW, Ankley GT. Editor’s highlight: sequence alignment to predict across species susceptibility (SeqAPASS): a web-based tool for addressing the challenges of cross-species extrapolation of chemical toxicity. Toxicological Sciences. 2016 Oct 1;153(2):228-45.

Perkins EJ, Ankley GT, Crofton KM, Garcia-Reyero N, LaLone CA, Johnson MS, Tietge JE, Villeneuve DL. Current perspectives on the use of alternative species in human health and ecological hazard assessments. Environmental health perspectives. 2013 Sep;121(9):1002-10.

Rivetti C, Houghton J, Basili D, Hodges G, Campos B. Genes-to-Pathways Species Conservation Analysis: Enabling the Exploration of Conservation of Biological Pathways and Processes Across Species. Environmental Toxicology and Chemistry. 2023 May;42(5):1152-66, https://doi.org/10.1002/etc.5600.

Schumann, P. et al. (2024), “Combination of computational new approach methodologies for enhancing evidence of biological pathway conservation across species”, Science of The Total Environment, Vol.

912, p. 168573, https://doi.org/10.1016/j.scitotenv.2023.168573.
Spurgeon D, Lahive E, Robinson A, Short S, Kille P. Species sensitivity to toxic substances: Evolution, ecology and applications. Frontiers in Environmental Science. 2020 Dec 1;8:588380.
Vliet SM, Markey KJ, Lynn SG, Adetona A, Fallacara D, Ceger P, Choksi N, Karmaus AL, Watson A, Ewans A, Daniel AB. Weight of evidence for cross-species conservation of androgen receptor-based biological activity. Toxicological Sciences. 2023 Jun 1;193(2):131-45. (Vliet et al., 2023a)

Online Tools and Databases

DrugBank. DrugBank Database. Retrieved from https://www.drugbank.ca/ [Accessed: October 2025]
National Center for Biotechnology Information (NCBI). Conserved Domain Database (CDD). Retrieved from https://www.ncbi.nlm.nih.gov/cdd/ [Accessed: October 2025]
Posit Software, PBC. RStudio Desktop. Retrieved from https://posit.co/download/rstudio-desktop/ [Accessed: October 2025]
Research Collaboratory for Structural Bioinformatics (RCSB). Protein Data Bank (PDB). Retrieved from https://www.rcsb.org/ [Accessed: October 2025]
Seacunilever. Genes-to-Pathways Species Conservation Analysis (G2P-SCAN) [GitHub repository]. Retrieved from https://github.com/seacunilever/G2P-SCAN [Accessed: October 2025]
STRING Consortium. STRING: Functional protein association networks. Retrieved from https://string-db.org/ [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). 2-Ethylhexanoic acid. Retrieved from https://comptox.epa.gov/dashboard/chemical/details/DTXSID9025293 [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). AOP-Database (AOP-DB). Retrieved from https://aopdb.epa.gov/ [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Butylparaben. Retrieved from https://comptox.epa.gov/dashboard/chemical/details/DTXSID3020209 [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). CompTox Chemicals Dashboard. Retrieved from https://comptox.epa.gov/dashboard/ [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Diethylstilbestrol. Retrieved from https://comptox.epa.gov/dashboard/chemical/details/DTXSID3020465 [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Dibutyl phthalate. Retrieved from https://comptox.epa.gov/dashboard/chemical/details/DTXSID2021781 [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Oxybenzone. Retrieved from https://comptox.epa.gov/dashboard/chemical/details/DTXSID3022405 [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). SeqAPASS Tool. Retrieved from https://seqapass.epa.gov/seqapass/ [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). SeqAPASS Login Guide. Retrieved from https://www.epa.gov/comptox-tools/guide-login-seqapass-users [Accessed: October 2025]
United States Environmental Protection Agency (US EPA). Topiramate. Retrieved from https://comptox.epa.gov/dashboard/chemical/details/DTXSID8023688 [Accessed: October 2025]
Wishart Research Group. Toxin and Toxin Target Database (T3DB). Retrieved from https://www.t3db.ca/
Cytoscape Consortium. Cytoscape Network Visualization Tool. Retrieved from https://cytoscape.org/ [Accessed: October 2025]

Annex D. READ ME to accompany ANNEX C Data for Case Studies

Please refer to the Excel file for more details on the data.

INPUT AND OUPUT FOR ALL 3 CASE STUDIES: PPAR, ESR1, AND GABRA1

Chemical Names are used to initiate Target Selection

Target Selection Tab

Captures information for identification of targets for chemicals of interest and provides reasoning for support of the chemical-protein interaction from existing scientific evidence. Chemical information columns A-C; Gene Target information columns D-H; Data describing target evidence columns I-U; Evidence sufficient to support target "Yes" or "No" in column V; justification for target call and sources columns W-X. Column D, *Putative_target_genes* is the input to G2P-SCAN. Columns E, *gene_name*, and H, *taxon_name*, or G, *ncbi_accession,* is the input into SeqAPASS Level 1.

SeqAPASS Levels 1-4 can be run on any known chemical-protein interaction for evaluating conservation and making predictions of chemical susceptibility across hundreds of species. This can be done at any time to define the biologically plausible taxonomic domain of applicability for a molecular initiating event (e.g., PPARα, ESR1, GABR1) in an adverse outcome pathway or to extrapolate toxicity knowledge across species to predict chemical susceptibility. Once targets are identified and species or the National Center for Biotechnology Information (NCBI) Protein accession are found, the SeqAPASS Level 1 can be queried. SeqAPASS Level 2 and Level 3 information was identified using literature and the NCBI conserved domain database link found in SeqAPASS.

SeqAPASS results for the 2-ethylhexanoic acid-PPARα interaction are found on the next 3 Tabs:

2-EHA_PPAR_SeqAPASS_L1; Level 1 Primary amino acid sequence comparisons and susceptibility predictions

Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

2-EHA_PPAR_SeqAPASS_L2; Level 2 Functional domain sequence comparisons and susceptibility predictions

Headers in the Level 2 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that

species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, Domain Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

SeqAPASS results for the Butylparaben-estrogen receptor 1 interaction are found on the next 3 Tabs:

ButylparabenESR1_SeqAPASS_L1; Level 1 Primary amino acid sequence comparisons and susceptibility predictions

Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

ButylparabenESR1_SeqAPASS_L2; Level 2 Functional domain sequence comparisons and susceptibility predictions

Headers in the Level 2 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, Domain Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

ButylparabenESR1_SeqAPASS_L3; Level 3 critical individual amino acid sequence comparisons and susceptibility predictions

Headers in the Level 3 Report dataset: Data Version (referring to SeqAPASS data version), Job Name, NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Scientific Name (of the species), Common Name (of the species), Protein Name, Analysis Completed, Similar Susceptibility as Template (susceptibility prediction in comparison to the user selected template species), Position 1 (position of the amino acid in the sequence), Amino Acid 1 (single letter abbreviation for the amino acid, e.g. R is arginine), Total Match 1 (both side chain and MW are Y = Yes; either side chain or MW are Y = Yes; both side chain and MW are N – No)

SeqAPASS results for the Diethylstilbestrol-estrogen receptor 1 interaction are found on the next 3 Tabs:

DES_ESR1_SeqAPASS_L1; Level 1 Primary amino acid sequence comparisons and susceptibility predictions

Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that

species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

DES_ESR1_SeqAPASS_L2; Level 2 Functional domain sequence comparisons and susceptibility predictions

Headers in the Level 2 datasets include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, Domain Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

DES_ESR1_SeqAPASS_L3; Level 3 critical individual amino acid sequence comparisons and susceptibility predictions

Headers in the Level 3 Report dataset: Data Version (referring to SeqAPASS data version), Job Name, NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Scientific Name (of the species), Common Name (of the species), Protein Name, Analysis Completed, Similar Susceptibility as Template (susceptibility prediction in comparison to the user selected template species), Position 1 (position of the amino acid in the sequence), Amino Acid 1 (single letter abbreviation for the amino acid, e.g. R is arginine), Total Match 1 (both side chain and MW are Y = Yes; either side chain or MW are Y = Yes; both side chain and MW are N – No)

SeqAPASS results for the Topiramate_GABRA1 interaction are found on the next 3 Tabs:

Topiramate_GABRA1_SeqAPASS_L1; Level 1 Primary amino acid sequence comparisons and susceptibility predictions

Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

Topiramate_GABRA1_SeqAPASS_L2; Level 2 Functional domain sequence comparisons and susceptibility predictions

Headers in the Level 2 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the

species), Common Name (of the species), Protein Name, Domain Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX

DB Versions Tab

Describes the inputs and Data sources and versions for G2P-SCAN evaluations in the case

PPARA CASE STUDY

PPARA-counts, PPARA-orthologs, PPARA-families, and PPARA mapped pathways Tabs contain output from G2P-SCAN for PPARA case example. Column B, Gene.Symbol, from PPARA-mapped pathways Tab is input to STRING to generate the pathway-specific protein-protein interaction (PPI) networks found in PPARA-PPI Network Tab.

PPARA-PPI Network and PPARA-Cytoscape_PPI net Tabs

Contains protein-protein interaction (PPI) network data for one example pathway (R-HSA-2426168). This table describes the connections between each node of the cluster with columns A-G providing scores across various metrics for determining interconnectivity and all subsequent columns describing the connections between nodes. This network was input to Cytoscape software for visualization and molecular complex detection (MCODE) analysis. MCODE evaluates the degree of protein interconnectedness within the PPI network.

PPARA-MCODE Tab

Contains MCODE output for the highest scoring molecular complex of the PPI network of the Reactome pathway R-HSA-2426168. The column “MCODE::Score (7)” contains the score for each node used to determine the degree of interconnectivity. The “name” column contains the gene ID of the protein.

PPARA-Top10 Tab

Contains the identification of the Top 10 MCODE scores with associated NCBI accessions in column G, accession. The accessions are input to SeqAPASS Level 1. The column “acceptable?” refers to whether the accession used for the protein is an acceptable input for SeqAPASS. More information on this is found in methods section 2.4 of the associated publication (Schumann et al., 2024).

PPARA_SREBF1_SeqAPASS Tab

The protein accessions from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_FDFT1_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic

Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_SREBF2_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_HMGCS1_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_RXRA_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_TBL1X_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using

reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_CARM1_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_NCOA1_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_NCOA2_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_TGS1_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

PPARA_2-EHA_ComboSeqAPASS Tab

R software is used to combine species from SeqAPASS, filtering and merging results from SeqAPASS output of PPARA across all 10 proteins evaluated representing the most interconnected proteins identified in MCODE. Data includes TaxID from NCBI Column B; species scientific name in column C; and the Taxonomic Group in Column D. These data define the biologically plausible taxonomic domain of applicability for susceptibility across the pathway identified from G2P-SCAN for 2-ethylhexanoic acid interaction with PPARα.

ESR1 CASE STUDY

ESR1-counts, ESR1-orthologs, ESR1-families, and ESR1 mapped pathways Tabs contain output from G2P-SCAN for ESR1 case example. Column B, *Gene.Symbol*, from ESR1-mapped pathways Tab is input to STRING to generate the pathway-specific protein-protein interaction (PPI) networks found in *ESR1-PPI Network* Tab

ESR1-PPI Network and ESR1-Cytoscape_PPI net Tabs

Contains protein-protein interaction (PPI) network data for one example pathway (R-HSA-9018519). This table describes the connections between each node of the cluster with columns A-G providing scores across various metrics for determining interconnectivity and all subsequent columns describing the connections between nodes. This network was input to Cytoscape software for visualization and molecular complex detection (MCODE) analysis. MCODE evaluates the degree of protein interconnectedness within the PPI network.

ESR1-MCODE Tab

Contains MCODE output for the highest scoring molecular complex of the PPI network of the Reactome pathway R-HSA-9018519. The column “MCODE::Score (1)” contains the score for each node used to determine the degree of interconnectivity. The “name” column contains the gene ID of the protein.

ESR1-Top10 Tab

Contains the identification of the Top 10 MCODE scores with associated NCBI accessions in column G, *accession*. The accessions are input to SeqAPASS Level 1. The column “acceptable ?” refers to whether the accession used for the protein is an acceptable input for SeqAPASS. More information on this is found in methods section 2.4 of the associated publication (Schumann et al., 2024).

ESR1_POLR2C_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_POLR2E_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the

species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_HIST1H2BJ_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_POLR2H_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_ HIST1H2BK_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_POLR2K_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for

susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_ HIST2H2AC_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_POLR2L_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_HIST1H2BD_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_HIST1H2BO_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

ESR1_Butylparaben_ComboSeqAPASS Tab

R software is used to combine species from SeqAPASS, filtering and merging results from SeqAPASS output of ESR1 in relation to butylparaben across all 10 proteins evaluated representing the most interconnected proteins identified in MCODE. Data includes TaxID from NCBI Column B; species scientific name in column C; and the Taxonomic Group in Column D. These data define the biologically plausible taxonomic domain of applicability for susceptibility across the pathway identified from G2P-SCAN for Butylparaben interaction with ESR1.

ESR1_Diethystilbestrol_ComboSeqAPASS Tab

R software is used to combine species from SeqAPASS, filtering and merging results from SeqAPASS output of ESR1 in relation to diethylstilbestrol across all 10 proteins evaluated representing the most interconnected proteins identified in MCODE. Data includes TaxID from NCBI Column B; species scientific name in column C; and the Taxonomic Group in Column D. These data define the biologically plausible taxonomic domain of applicability for susceptibility across the pathway identified from G2P-SCAN for Diethylstilbestrol interaction with ESR1.

GABRA1 CASE STUDY

GABRA1-counts, GABRA1-orthologs, GABRA1-families, and GABRA1 mapped pathways Tabs contain output from G2P-SCAN for GABRA1 case example. Column B, Gene.Symbol, from GABRA1-mapped pathways Tab is input to STRING to generate the pathway-specific protein-protein interaction (PPI) networks found in GABRA1-PPI Network Tab

GABRA1-PPI Network and GABRA1-Cytoscape_PPI net Tabs

Contains protein-protein interaction (PPI) network data for one example pathway (R-HSA-1236394). This table describes the connections between each node of the cluster with columns A-G providing scores across various metrics for determining interconnectivity and all subsequent columns describing the connections between nodes. This network was input to Cytoscape software for visualization and molecular complex detection (MCODE) analysis. MCODE evaluates the degree of protein interconnectedness within the PPI network.

GABRA1-MCODE Tab

Contains MCODE output for the highest scoring molecular complex of the PPI network of the Reactome pathway R-HSA-1236394. The column “MCODE:Score (4)” contains the score for each node used to determine the degree of interconnectivity. The “name” column contains the gene ID of the protein.

GABRA1-Top10 Tab

Contains the identification of the Top 10 MCODE scores with associated NCBI accessions in column G, accession. The accessions are input to SeqAPASS Level 1. The column “acceptable?” refers to whether the accession used for the protein is an acceptable input for SeqAPASS. More information on this is found in methods section 2.4 of the associated publication (Schumann et al., 2024).

GABRA1_NRG4_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for

susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_HBEGF_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_NRG3_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_NRG2_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_BTC_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_NRG1_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_ERB4_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_EREG_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_GABRB2_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_GABRG2_SeqAPASS Tab

Protein Accession from Top 10 MCODE for Human Protein Target was used to query SeqAPASS Level 1. Headers in the Level 1 dataset include: Data Version (referring to SeqAPASS data version), NCBI Accession (Protein ID derived from NCBI), Protein Count (# of protein sequences associated with that species in NCBI), Species Tax ID (Species ID derived from NCBI), Taxonomic Group, Filtered Taxonomic Group (User has the option to filter to a different level of the taxonomic hierarchy), Scientific Name (of the species), Common Name (of the species), Protein Name, BLASTp Bitscore (metric derived from NCBI BLAST tools to describe overall quality of the alignment), Ortholog Candidate (derived from using reciprocal best hit BLAST), Ortholog Count (total orthologs found in the dataset), Cut-off (used for susceptibility prediction), Percent Similarity (of sequence compared to the query sequence in the top row), Susceptibility Prediction, Analysis Completed, Eukaryote, ECOTOX.

GABRA1_TopiramateComboSeqAPASS Tab

R software is used to combine species from SeqAPASS, filtering and merging results from SeqAPASS output of GABRA1 across all 10 proteins evaluated representing the most interconnected proteins identified in MCODE. Data includes TaxID from NCBI Column B; species scientific name in column C; and the Taxonomic Group in Column D. These data define the biologically plausible taxonomic domain of applicability for susceptibility across the pathway identified from G2P-SCAN for Topiramate interaction with GABRA1.