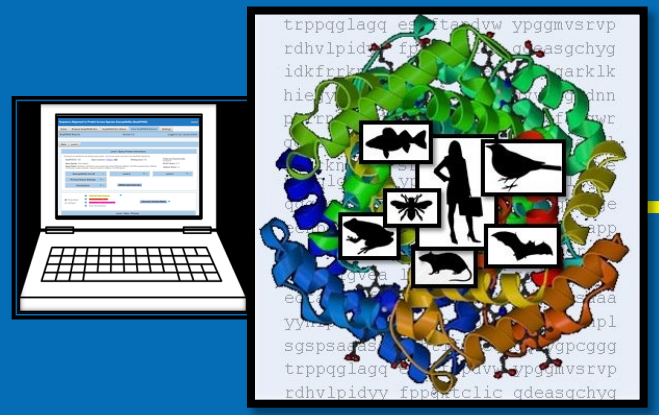
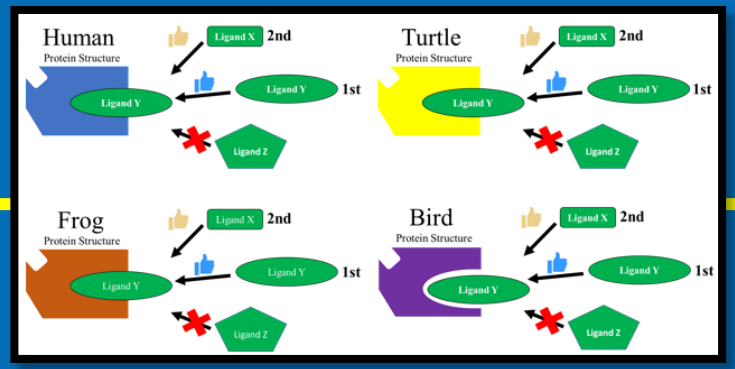
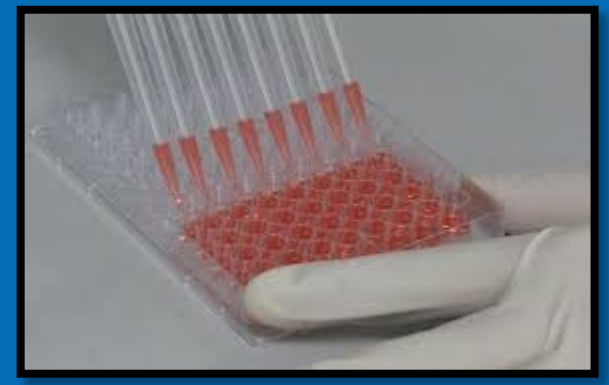
United States Environmental Protection Agency

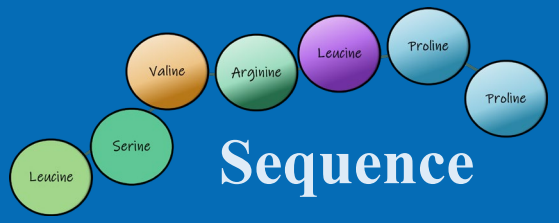
Advances in bioinformatics to improve cross species extrapolation of toxicity information for chemical safety evaluations

SeqAPASS v6

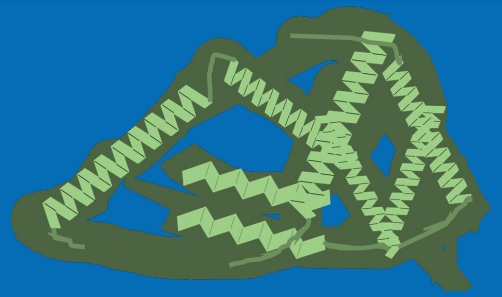




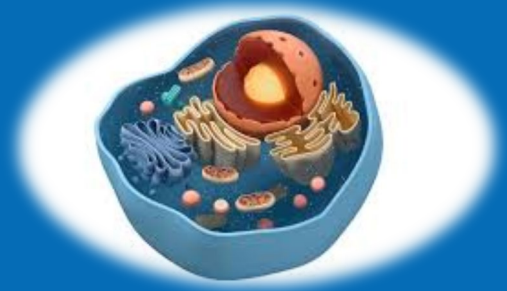




Sequence



Structure



Function

Presenter: Carlie A. LaLone, Ph.D.

March 2022

Office of Research and Development
Center for Computational Toxicology and Exposure, Great Lakes Toxicology and Ecology Division

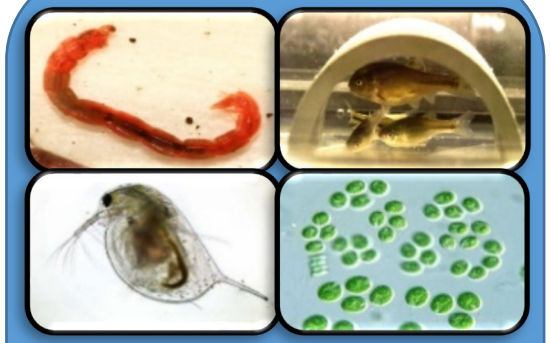
Overview

- Need for cross species extrapolation
- Bioinformatics to advance extrapolation
- The Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS) tool
- Demonstrated applications
- Future vision for incorporating structural evaluations
- Bringing champions in the field together to advance the science for action through an international consortium

Toxicity Testing to Understand Chemical Safety

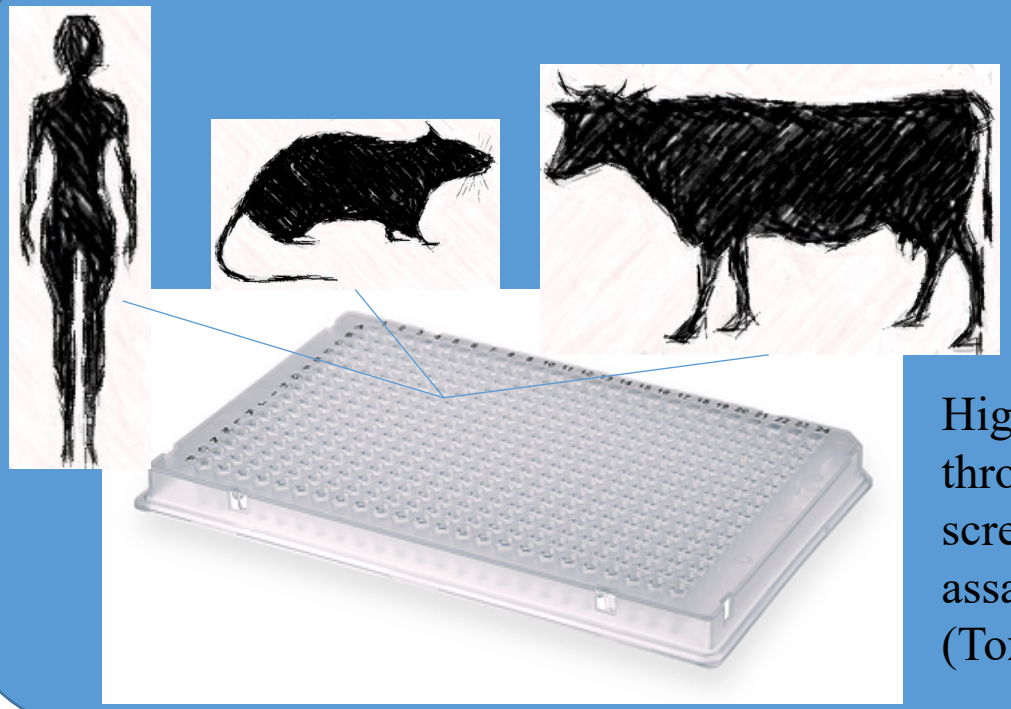
- **<u>US EPA Examples:</u>**
- *Clean Air Act*
- *Clean Water Act*
- *Resource Recovery Act*
- *Endangered Species Act*
- *Food Quality Protection Act*
- *Endocrine Disruptor Screening Program*
- *Federal Insecticide, Fungicide, and Rodenticide Act*
- *Frank R. Lautenberg Chemical Safety for the 21st Century Act*
- *Comprehensive Environmental Response, Compensation, and Liability Act*
- *Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and Their Uses*

Need for Advances in Species Extrapolation



High throughput transcriptomics

Whole organism toxicity testing



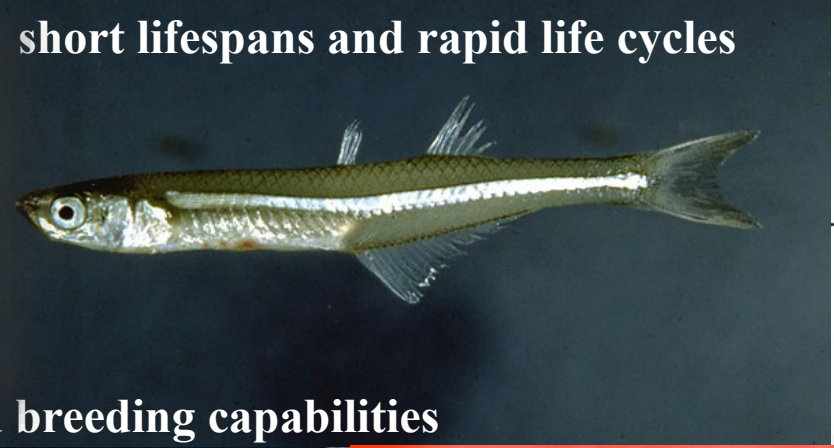
High-throughput screening assays (ToxCast)

Define the taxonomic domain of applicability in AOP development

Use of model organisms as surrogates representing the diversity of species in the environment

cheap and readily available

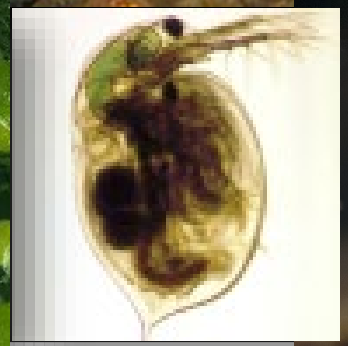
short lifespans and rapid life cycles



easy maintenance and good breeding capabilities

ability to control diet and surroundings

requires least space and time-consuming care



Species Extrapolation

What is it?

- Using existing knowledge about one species to estimate, predict, project, or infer the effect, impact, or trajectory of another species
    - For chemical safety typically dealing with toxicity

Why is it important:

- Limited or no toxicological data for the animal or plant species of interest – reliance on surrogate (model organisms)
    - Impractical to generate new data for all species
- Testing resources are limited
    - International interest to reduce animal use
    - Ever-increasing demand to evaluate more chemicals in a timely and sometimes expedited manner
- Sensitivity of species must be estimated based on scientifically-sound methods of cross-species extrapolation
    - Immense diversity of species in the wild
    - Important challenge for species listed under the Endangered Species Act

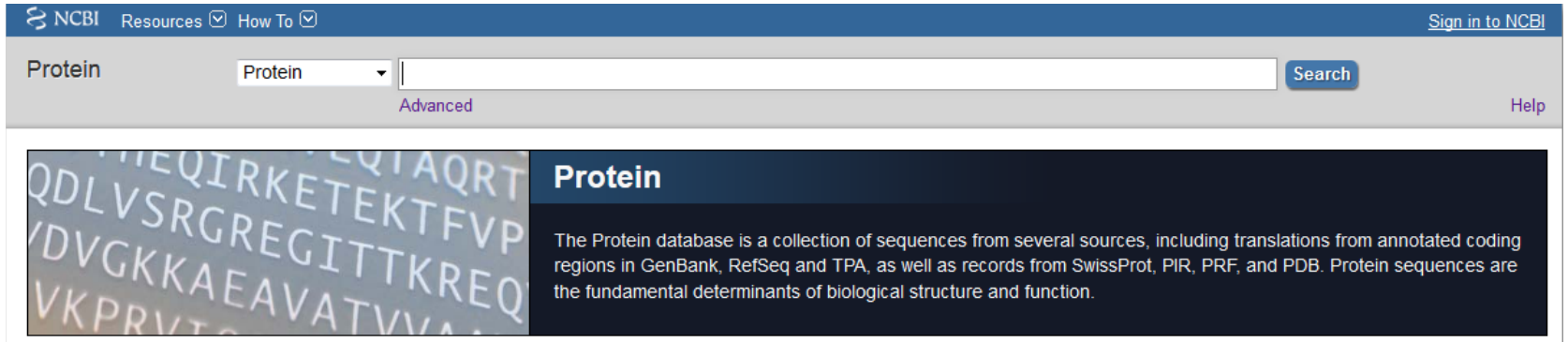
Where could we begin in understanding species similarities and differences?

Look for existing, expanding data that does not require the destruction of live organisms

Sequence and structural data: New tools and technologies have emerged

- Improved sequencing technologies
- Large databases of sequence data

NCBI: 224,211,842 Proteins representing 117,030 Organisms



EPA United States Environmental Protection Agency

Bioinformatics

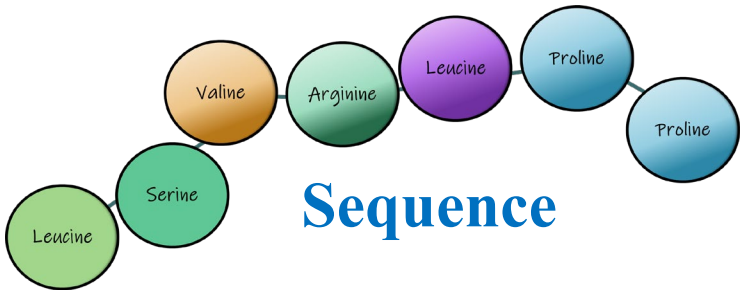
- Combines mathematics, information science, and biology to **answer biological questions**

- Developing methodology and analysis tools to **explore large volumes of biological data**
  - Query, extract, store, organize, systematize, annotate, visualize, mine, and interpret complex data
    - Usually **pertains to DNA and amino acid sequences**

Let the computers do the work

EPA United States Environmental Protection Agency

Begin Simple and Advance as the Science Advances



Sequence

Structure

Function

Consider sequence and structural attributes to understand protein conservation across species

EPA United States Environmental Protection Agency

TOXICOLOGICAL SCIENCES, 153(2), 2016, 228–245

OXFORD SOT Society of Toxicology www.toxsci.oxfordjournals.org

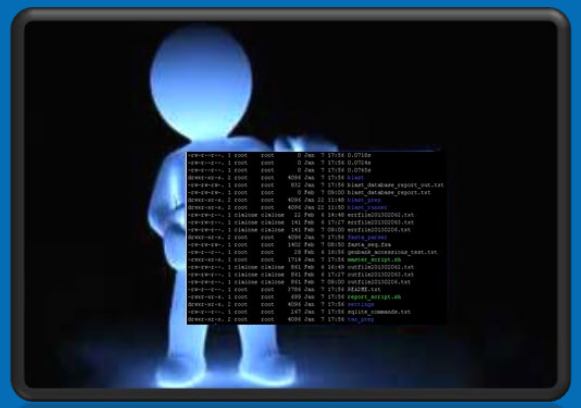
doi: 10.1093/toxsci/kfw119
Advance Access Publication Date: June 30, 2016
Research article

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS): A Web-Based Tool for Addressing the Challenges of Cross-Species Extrapolation of Chemical Toxicity

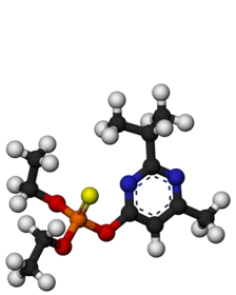
Carlie A. LaLone,*,[1] Daniel L. Villeneuve,* David Lyons,† Henry W. Helgen,‡ Serina L. Robinson,§,[2] Joseph A. Swintek,¶ Travis W. Saari,* and Gerald T. Ankley*

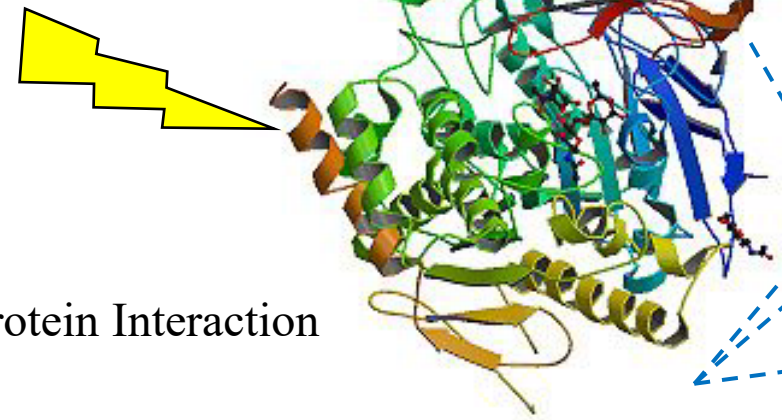
https://seqapass.epa.gov/seqapass/

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)



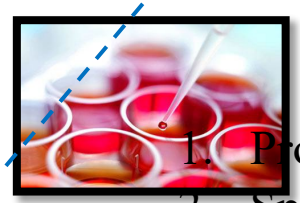
What information is required for a SeqAPASS query?





Chemical-Protein Interaction

Knowledge of a sensitive or targeted species

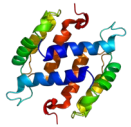


Knowledge of the model organism used in an *in vitro* assay

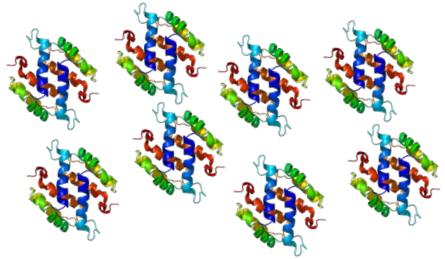
1. Protein
2. Species

Knowledge of the species for which the Key Event was developed

Chemical Molecular Target in Target Species



≅



Compare to Millions of Proteins From Thousands of Species

Greater similarity = Greater likelihood that chemical can act on the protein

Line of Evidence: Predict Potential Chemical Susceptibility Across Species

Flexible Analysis Based On Available Data

Level 1 Primary Amino Acid Sequence Alignments

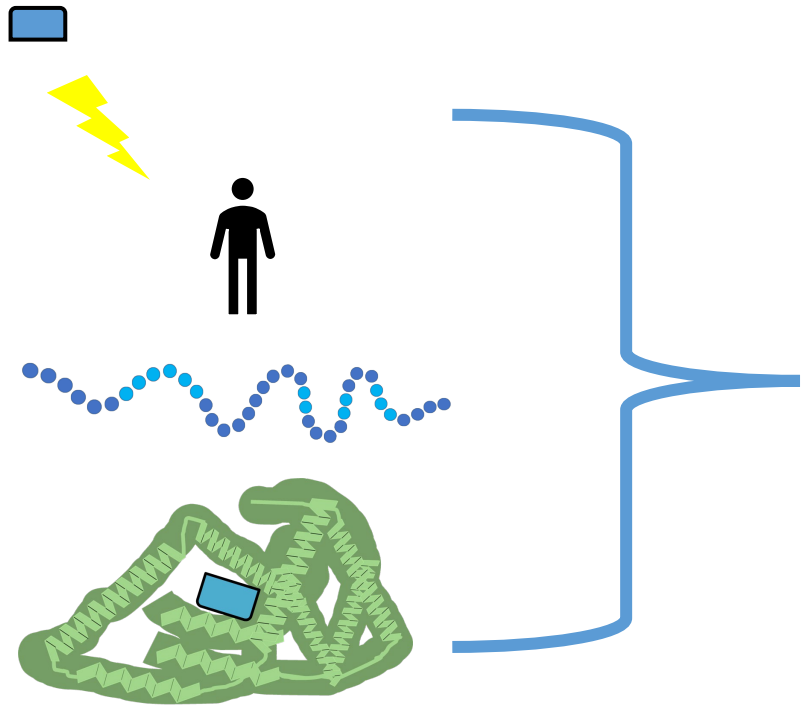
Level 2 Conserved Functional Domain Alignments

Level 3 Critical (Close Contact) Amino Acid Conservation

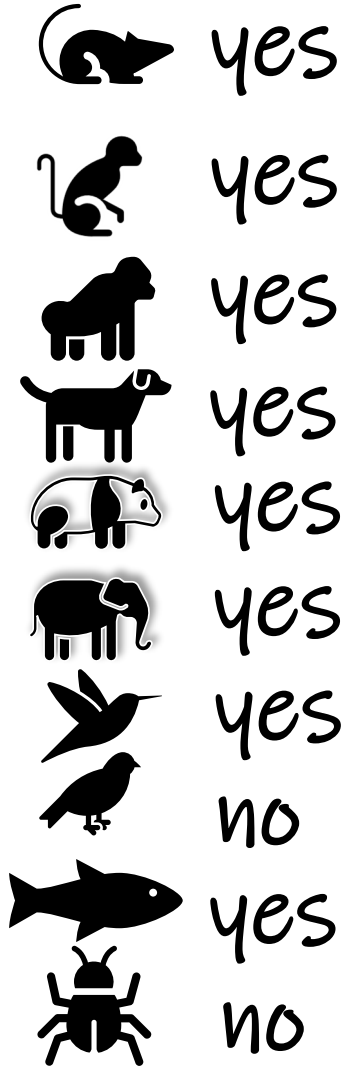
seqapass.epa.gov/seqapass/

Gather Lines of Evidence Toward Protein Conservation

SeqAPASS Predicts Likelihood of Similar Susceptibility based on Sequence Conservation:



yes

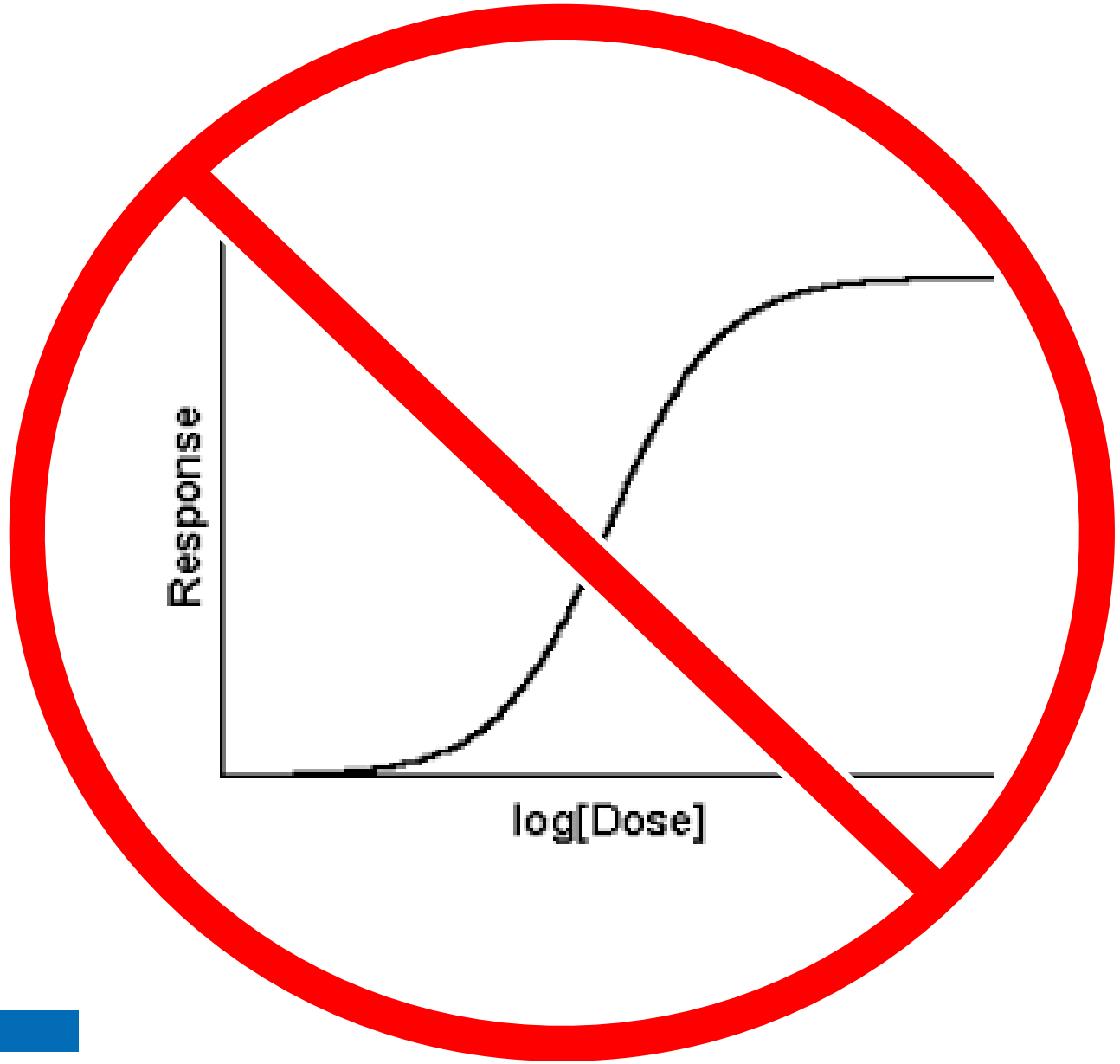


yes
yes
yes
yes
yes
yes
yes
no
yes
no

Line(s) of evidence indicate

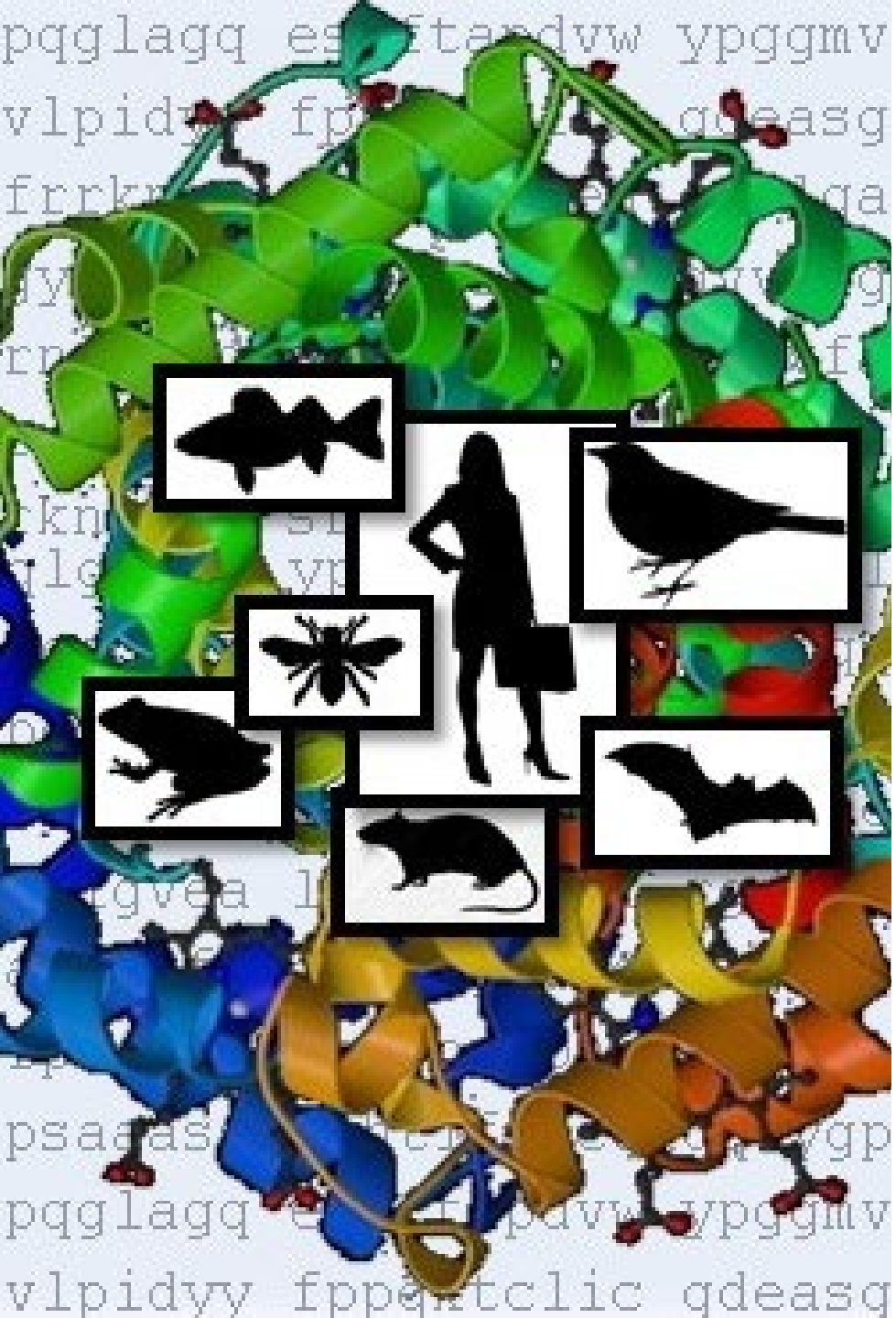
- The protein is conserved
- The protein is NOT conserved

SeqAPASS DOES NOT predict the degree of sensitivity/susceptibility:



Factors that make a species sensitive

- Exposure
- Dose
- ADME
- Target receptor availability
- Life stage
- Life history
- etc.
- etc.



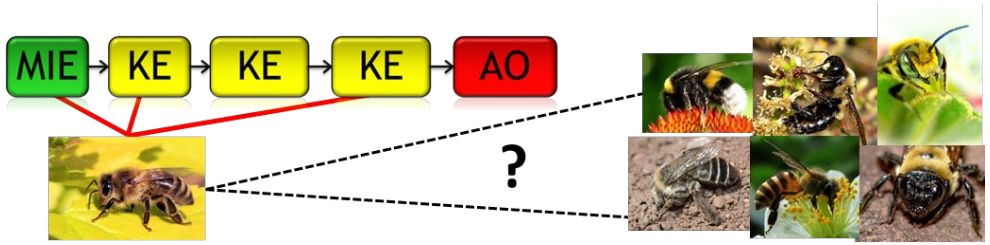
Strengths of SeqAPASS

- **Publicly available** to all
- Lines of evidence for conservation for **100s-1000s of species** rapidly
- Takes advantage of **well-established tools and databases**
- Streamlined, consistent, transparent, and published methods
  - **Case examples** to demonstrate applications
- **Guides users** to appropriate input
- **Evolves** as bioinformatics approaches become more user friendly
  - Smart automation or semi-automation


Application of SeqAPASS
MIE
KE
KE
KE
AO

Applications of Bioinformatics: Case Studies

- **Extrapolate adverse outcome pathway knowledge across species**
  - Define the taxonomic relevance: Apis vs Non-Apis bees



- **Extrapolate high throughput screening data**
  - Chemicals that target human estrogen receptor alpha, androgen receptor, steroidogenic enzymes, thyroid axis proteins
  - All ToxCast Assay targets
- **Predict relative intrinsic susceptibility**
  - Pesticides
  - Endangered Species Act
  - Derivation of Aquatic Life Criteria
- **Predict chemical bioaccumulation across species**
  - Chemicals of concern: PFAS
- **Generate research hypotheses** Strobilurin fungicides
- **Prioritization strategies** Pharmaceuticals

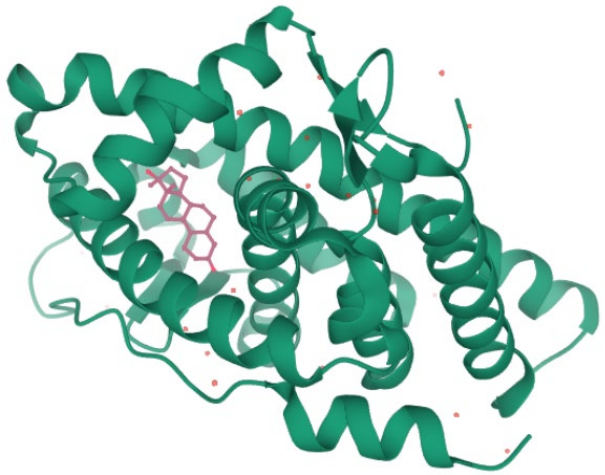
EPA
United States
Environmental Protection
Agency

Advances in Bioinformatics –
Future of SeqAPASS

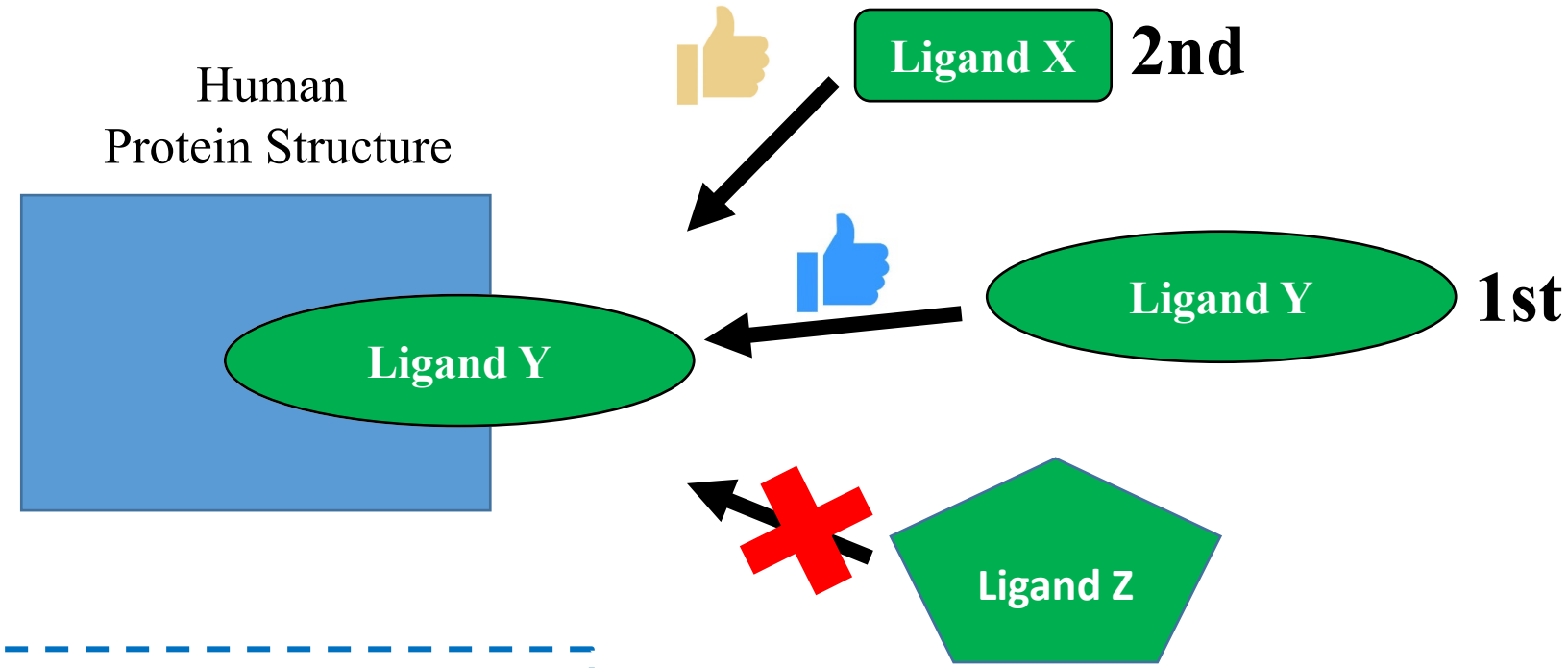
Always Look Several Steps Ahead

Advances in Drug Discovery/Development

(COVID-19 has led to advances)



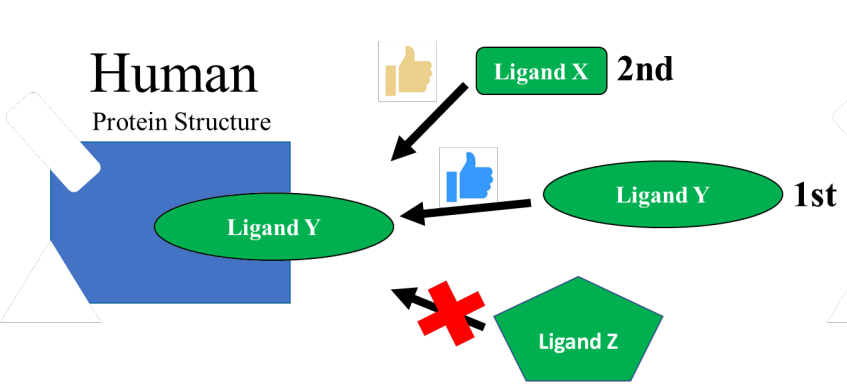
Structure derived from X-ray crystallography

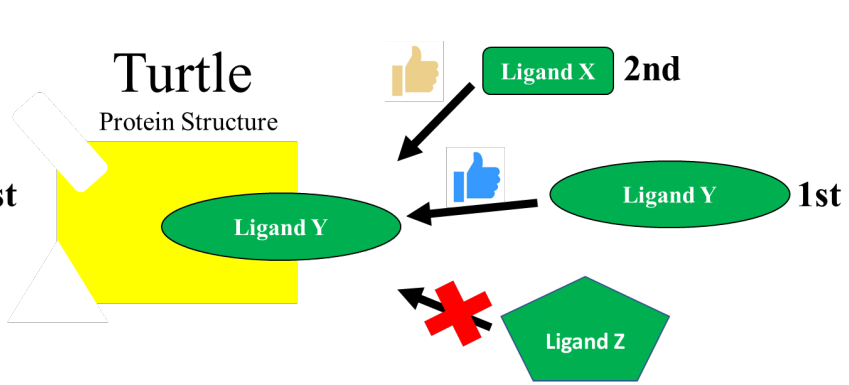


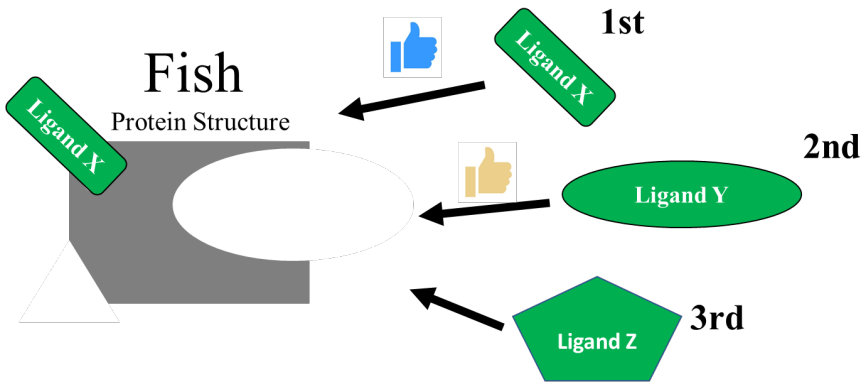
Bioinformatics Toolbox:

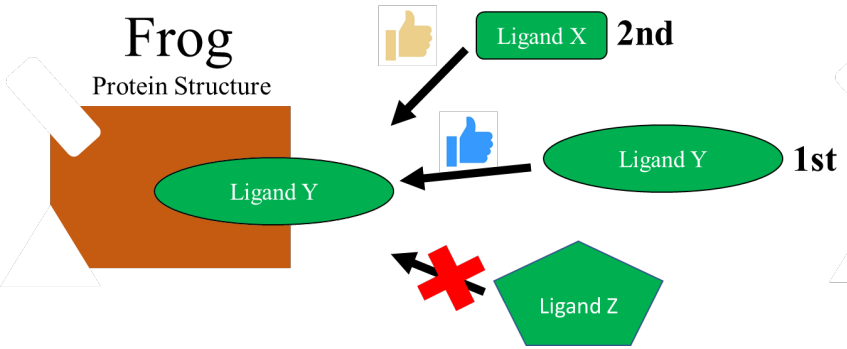
Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations

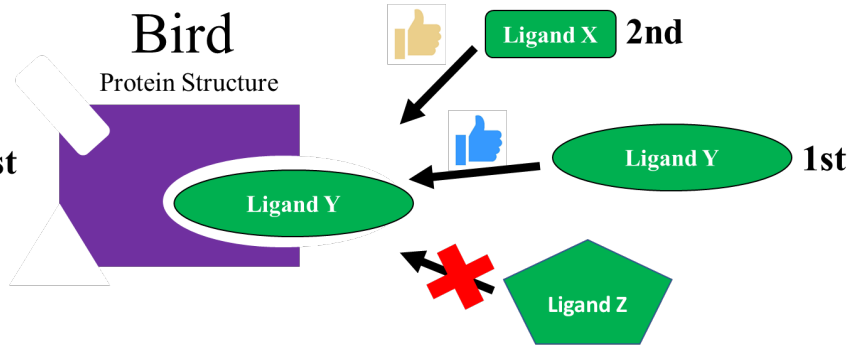
Application to Species Extrapolation

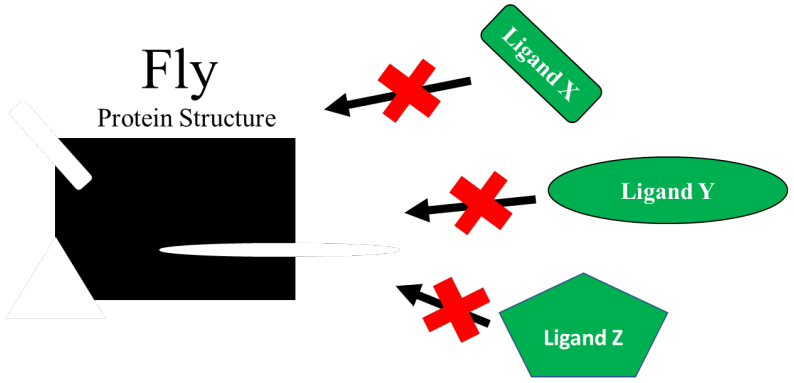








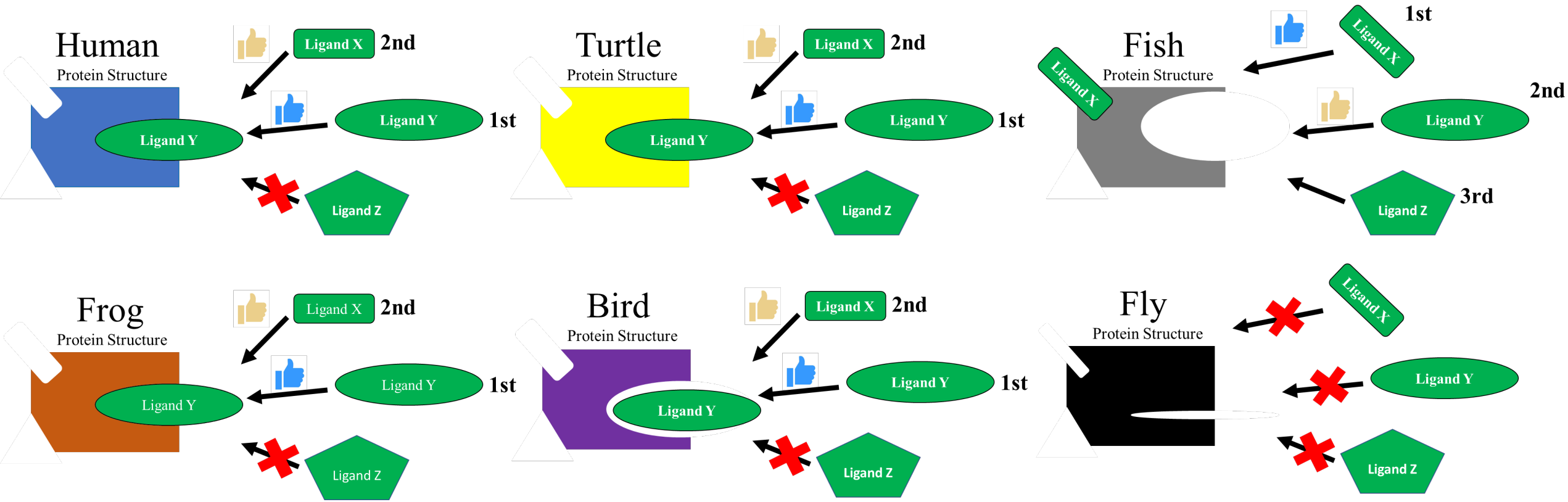




Bioinformatics Toolbox:

Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations

Application to Species Extrapolation



Bioinformatics Toolbox:

Molecular modeling

Molecular docking

Virtual screening

Molecular dynamic simulations

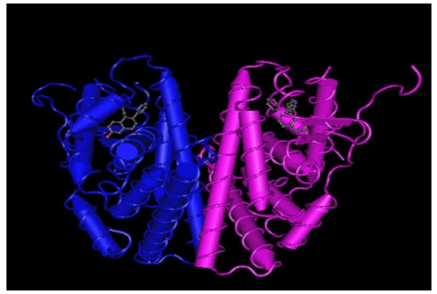
Thousands/Millions/Billions of Chemicals

How to begin:

Sequence

MTMTLHTKASGMALLHQIQGNELEPLNRPQLKIPLERPLGE
VYLDSSKPAVYNYPEGAAYEFNAAAAANAQVYGQTGLPYG
PGSEAAAFGSNGLGGFPPLNSVSPSPLMLLHPPPQLSPFLQ
PHGQQVPYYLENEPSGYTVREAGPPAFYRPNSDNRRQGGR
ERLASTNDKGSMAMESAKETRYCAVCNDYASGYHYGVWSC
EGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRLR
KCYEVGMMKGGIRKDRRGGRMLKHKRQRDDGEGRGEVG
SAGDMRAANLWPSPLMIKRSKKNSLALSLTADQMVSALLA
EPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKV
PGFVDLTLHDQVHLLECAWLEILMIGLVWRSMEHPGKLLFA
PNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEF
VCLKSIILLNSGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLM

Structure



SeqAPASS Results from Level 1

Query Sequence FASTA + FASTA from 100s of Aligned Sequences Across Taxa

>NP_001434.1 Protein X [Homo sapiens]
MSFSGKYQLQSQENFEAFMKAIGLPEELIQKGKDI
KGVSEIVQNGKHFKFTITAGSKVIQNEFTVGEECE
LETMTGEKVKTVVQLEGDNKLVTTFKNIKSVTELN
GDIITNTMTLGDIVFKRISKRI

>NP_787011.1 Protein X [Bos taurus]
MNFSGKYQVQTQENYEAFMKAVGMPDDIIQKGK
DIKGVSEIVQNGKHFKFIITAGSKVIQNEFTLGEECE
MEFMTGEKIKAVVQQEGDNKLVTTFKGIKSVTEFN
GDTVTSTMTKGDVVFKRVSKRI

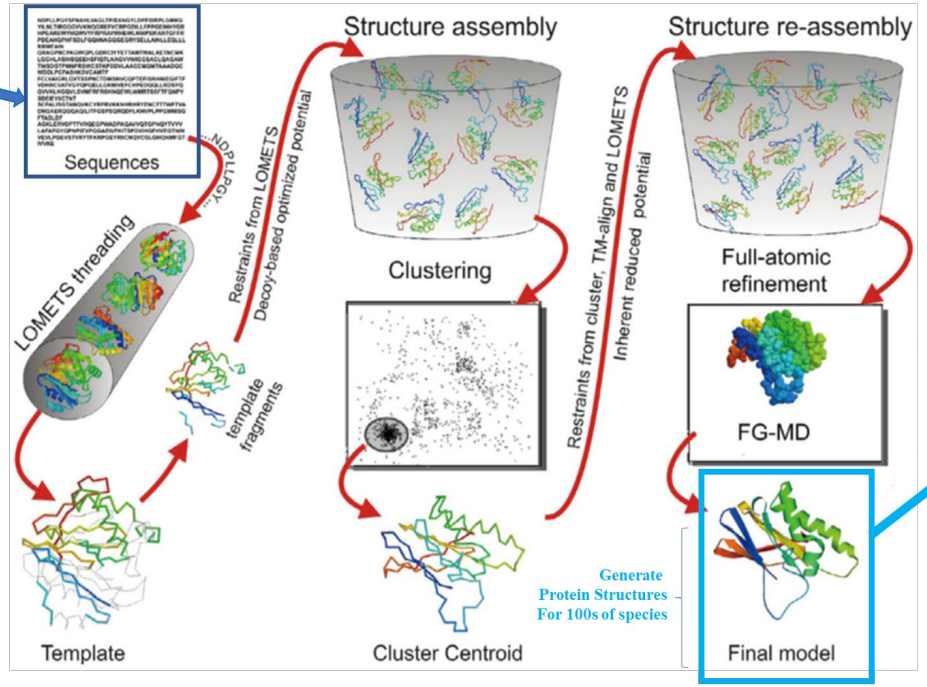
>KFQ76585.1 Protein X [Phoenicopterus ruber ruber]
MSFTGKYELQSQENFEPFMKALGLPDDQIQKGKD
IKSISEIVQDGKKFKVTVTTGSKVMQNEFTIGEECD
IEMLTGEKVKAVVQMEGNNRLVANLKGLKSVTEL
NGDIITHTMTMGDLTYKRISKRI

>NP_001116883.1 Protein X [Xenopus tropicalis]
MAFAGKYELVHQENFETFMKAIGLSDELIQKGKDV
KSVTEIQQNGKHFIVTVTTGSKVLRNEFTIGEEAE
LETPTGEKVKSVVKLEGDNKLVVQLKAITSTTELSG
DTITHVLTLNNLVFKRVSKRV

100s of FASTA

Iterative Threading ASSEmbly Refinement

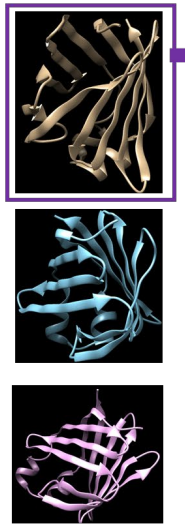
Develop Models for 100s of Species Based on Aligned Sequences (I-TASSER; https://zhanglab.ccmb.med.umich.edu/I-TASSER/)



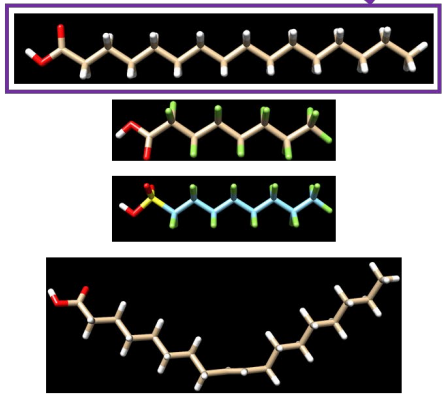
UCSF Chimera

DockPrep Structures and Minimize Ligands

Protein Structure Models From 100s of Species

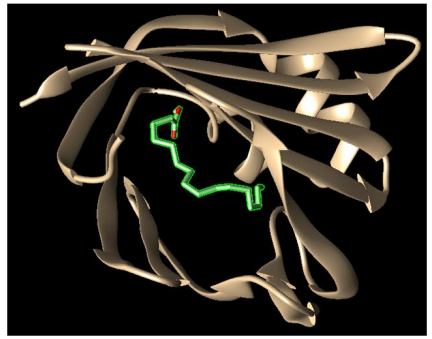


Ligands of Interest for Docking



AutoDock Vina

Dock Multiple Ligands to Protein Structures



Collect Predicted Binding Affinity

S	Score	RMSD lb.	RMSD u.b.	HBonds (all)	HBond Ligand Atoms	HBond Receptor Atoms
V	-7.1	0.0	0.0	0	0	0
V	-7.0	1.212	2.436	0	0	0
V	-7.0	2.148	6.837	1	1	1
V	-6.9	1.128	2.04	0	0	0
V	-6.9	4.472	7.133	0	0	0
V	-6.7	3.27	7.552	0	0	0
V	-6.7	2.637	3.461	2	2	2
V	-6.6	1.572	3.516	0	0	0
V	-6.6	1.725	3.368	0	0	0

Chimera Model #3.1

```
REMARK VINA RESULT:      -7.1      0.000      0.000
REMARK  15 active torsions:
REMARK  status: ('A' for Active; 'I' for Inactive)
REMARK    1  A    between atoms: C2_2  and  C3_3
REMARK    2  A    between atoms: C3_3  and  C4_4
REMARK    3  A    between atoms: C4_4  and  C5_5
REMARK    4  A    between atoms: C5_5  and  C6_6
REMARK    5  A    between atoms: C6_6  and  C7_7
REMARK    6  A    between atoms: C7_7  and  C8_8
REMARK    7  A    between atoms: C8_8  and  C9_9
REMARK    8  A    between atoms: C10_10  and  C9_9
REMARK    9  A    between atoms: C10_10  and  C11_11
REMARK   10  A    between atoms: C11_11  and  C12_12
REMARK   11  A    between atoms: C12_12  and  C13_13
REMARK   12  A    between atoms: C13_13  and  C14_14
REMARK   13  A    between atoms: C14_14  and  C15_15
REMARK   14  A    between atoms: C15_15  and  C16_16
REMARK   15  A    between atoms: C16_16  and  O2_18
```

Graphic Modified from Zhang et al., 2019 I-TASSER gateway: A protein structure and function prediction server powered by XSEDE Figure 1

Predicting Binding Affinity

Government

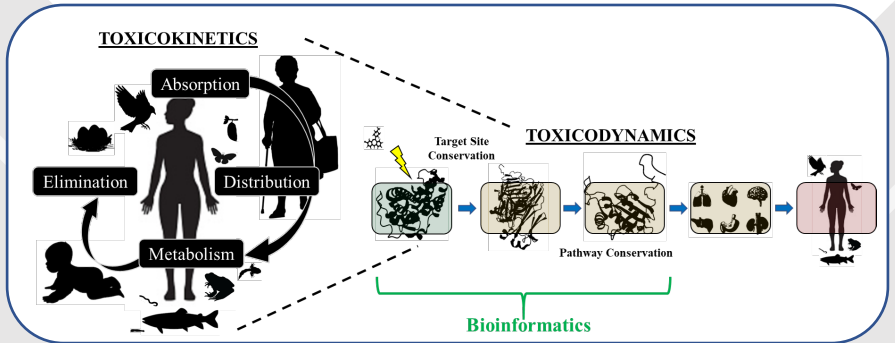
Industry

Consortium to Advance Cross Species Extrapolation in Regulation

Steering Committee:
Carlie LaLone (US EPA)
Geoff Hodges (Unilever)
Nil Basu (McGill U)
Steve Edwards (RTI)
Fiona Sewell (NC3Rs)
Michelle Embry (HESI)
Patience Browne (OECD)

1. Define the taxonomic domain of applicability
2. Define the global regulatory landscape/need
3. Develop a bioinformatics toolbox
4. Communicate a shared scientific vision

Interested in Learning more or Joining: Contact LaLone.Carlie@epa.gov or Geoff.Hodges@unilever.com



Academia

NGO

Acknowledgements

U.S. EPA, ORD

Marissa Jensen (University of Minnesota Duluth)
Sally Mayasich (University of Wisconsin)
Donovan Blatz (ORISE)
Monique Hazemi (ORISE)
Sara Vliet (US EPA)
Jon Doering (U of Lethbridge)
Colin Finnegan (Iowa State University)

GDIT

Thomas Transue
Cody Simmons
Audrey Wilkinson
Wilson Menendez

SeqAPASS v6.0 (Released Sept. 2021)

LaLone.Carlie@epa.gov

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