United States Environmental Protection Agency

Bioinformatics is for Everyone: Applications To Challenges in Ecotoxicology

Carlie A. LaLone, Ph.D.
Research Bioinformaticist

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

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the US EPA

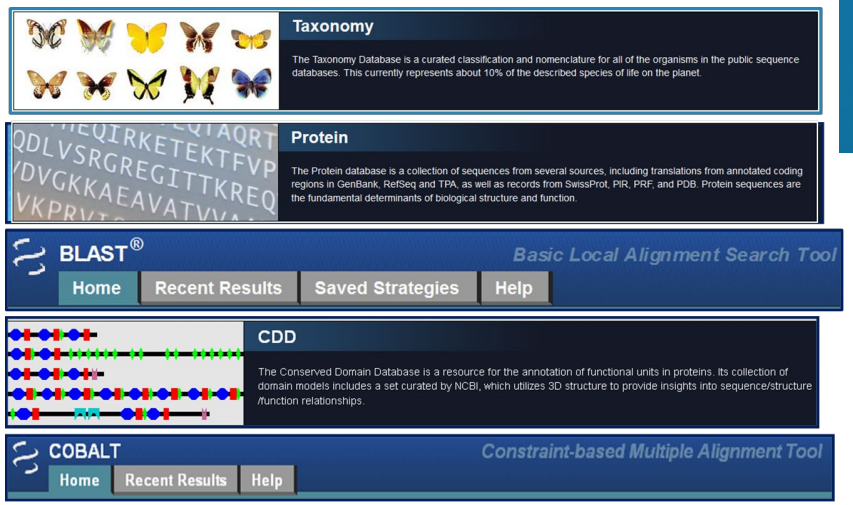
May 5th, 2022

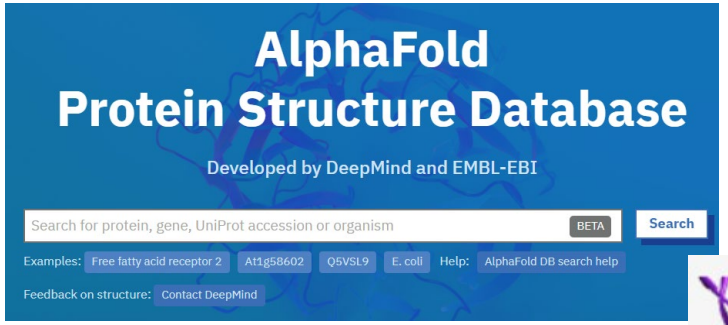
Bioinformatics

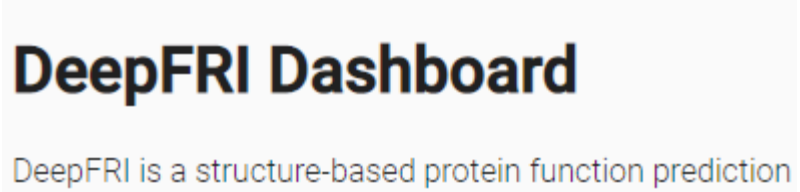
- Combines mathematics, information science, and biology to answer biological questions
- Developing methodology and analysis tools to explore large volumes of biological data
  - Query, extract, store, organize, systematize, annotate, visualize, mine, and interpret complex data
    - Usually pertains to DNA and amino acid sequences

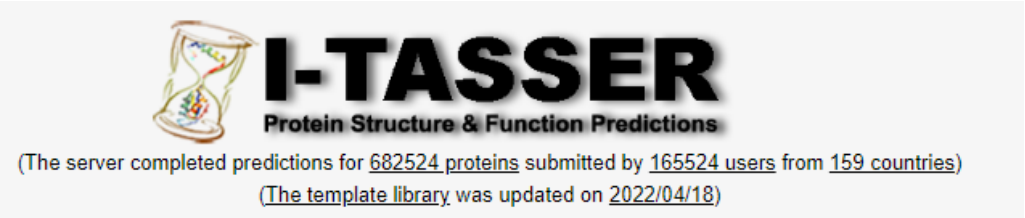
Let the computers do the work

EPA United States Environmental Protection Agency





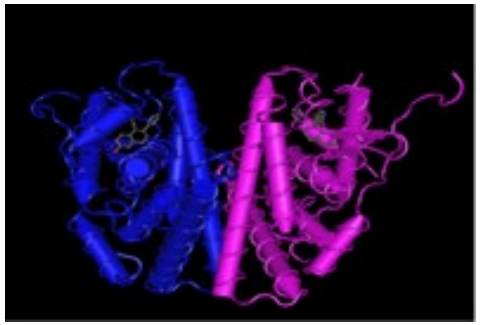




Sequence

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

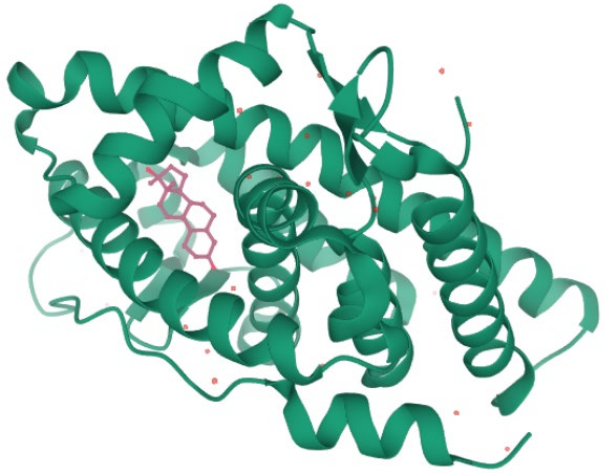
Structure



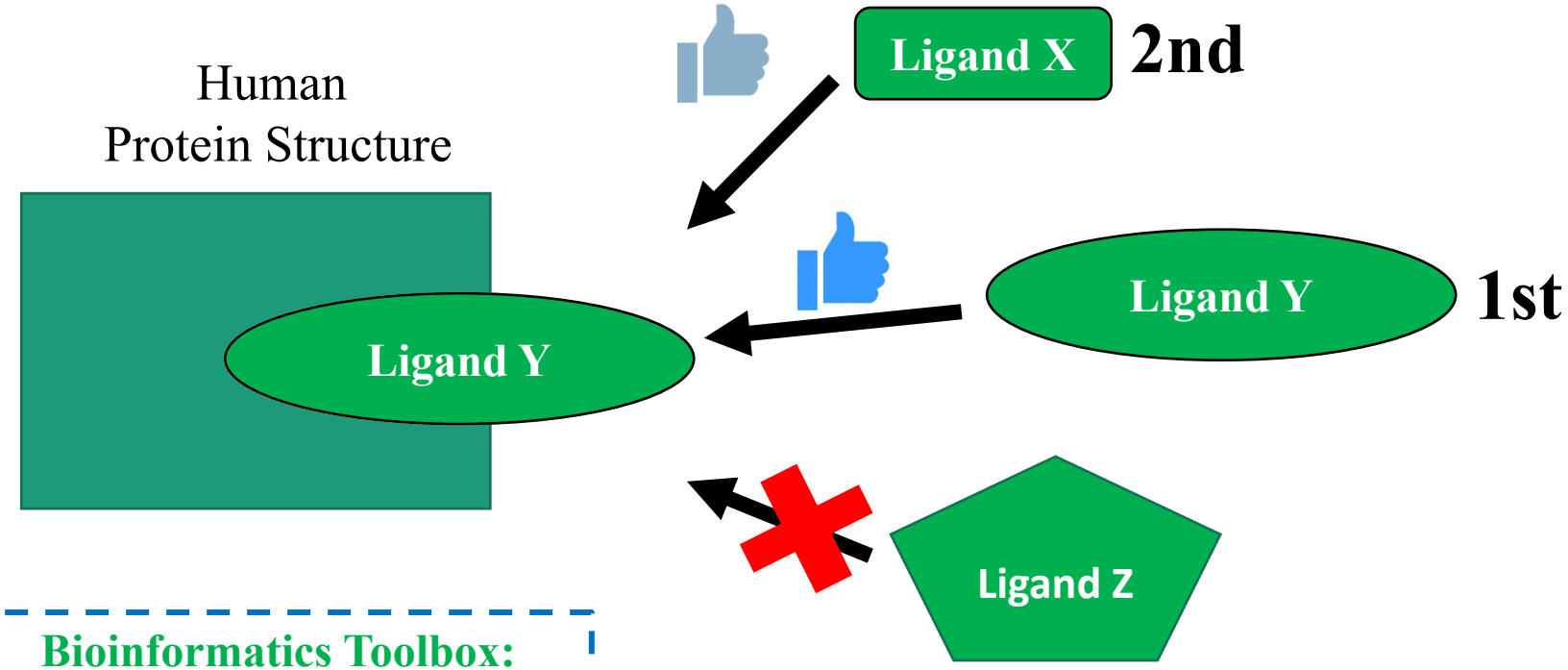
Function

Bioinformatics

Advances in Drug Discovery/Development



Structure derived from X-ray crystallography



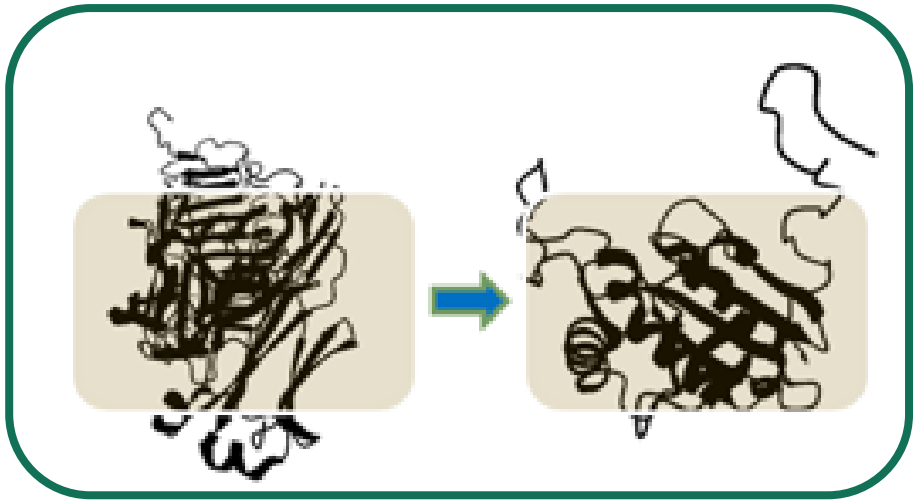
Bioinformatics Toolbox:

Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations
Functional prediction

Numerous bioinformatics approaches

- My interest and focus has been chemical-protein and protein-protein Interactions
  - Utility for cross species extrapolation

Chemical-Protein



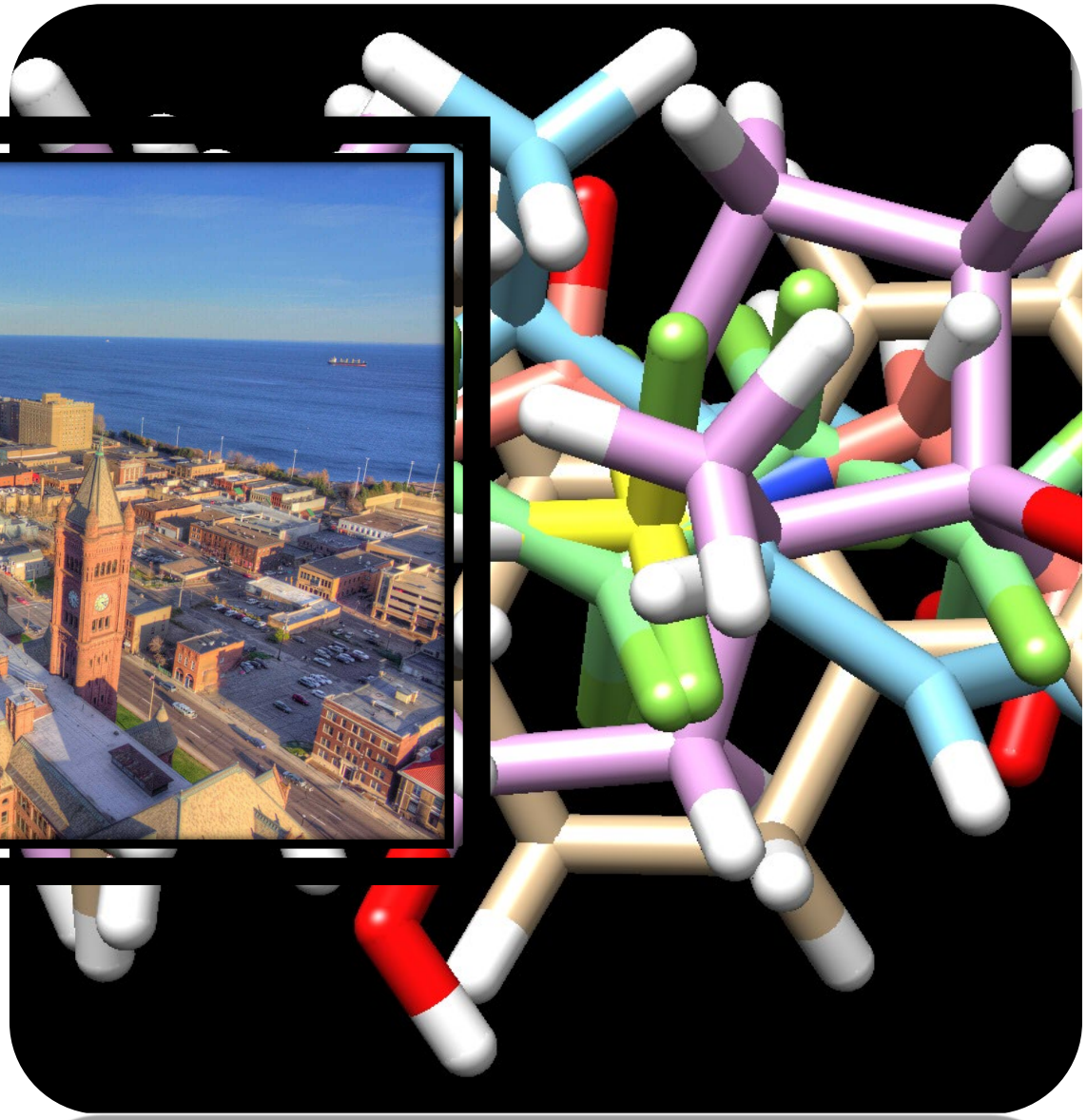
Protein-Protein

United States Environmental Protection Agency

Overview Questions

- What species do we rely on for toxicity testing and why?
- Why consider predictive and computational approaches?
- How can bioinformatics help for chemical safety evaluations?
- What tools are available now and moving forward?
- How do we incorporate bioinformatics in decision making?

Chemicals make up the world around us – necessary for our modern society



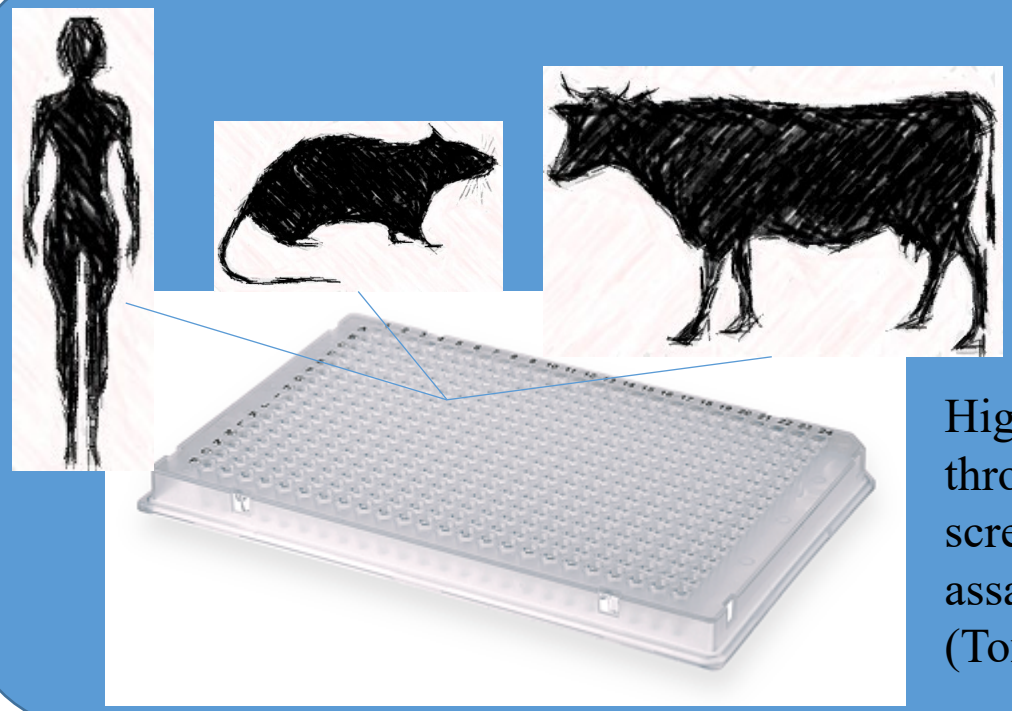
Toxicity Testing to Understand Chemical Safety

- **US EPA Examples:**
- *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

Historic whole organism toxicity testing

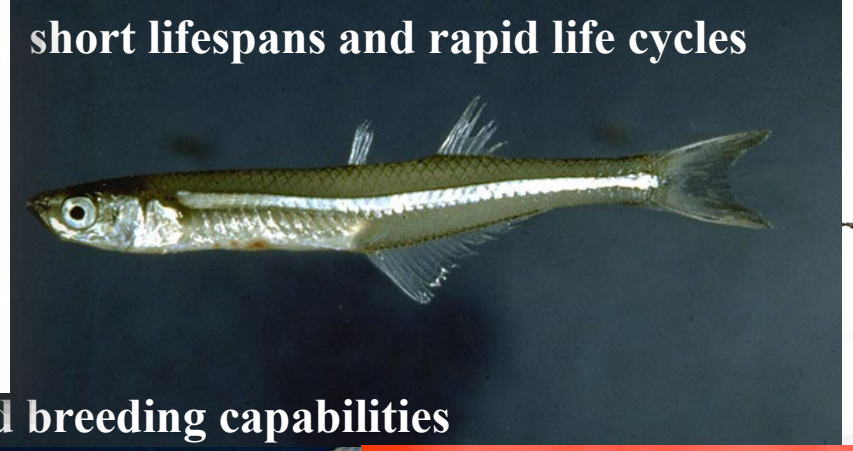


High-throughput screening assays (ToxCast)

Define the taxonomic domain of applicability in AOP development

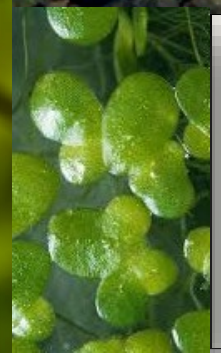
Use of model organisms as surrogates representing the diversity of species in the environment

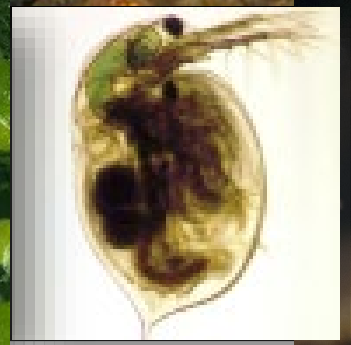

cheap and readily available


short lifespans and rapid life cycles


easy maintenance and good breeding capabilities


requires least space and time-consuming care






ability to control diet and surroundings

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:

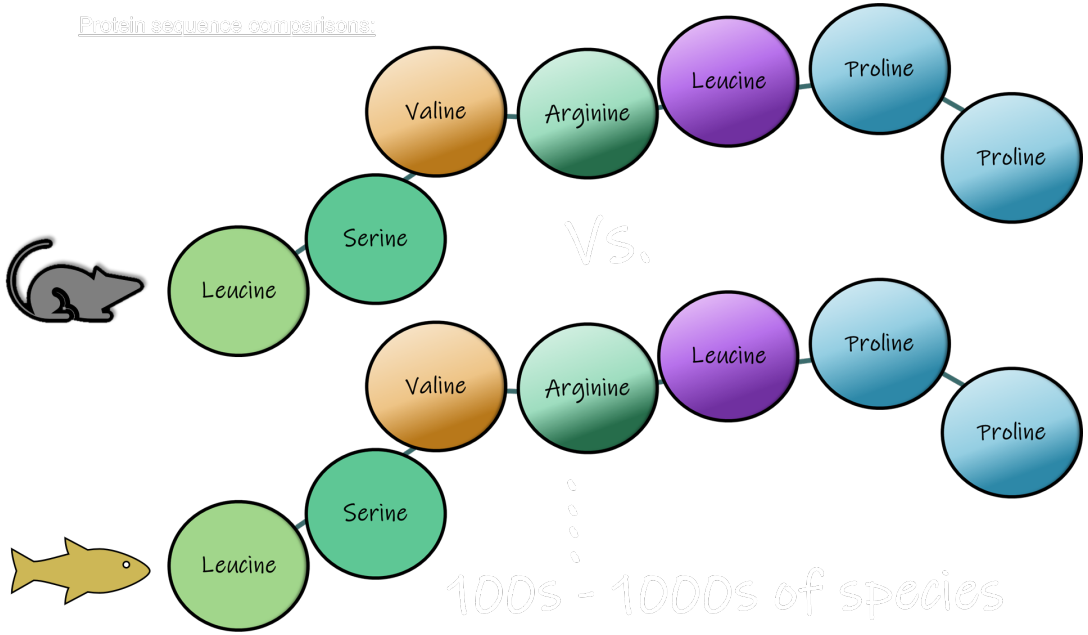
- Limited or no toxicological data for the animal or plant species of interest – reliance on surrogate (model organisms)
  - Impractical to generate new data for all species
- Testing resources are limited
  - International interest to reduce animal use
  - Ever-increasing demand to evaluate more chemicals 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

Bioinformatics for species extrapolation?


Environ. Sci. Technol. 2008, 42, 5807–5813

Evolutionary Conservation of Human Drug Targets in Organisms used for Environmental Risk Assessments

LINA GUNNARSSON,[†] ALEXANDRA JAUHIAINEN,[‡,⊥] ERIK KRISTIANSSON,[†,‡,§] OLLE NERMAN,[‡,⊥] AND D. G. JOAKIM LARSSON*,[†]



- Begin Simple and Advance as the Science Advances
- Consider sequence and structural attributes to understand protein conservation across species

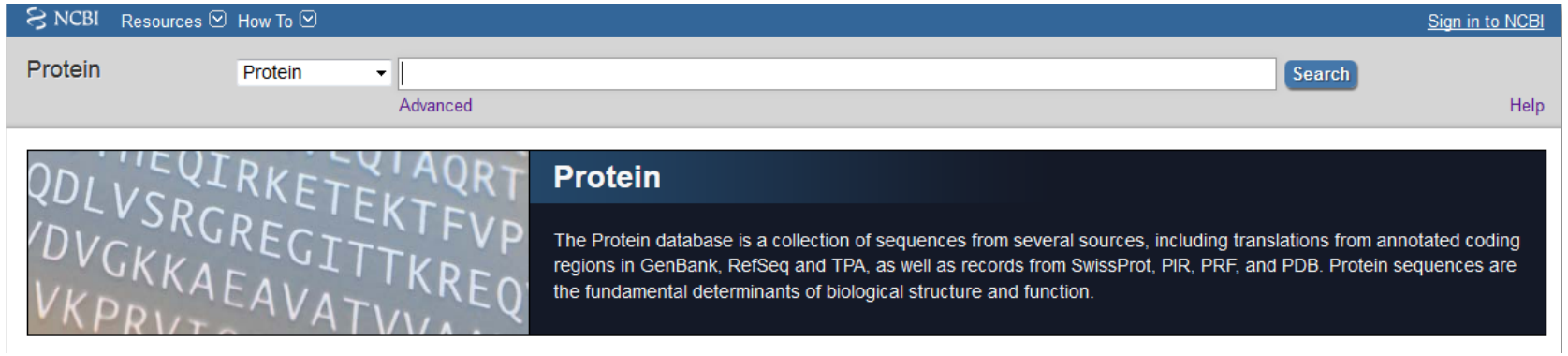
Where could we begin in understanding species similarities and differences?

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

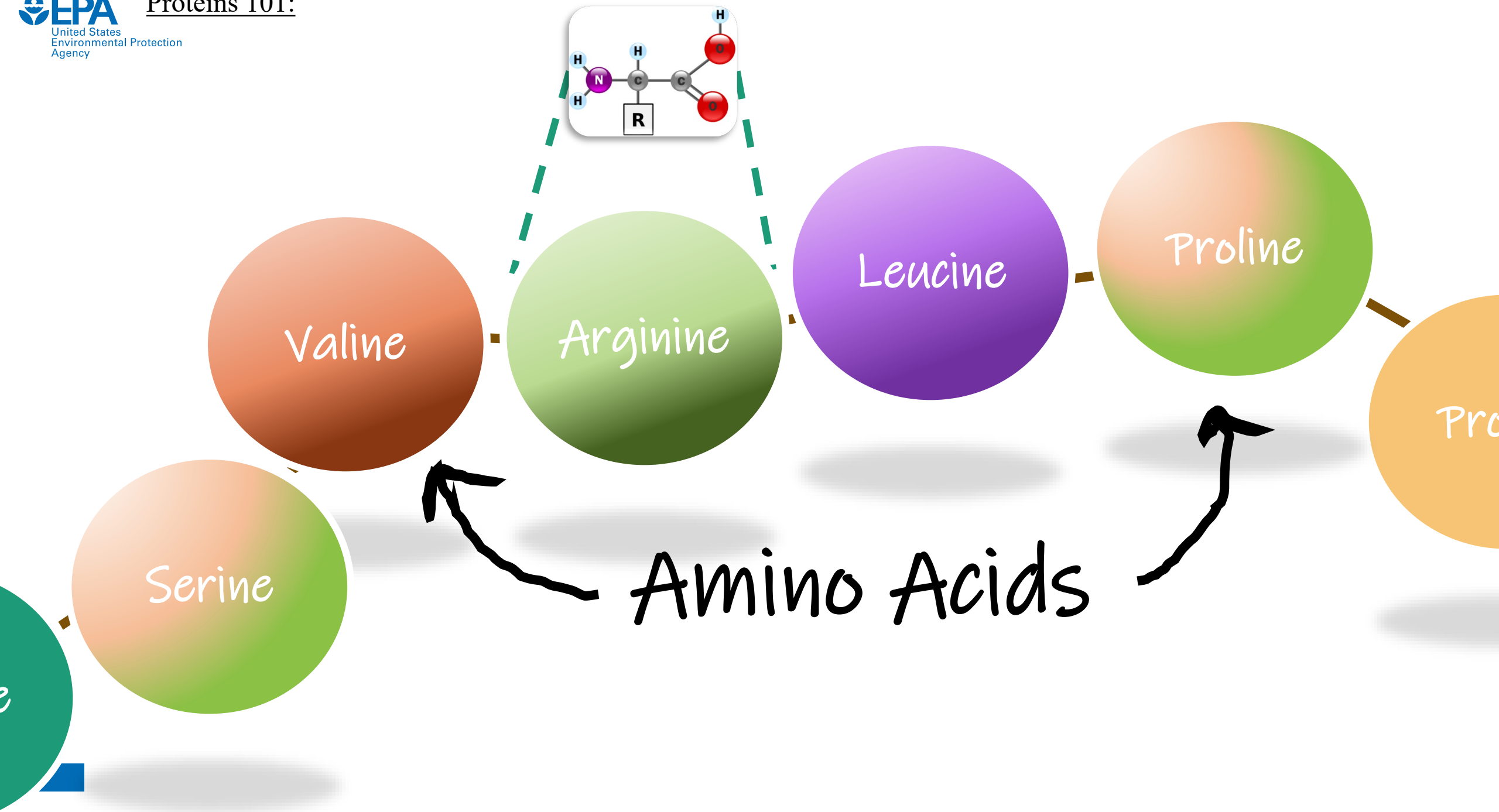
Sequence and structural data: New tools and technologies have emerged

- Improved sequencing technologies
- Large databases of sequence data

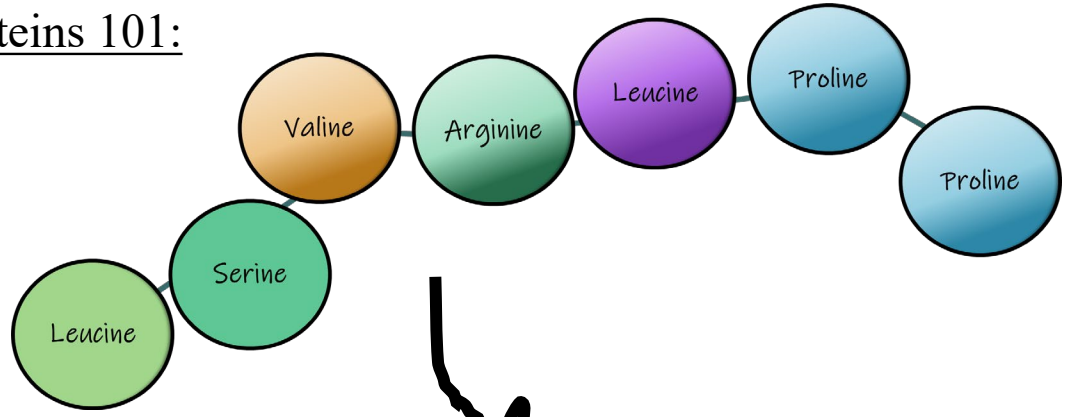
NCBI: 224,211,842 Proteins representing 117,030 Organisms



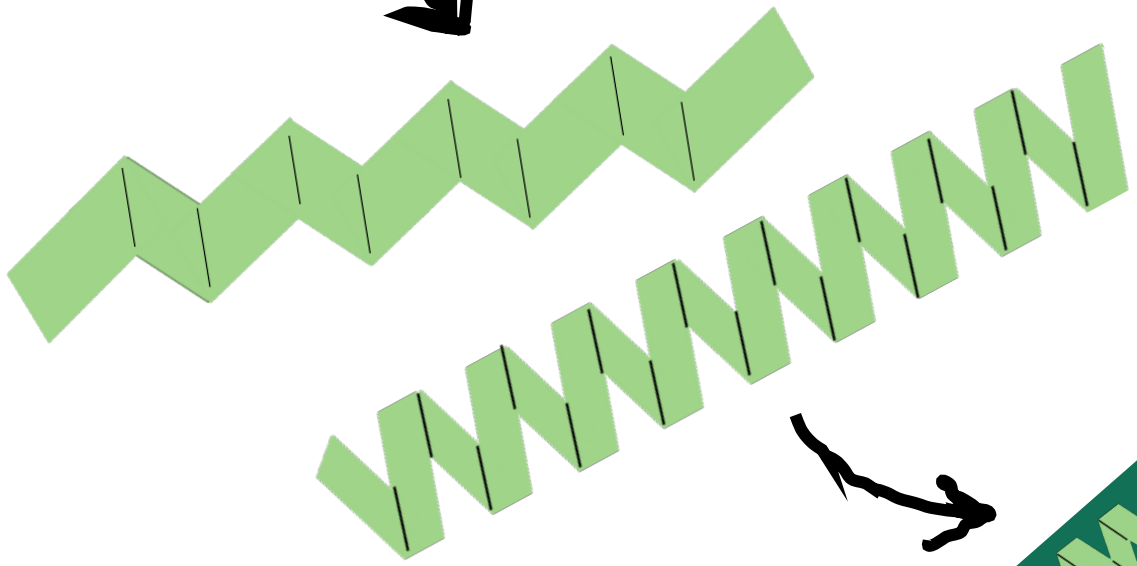
Proteins 101:



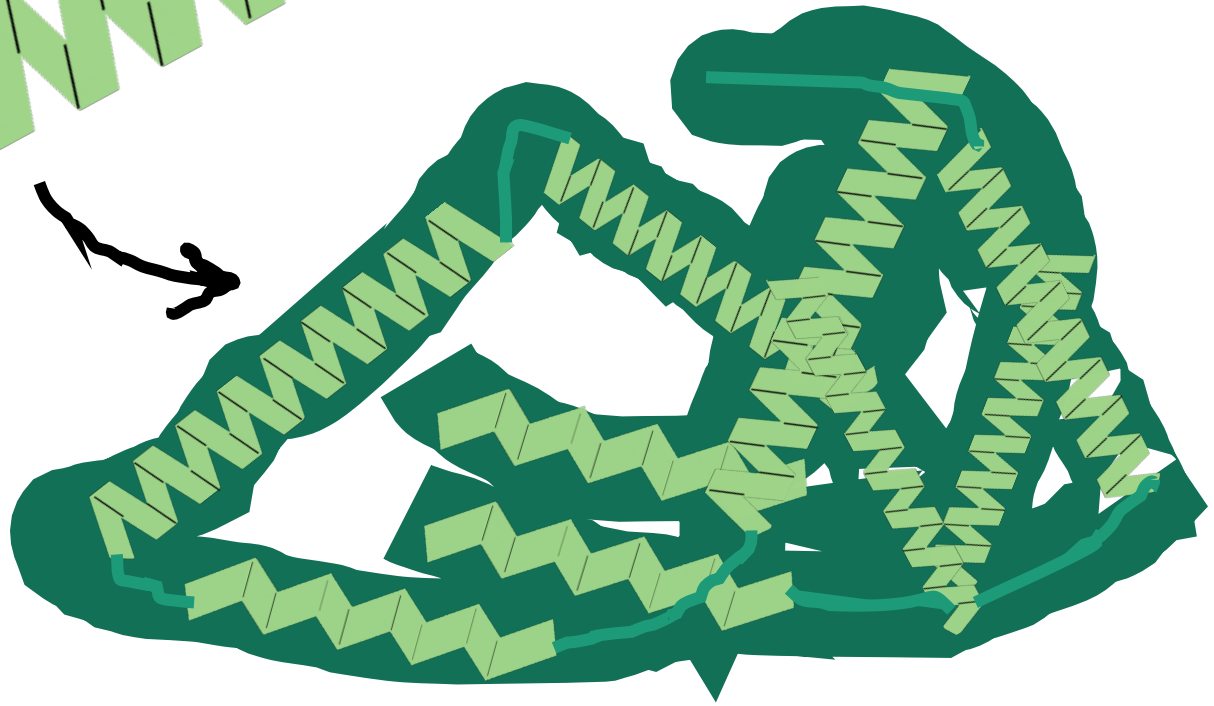
Proteins 101:



Primary amino acid sequence

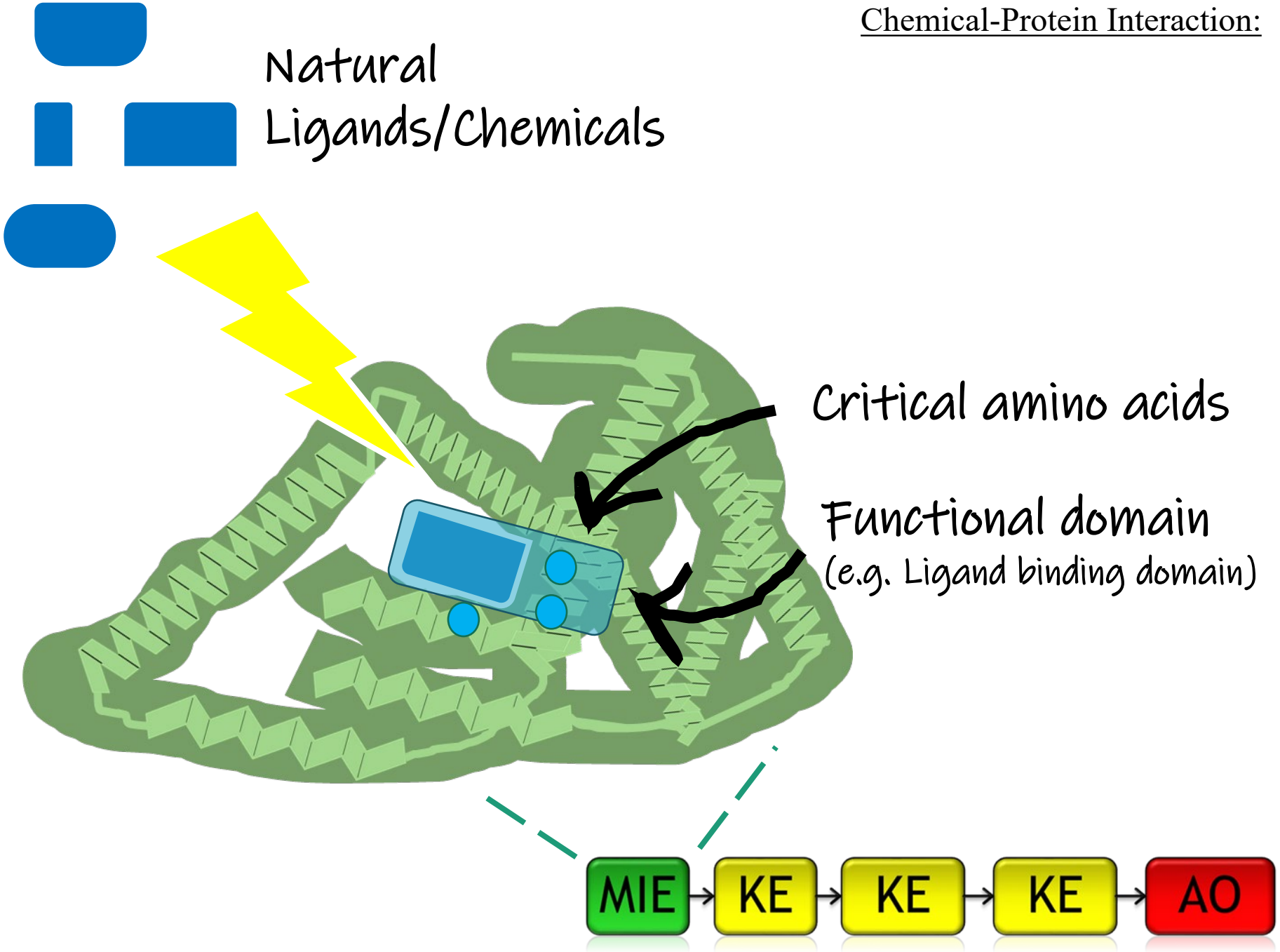


Secondary Structure



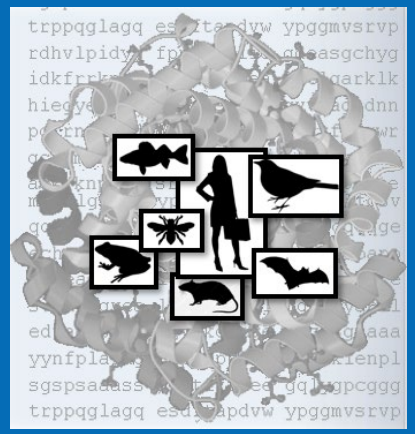
Tertiary Structure

Chemical-Protein Interaction:



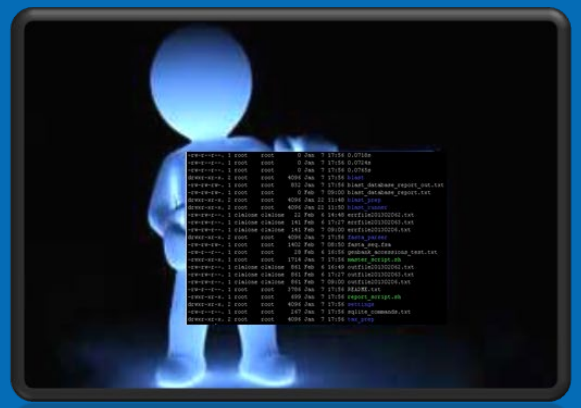
Similarity across species at the molecular level

EPA
United States
Environmental Protection
Agency



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

Sequence Alignment to Predict Across Species Susceptibility (SeqAPASS)



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

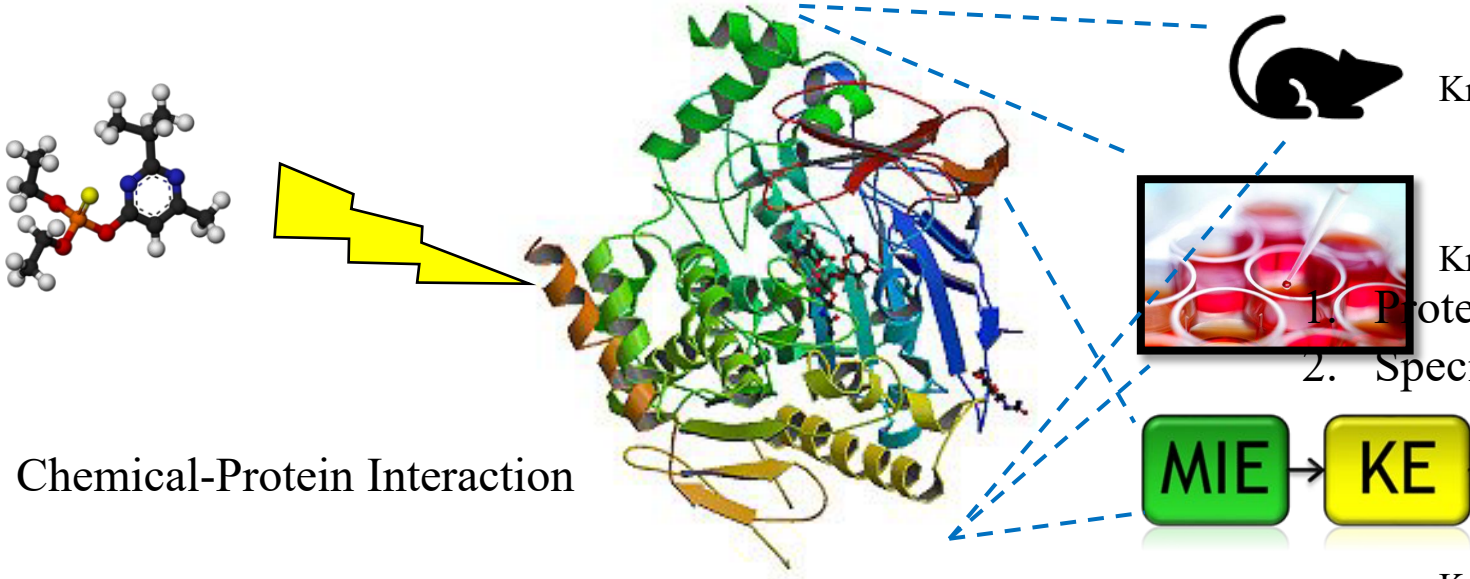
TOXICOLOGICAL SCIENCES, 153(2), 2016, 228–245

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*

What information is required for a SeqAPASS query?



Chemical-Protein Interaction

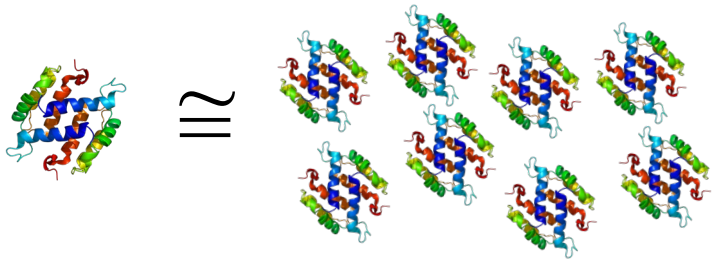
Knowledge of a sensitive or targeted species

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

Animation by: Miguel Moravec (EPA CSS) & Andrew Patterson

SeqAPASS

United States Environmental Protection Agency

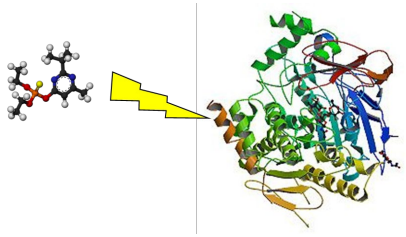
SeqAPASS

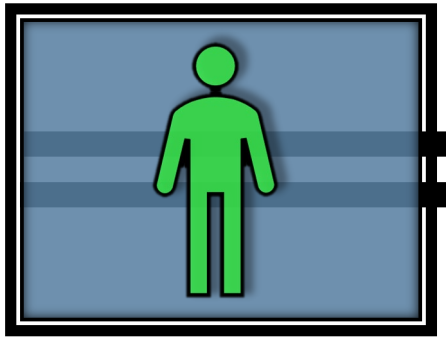
United States Environmental Protection Agency

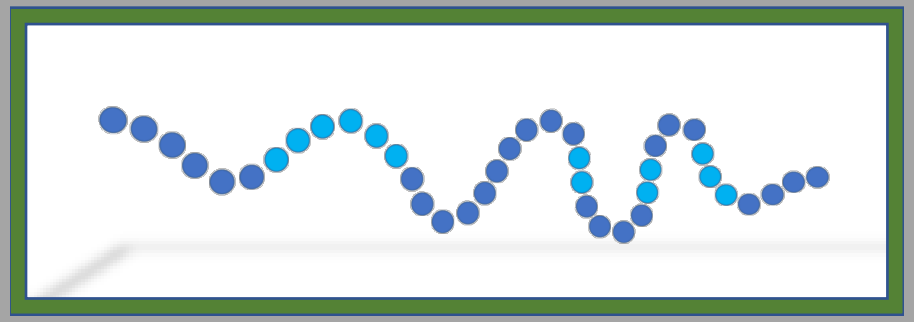
SeqAPASS Level 1

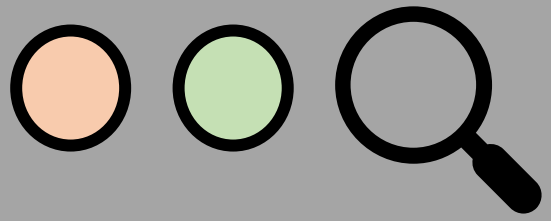
SeqAPASS Level 1

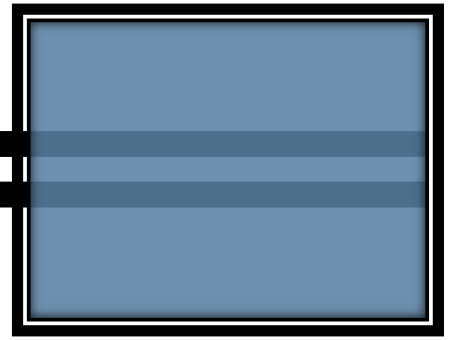
Human Protein Target

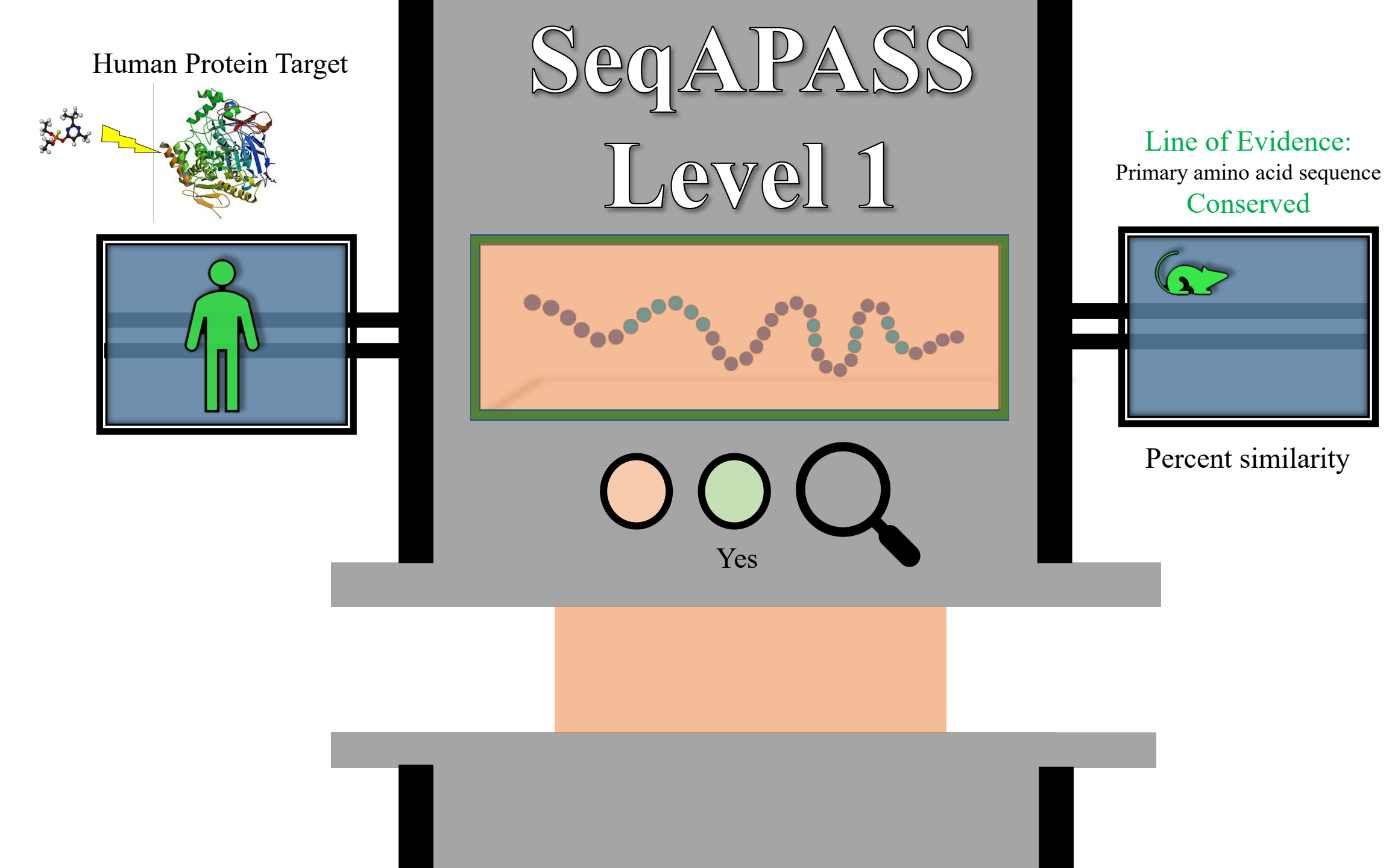


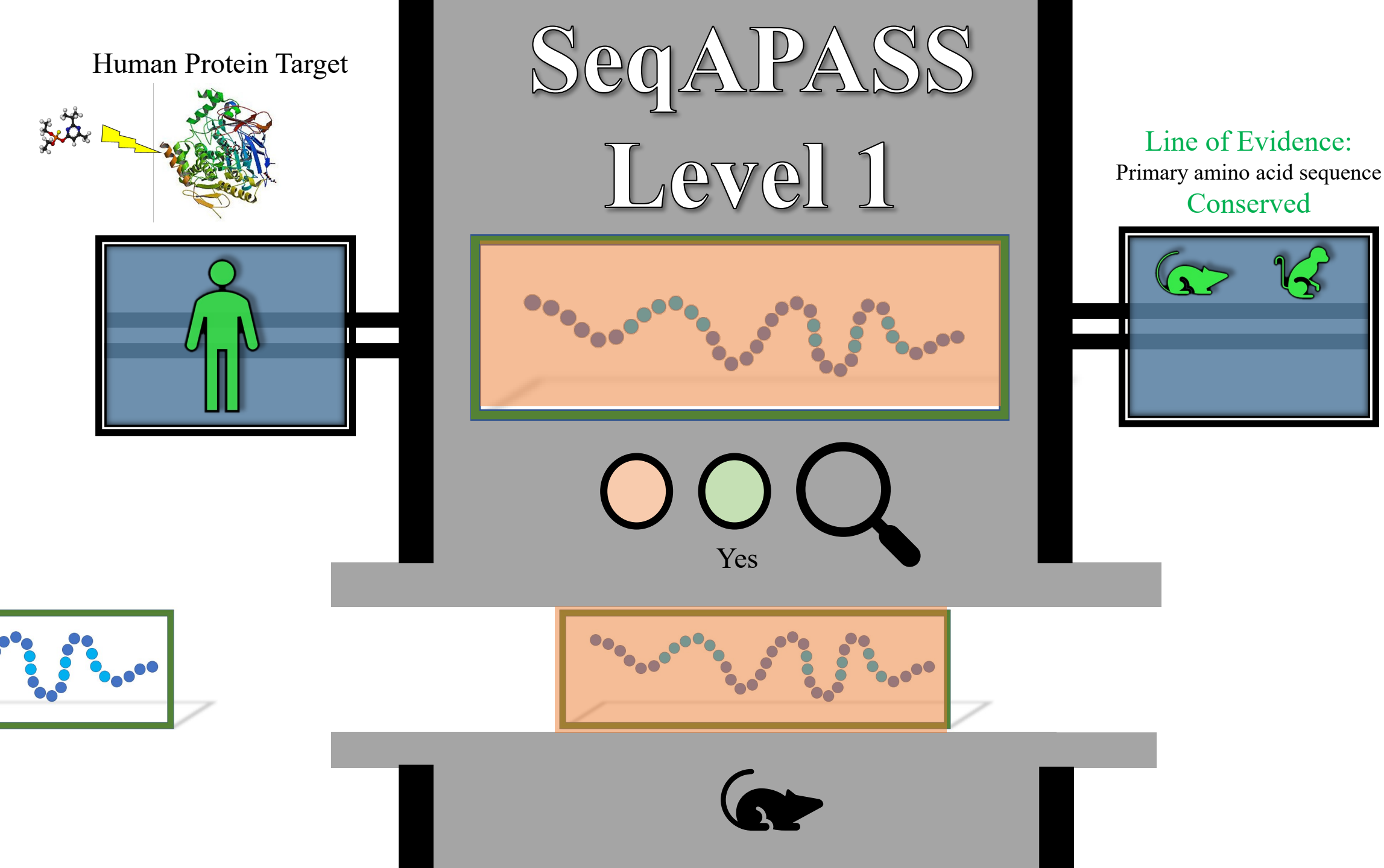


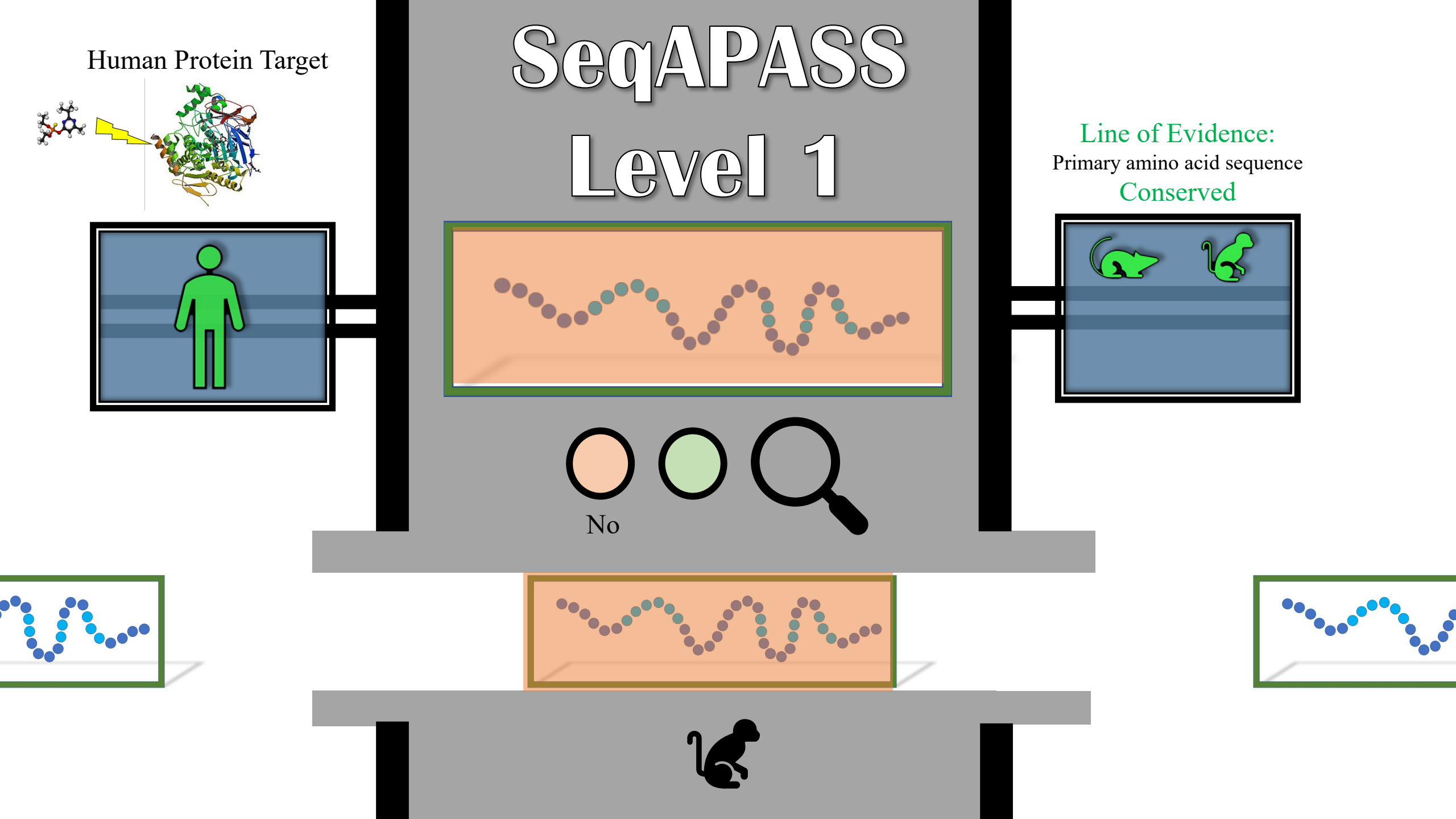




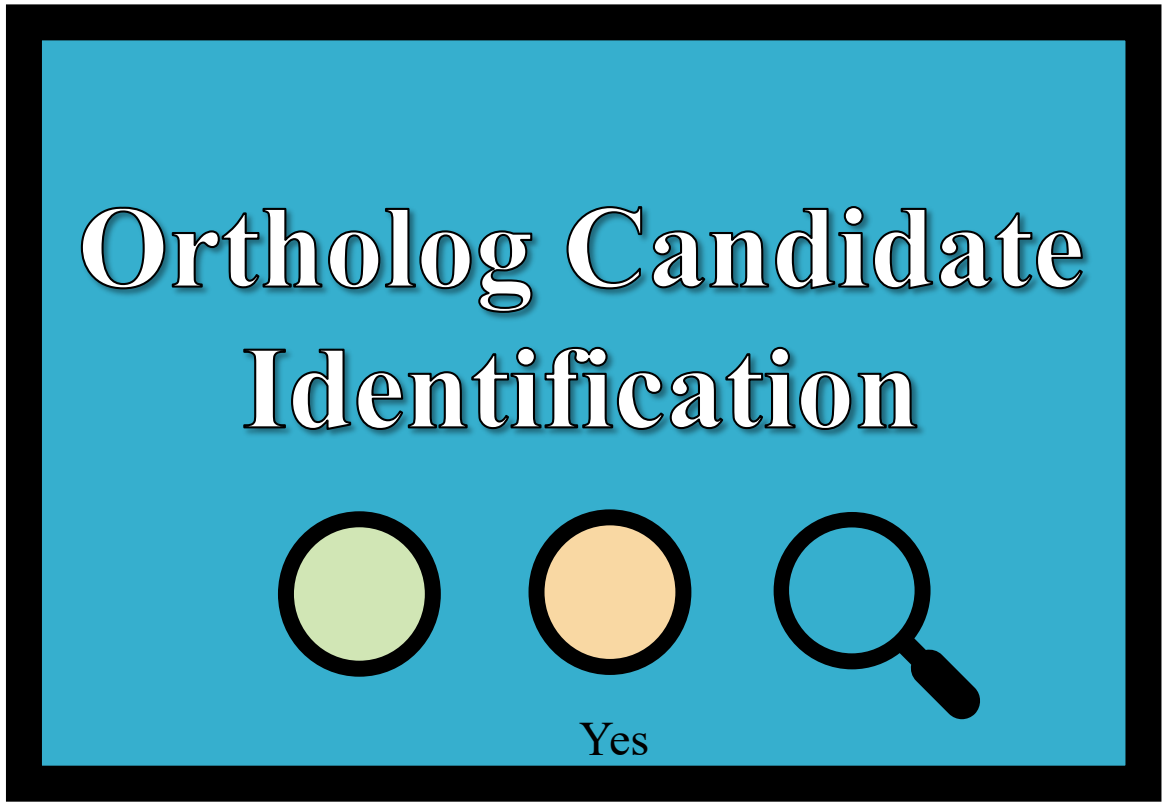


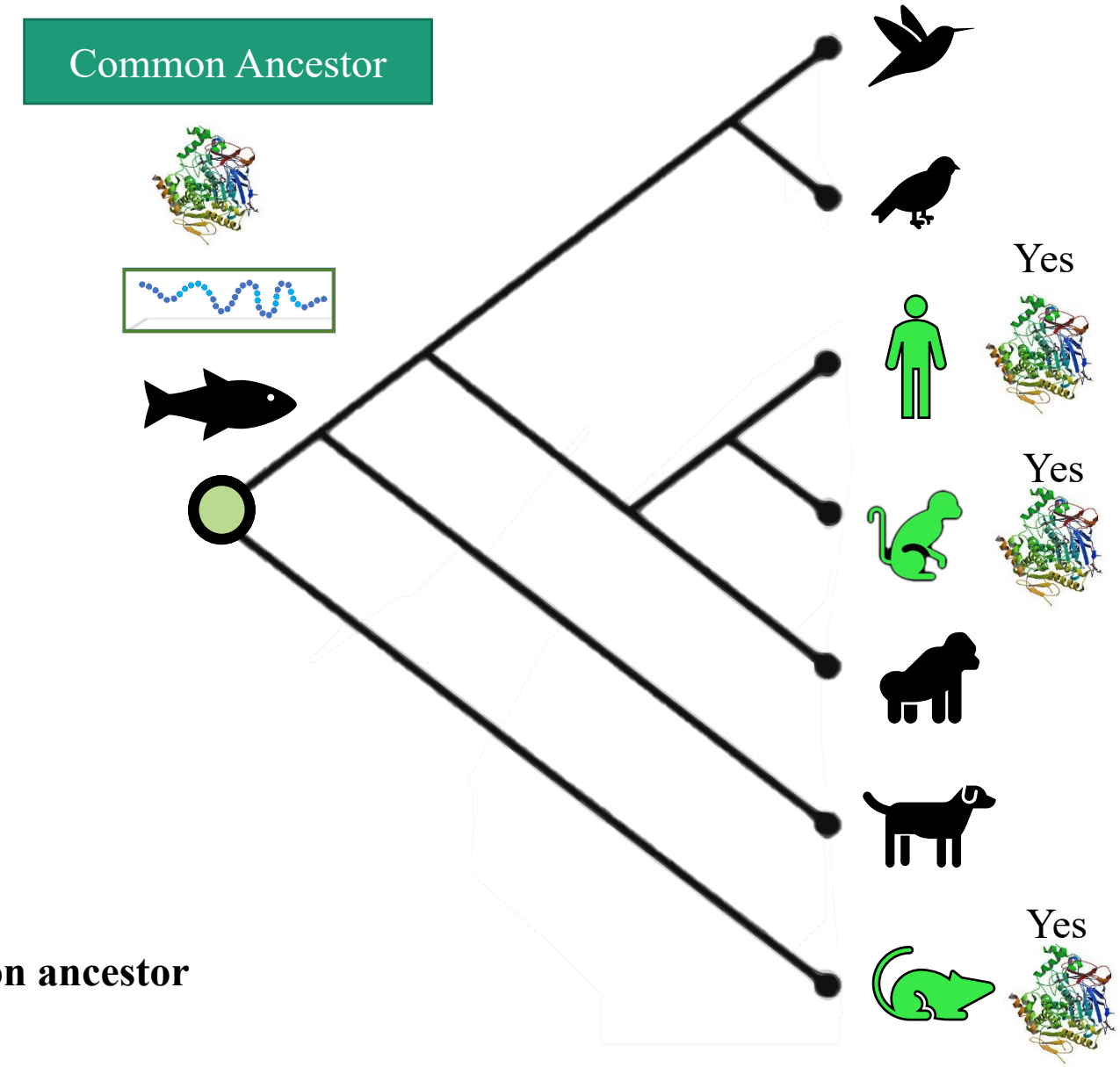

Human Protein Target
SeqAPASS
Level 1
Line of Evidence:
Primary amino acid sequence
Conserved
Percent similarity
Yes


Human Protein Target
SeqAPASS
Level 1
Line of Evidence:
Primary amino acid sequence
Conserved
Yes


Human Protein Target
SeqAPASS
Level 1
Line of Evidence:
Primary amino acid sequence
Conserved
No

SeqAPASS Level 1

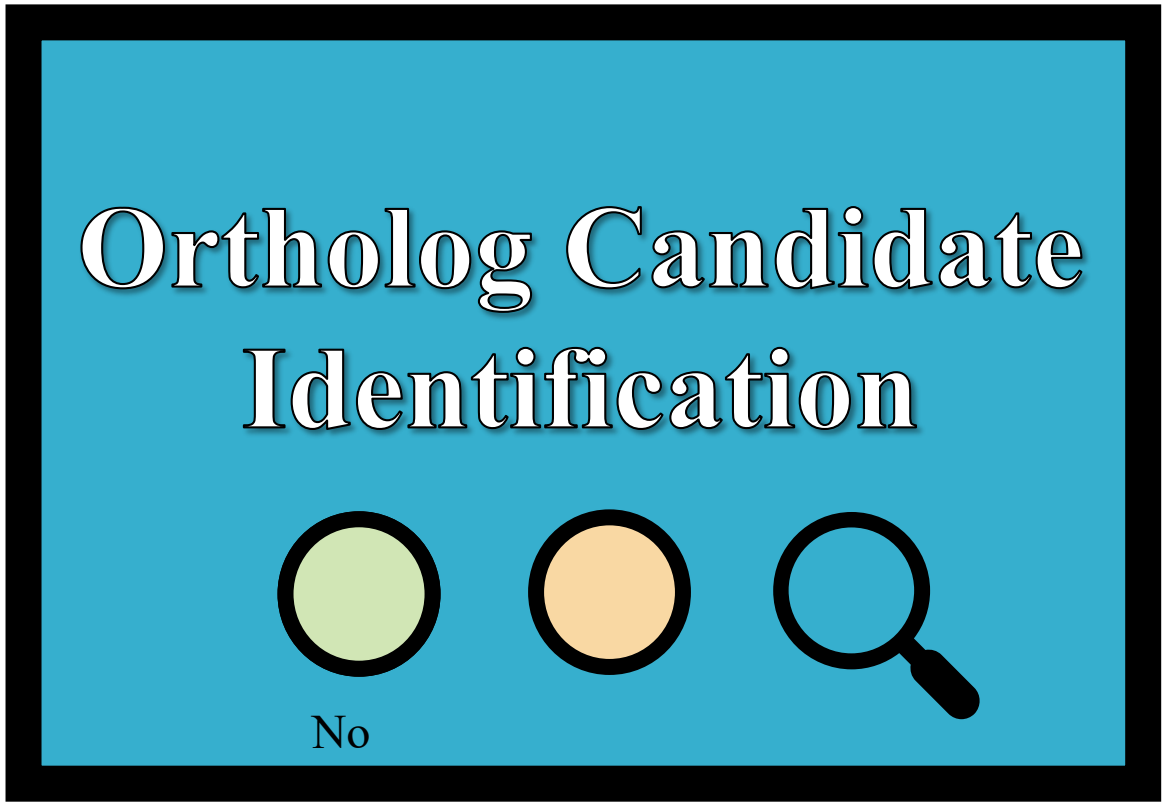


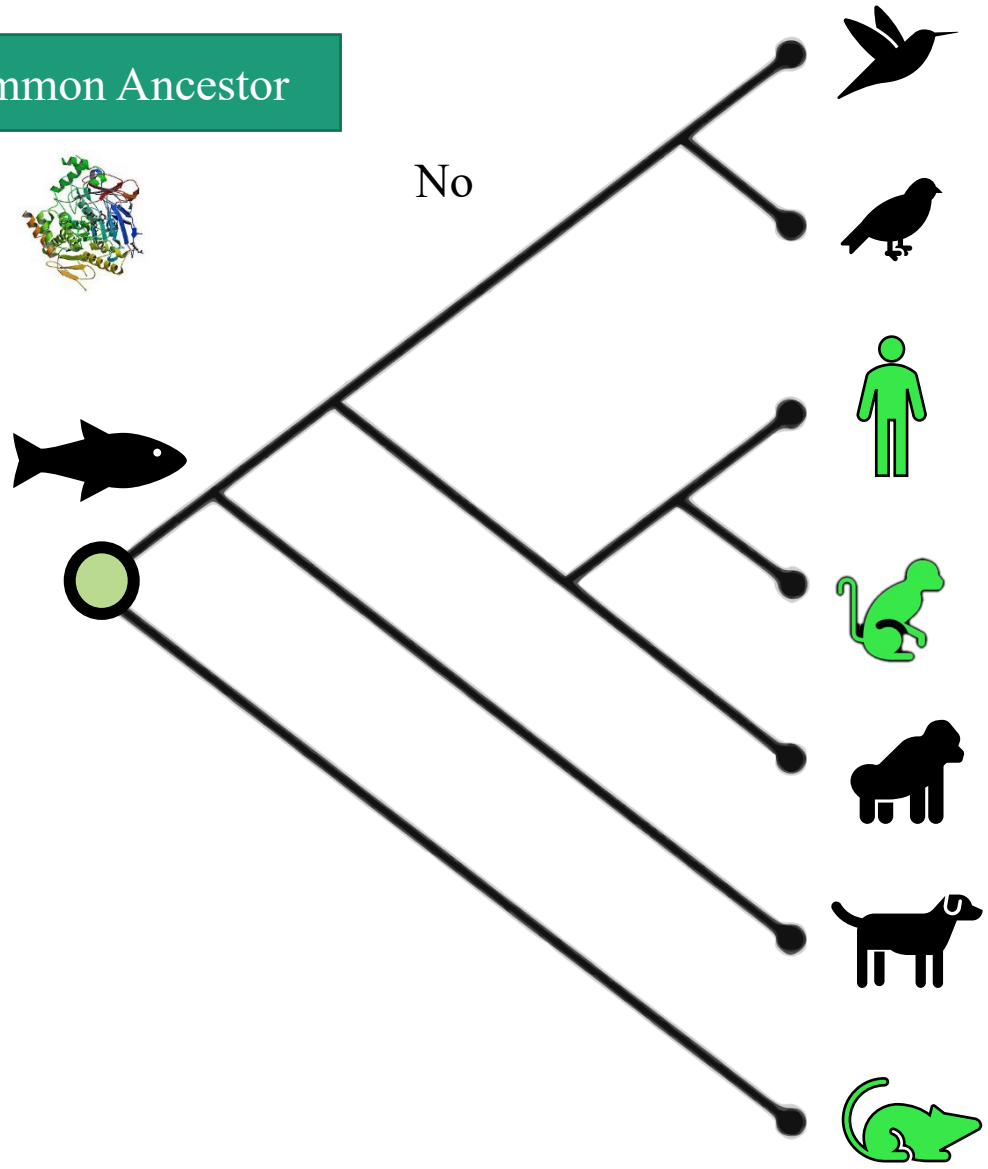


Proteins in different species that evolved from a common ancestor

Typically maintain similar function

SeqAPASS Level 1





SeqAPASS Level 1

Common Name	Ortholog Candidate	Cut-off	Percent Similarity
Human	Y	33.15	100
Florida manatee	Y	33.15	98.8
Mallard	Y	33.15	82.29
Rock pigeon	Y	33.15	80.93
Green anole	Y	33.15	80.65
Pacific transparent sea squirt	Y	33.15	33.15
Yesso scallop	N	33.15	32.87
Purple sea urchin	N	33.15	26.05
Human whipworm	N	33.15	23.53
Bed bug	N	33.15	21.62

Lowest % Similarity that is still an ortholog

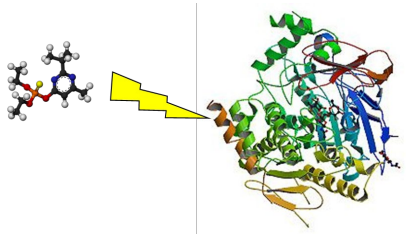
Example:

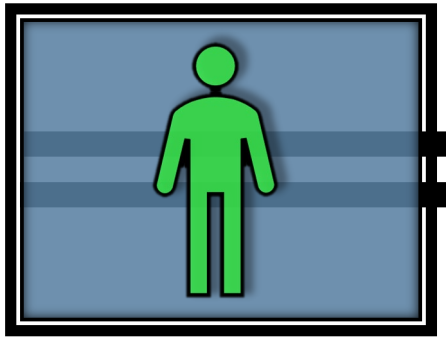
Susceptibility Cut-off: Set at 33.15

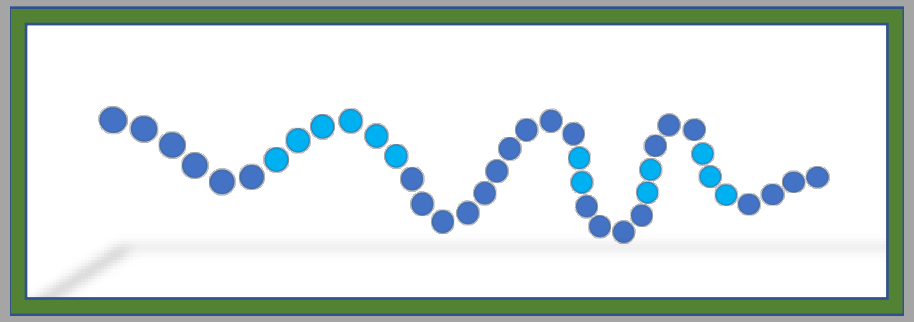
Above cut-off: More likely to be susceptible base on similar **FUNCTION**

SeqAPASS Level 1

Human Protein Target



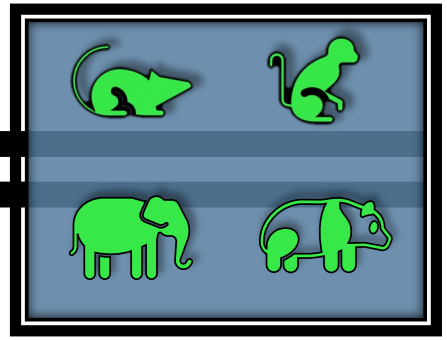




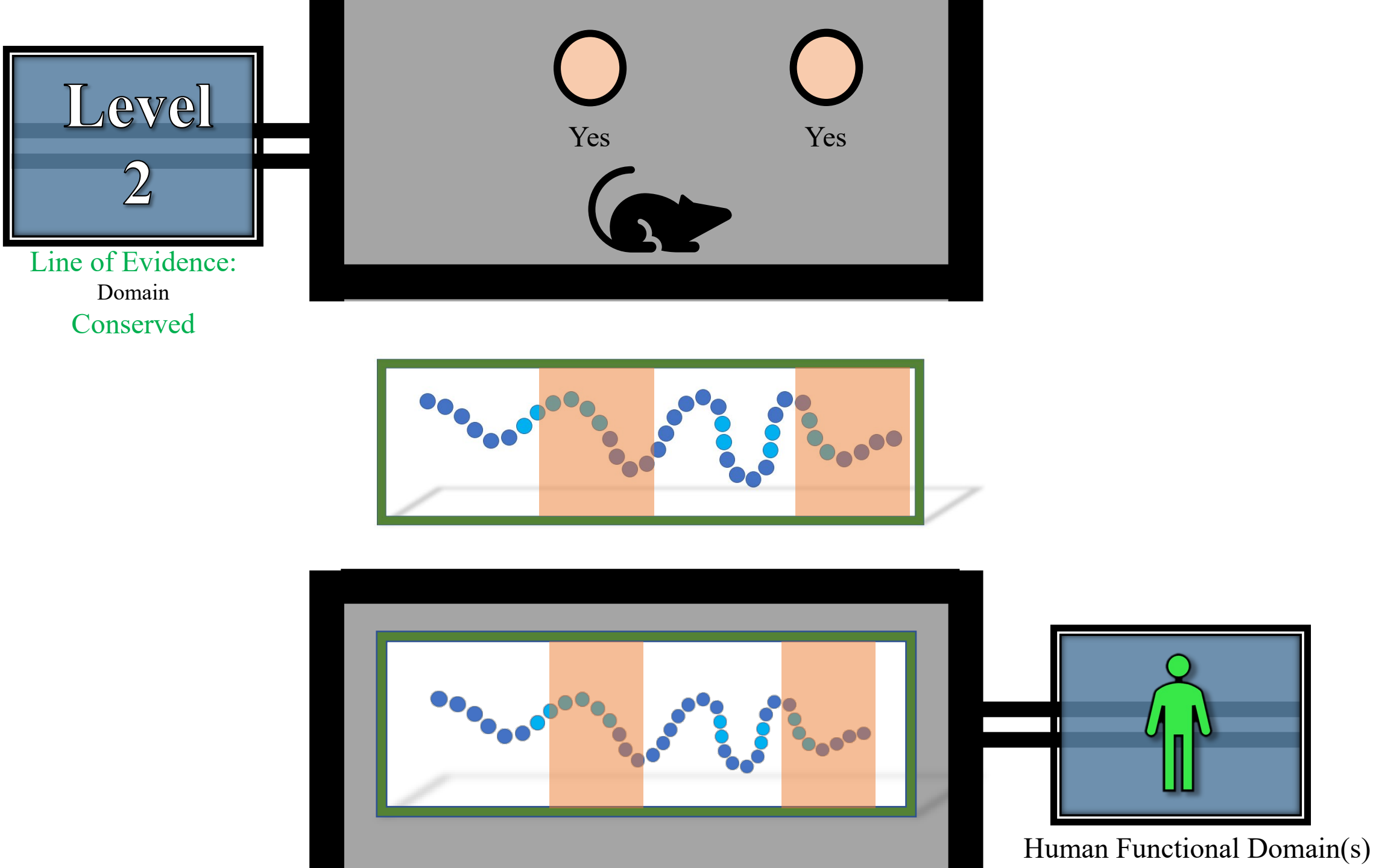
Line of Evidence:

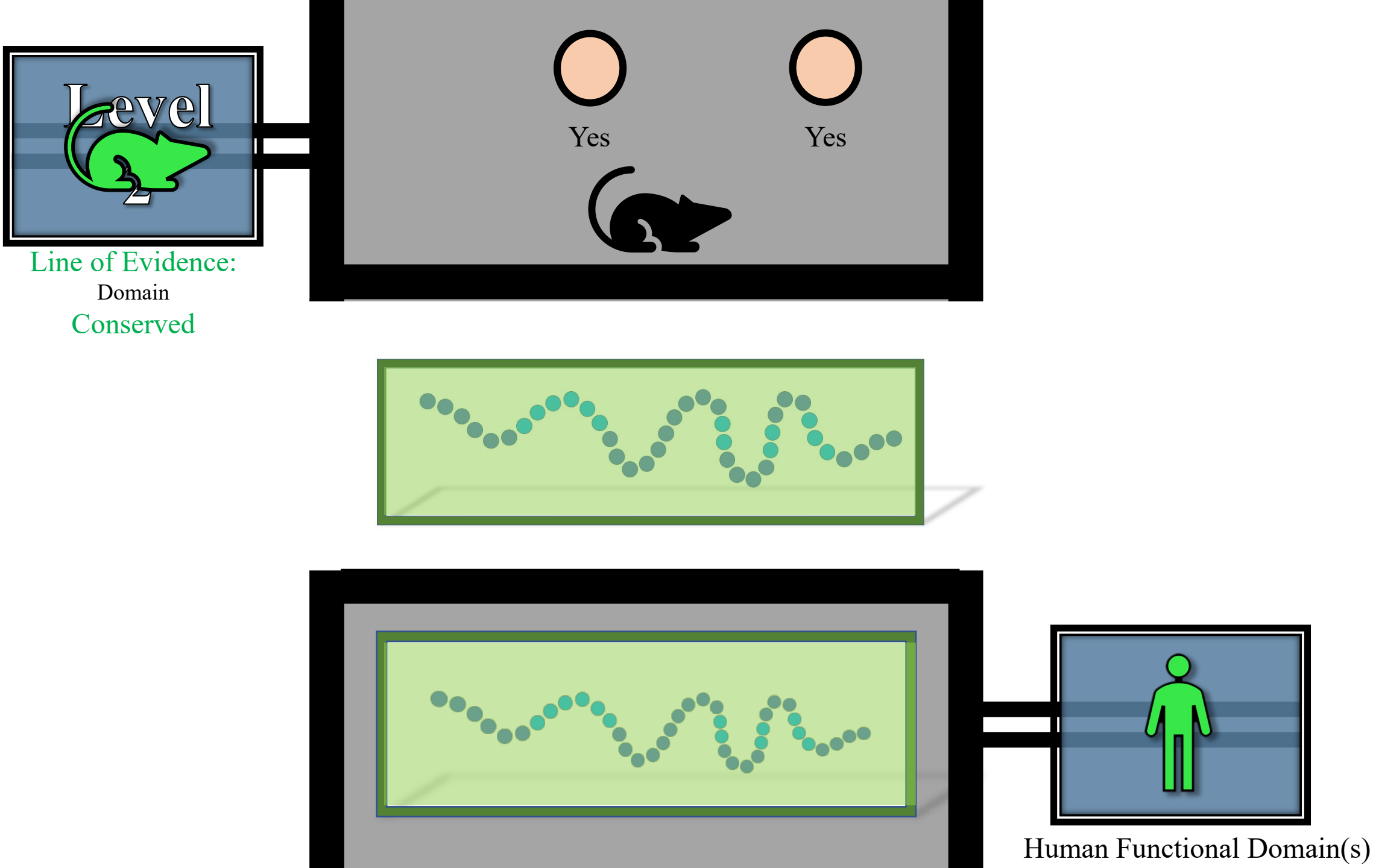
Primary amino acid sequence

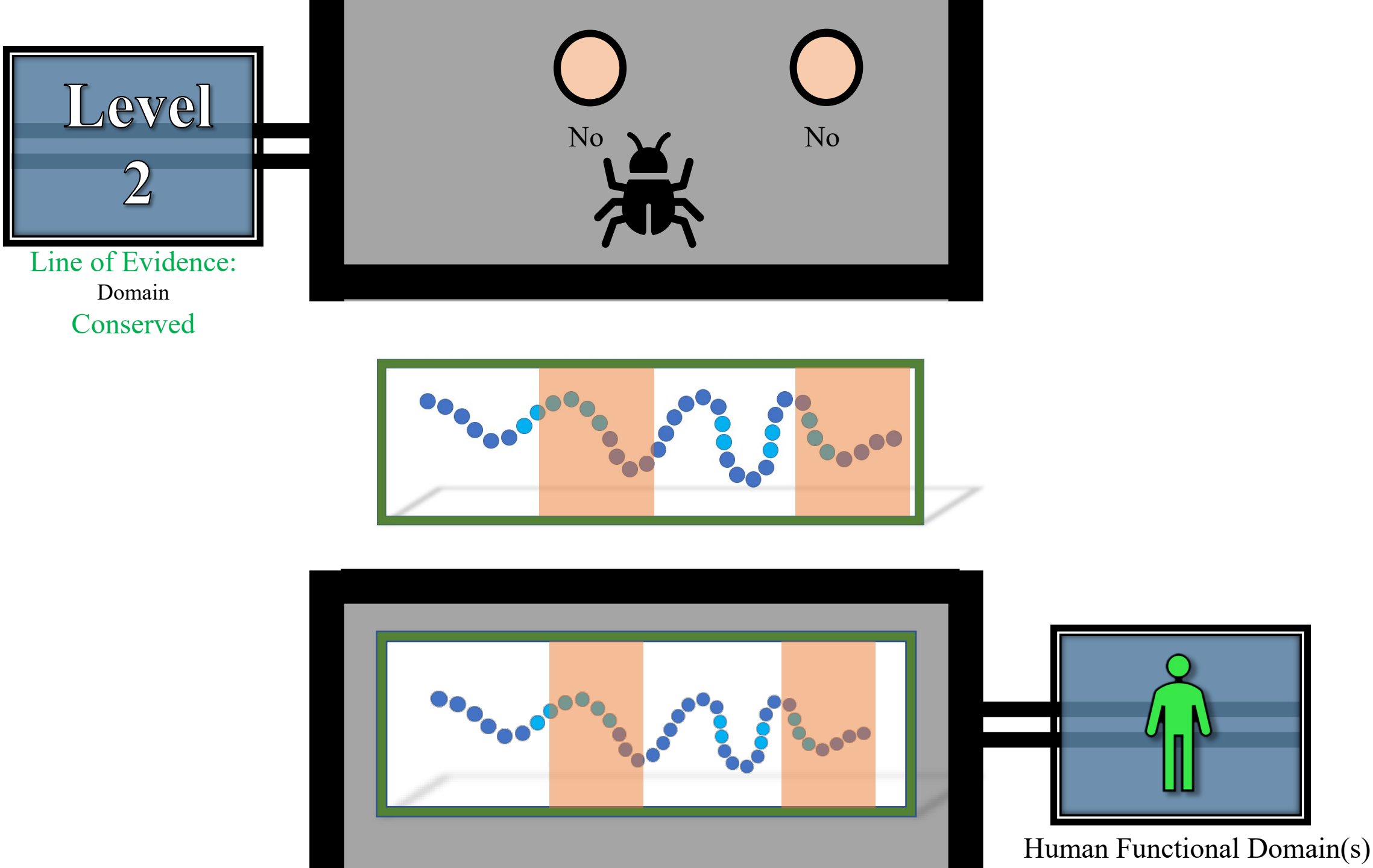
Conserved

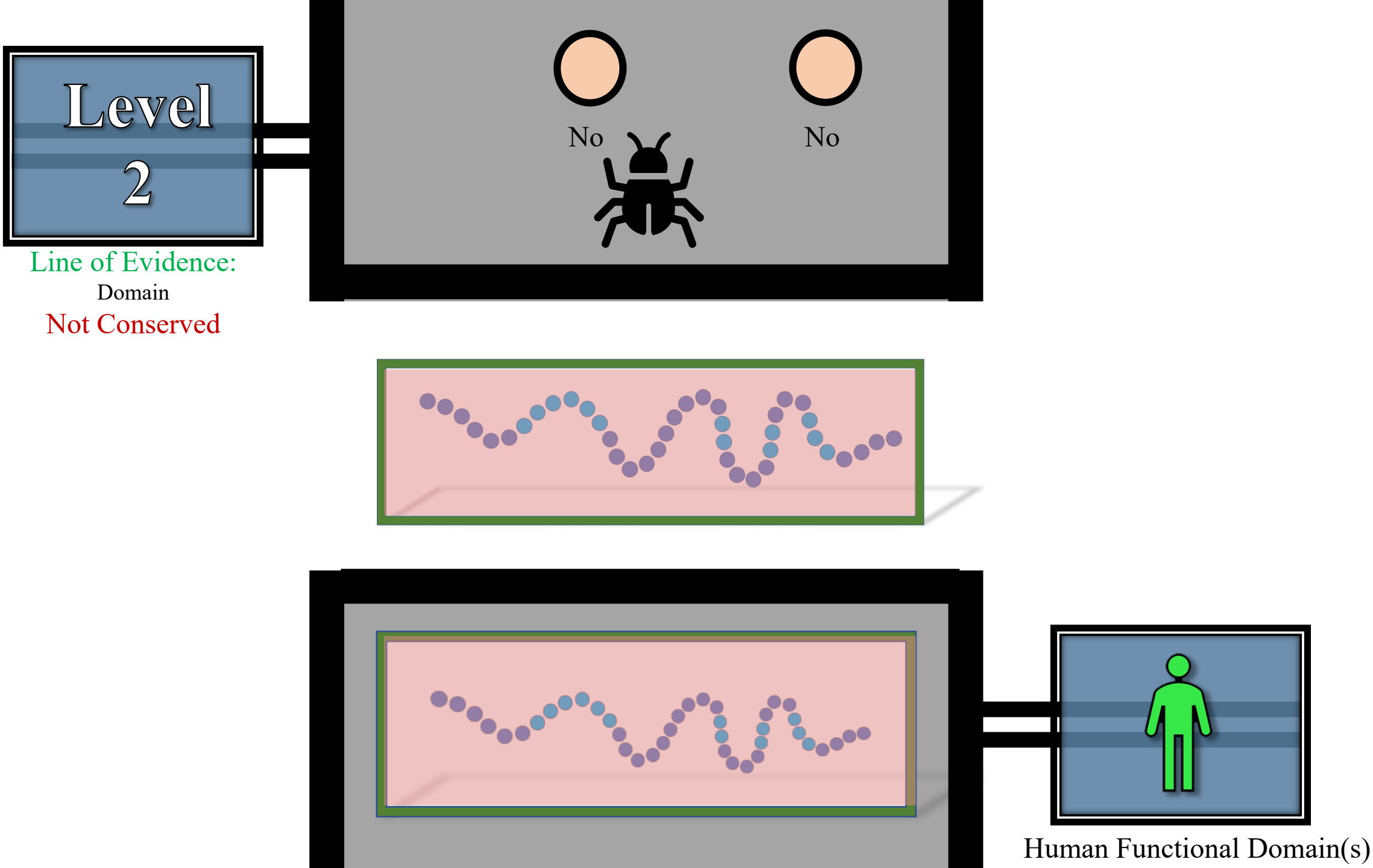


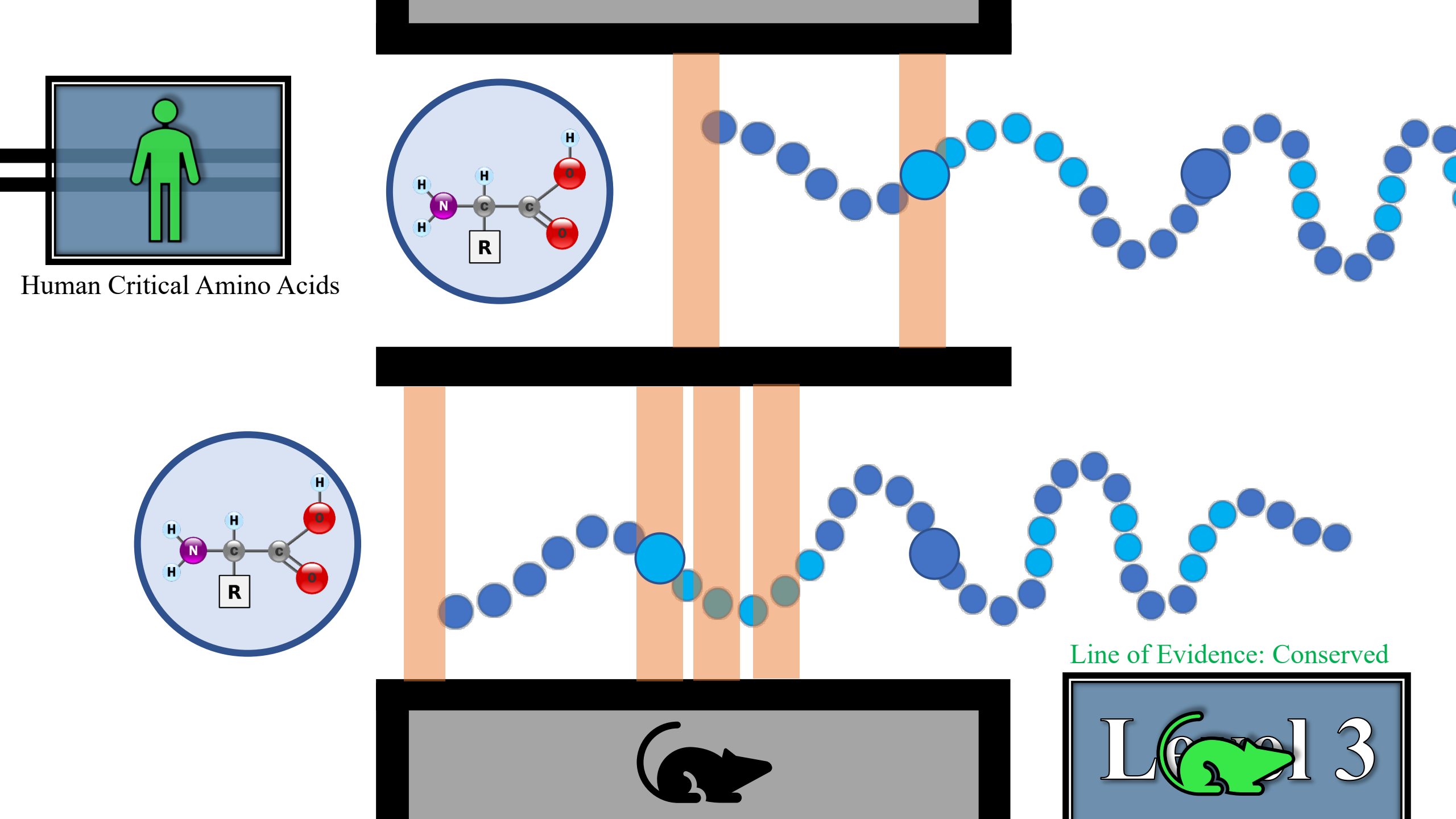
Hundreds to Thousands of Species

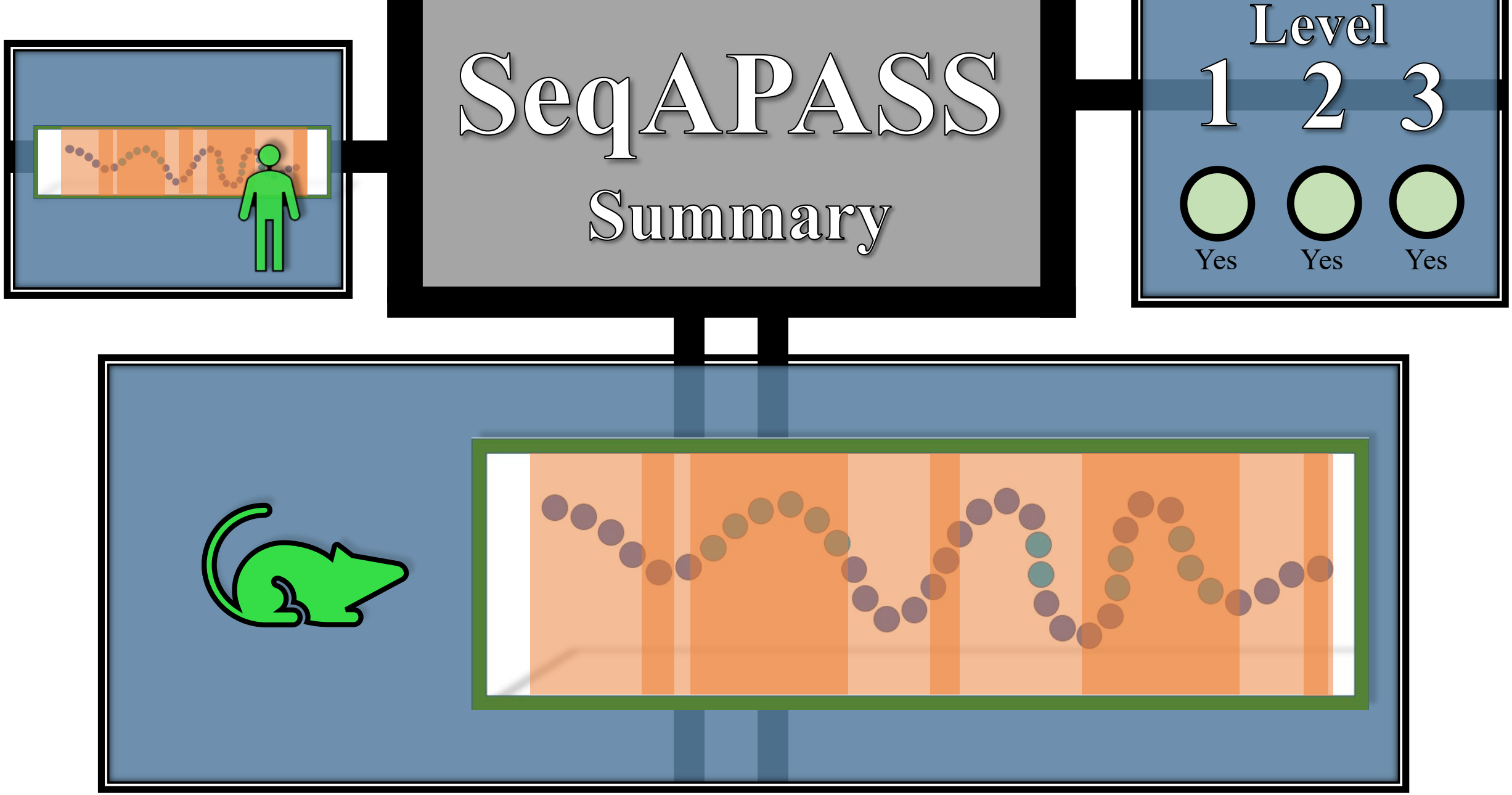

Level 2
Line of Evidence:
Domain
Conserved
Yes
Yes
Human Functional Domain(s)


Level
2
Line of Evidence:
Domain
Conserved
Yes
Yes
Human Functional Domain(s)


Level 2
Line of Evidence:
Domain
Conserved
No
No
Human Functional Domain(s)

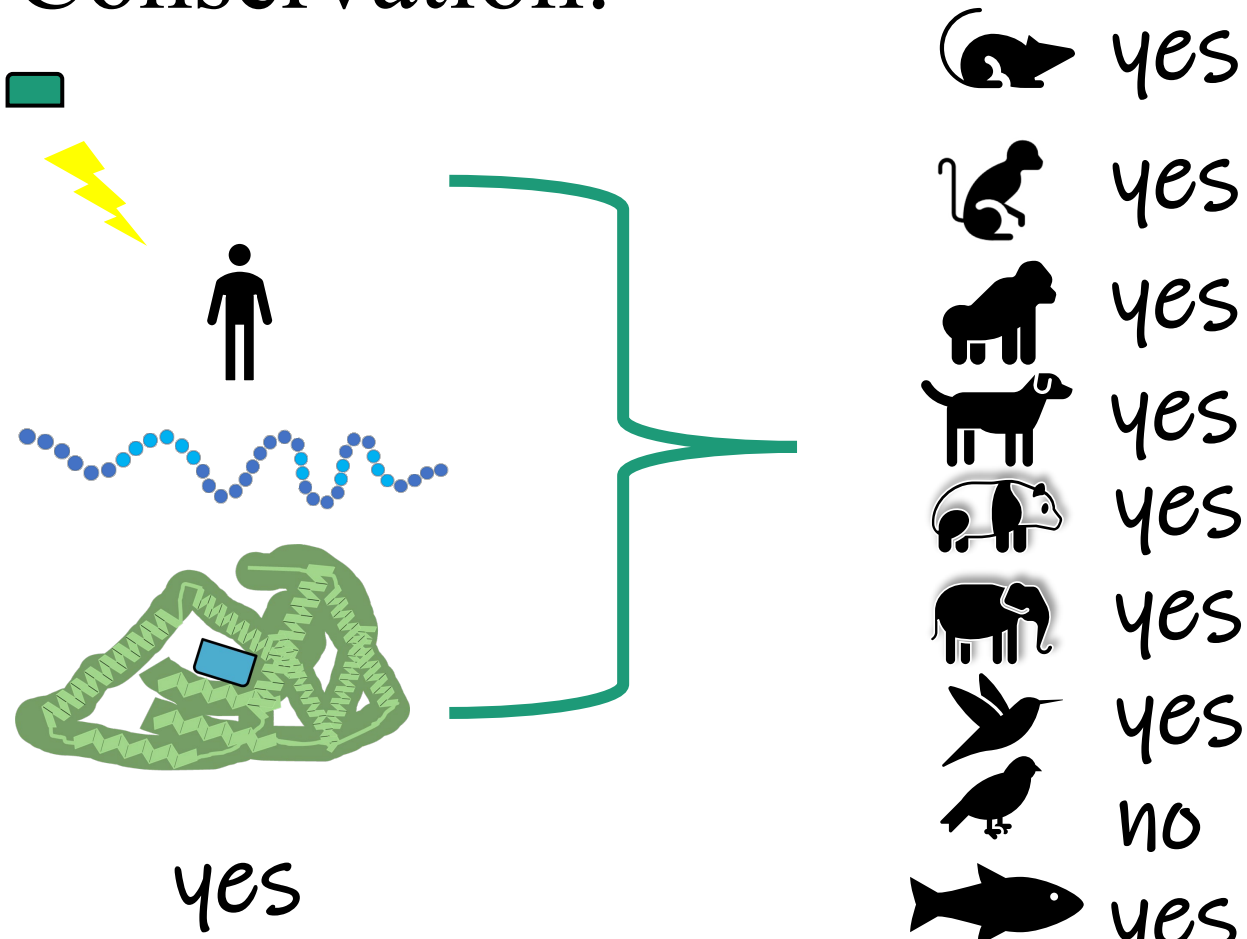

Level 2
Line of Evidence:
Domain
Not Conserved
No
No
Human Functional Domain(s)


Human Critical Amino Acids
H
N
C
C
O
O
R
Line of Evidence: Conserved
Level 3



Gather Lines of Evidence for Conservation of Protein Target:
Susceptibility Prediction: Yes or No

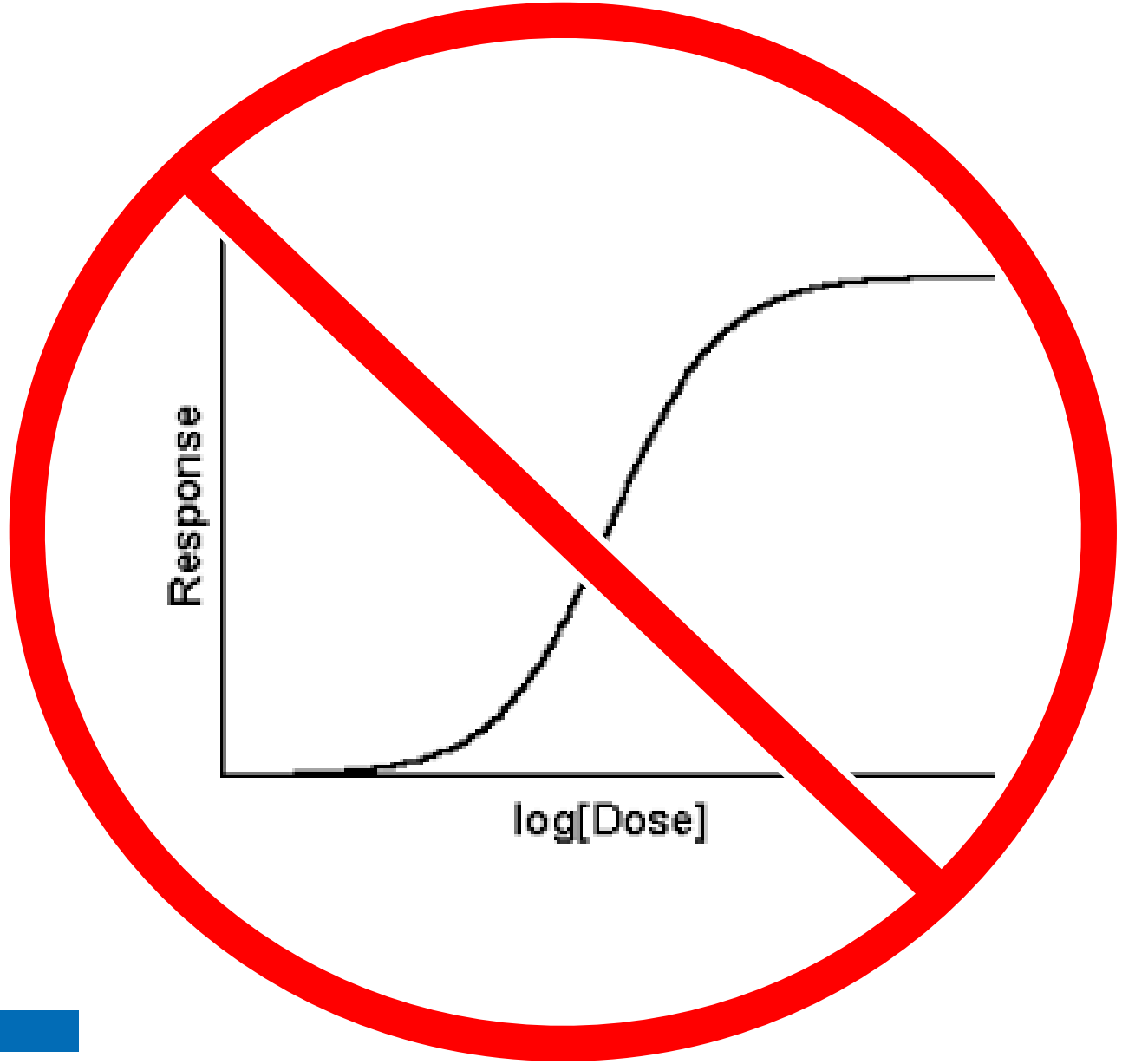
SeqAPASS Predicts Likelihood of Similar Susceptibility based on Sequence Conservation:



Line(s) of evidence indicate

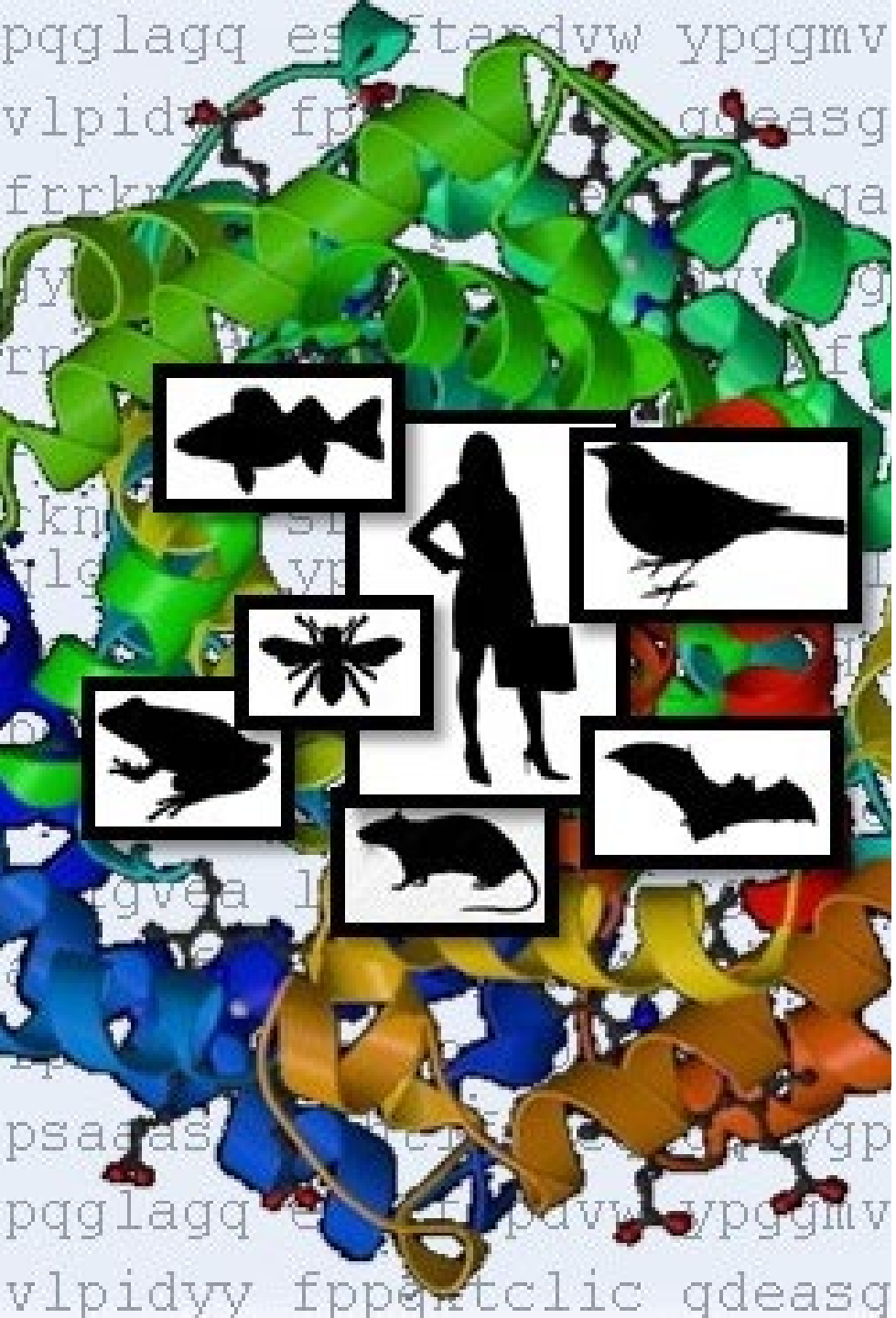
- The protein is conserved
- The protein is NOT conserved

SeqAPASS DOES NOT predict the degree of sensitivity/susceptibility:



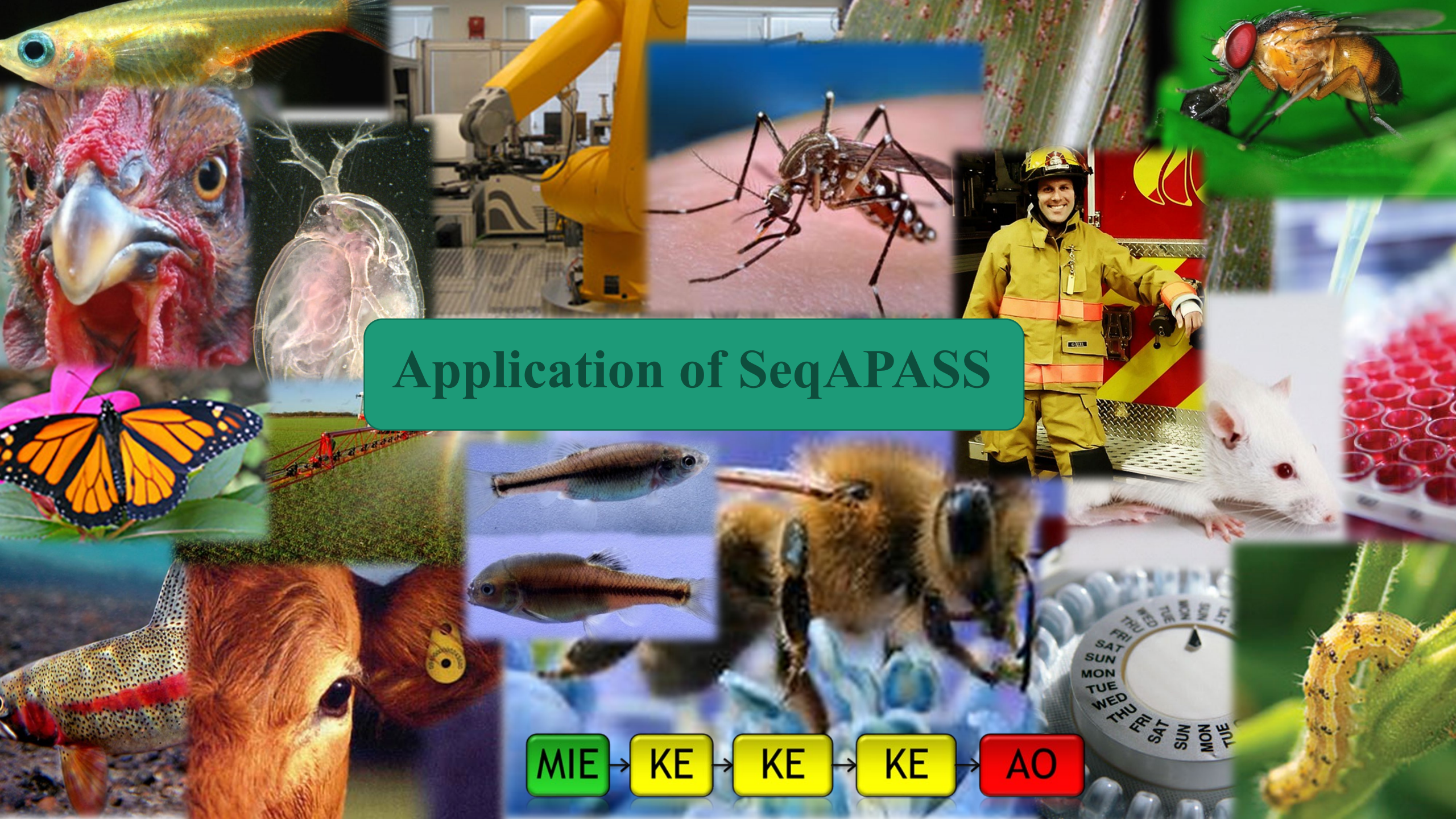
Factors that make a species sensitive

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



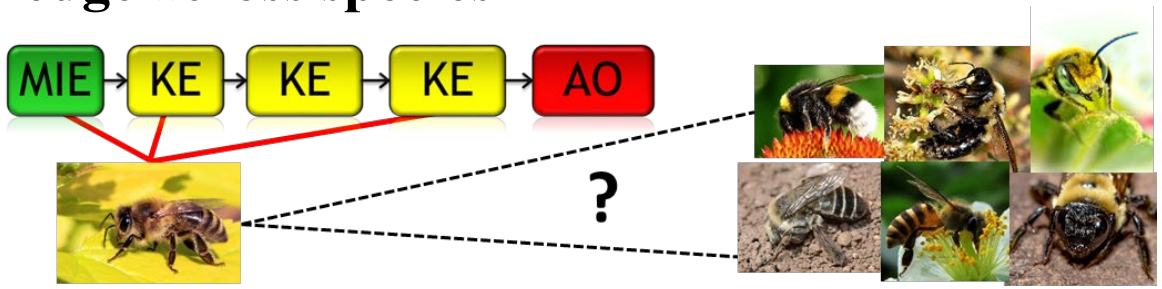
Strengths of SeqAPASS

- **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

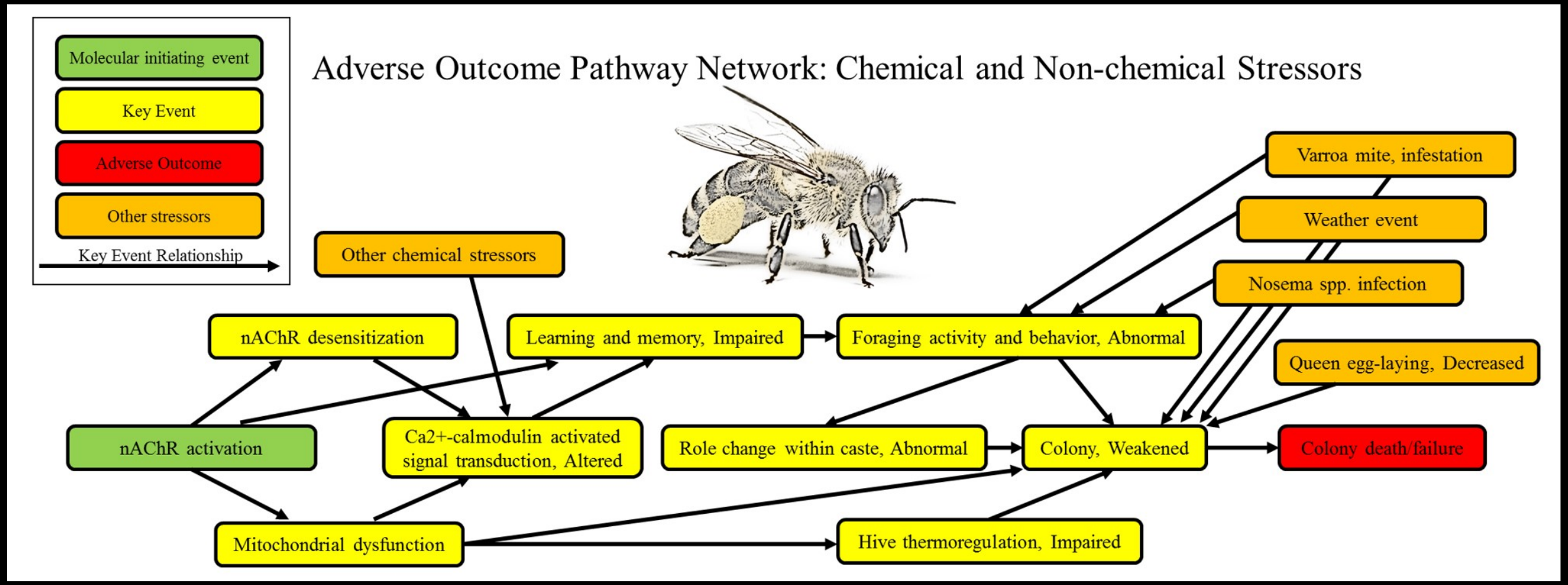
Applications of Bioinformatics: Case Studies

- **Extrapolate adverse outcome pathway knowledge across species**
  - **Define the taxonomic domain of applicability**
    - 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

LaLone et al., 2017. STOTEN 584-585, 751-775



Define Knowledge Gaps
Understand nodes that may be impacted by multiple stressors
Assists in development of mitigation strategies

How to define the taxonomic relevance of an AOP?

AOP Wiki Key Event IDs

AOP Wiki AOP IDs

	KE 559	KE 663	KE 177	KE 1243	KE 341	KE 560	KE 1108	KE 568	KE 1107	KE 563
AOP 88	MIE	KE		KE	KE	KE			KE	AO
AOP 89	MIE	KE		KE	KE	KE	KE		KE	AO
AOP 77	MIE		KE	KE	KE	KE			KE	AO
AOP 87	MIE		KE	KE	KE	KE	KE		KE	AO
AOP 79	MIE		KE					KE	KE	AO
AOP 178	MIE		KE						KE	AO

Modified from LaLone et al., 2017

MIE	KER1	KE1	KER2	KE2	KER3	KE3	KER4	KE4	KER5	KE5	KER6	KE6	KER7	AO
nAChR activation		nAChR desensitization		Ca^{2+}-calmodulin activated signal transduction, Altered		Learning and memory, Impairment		Foraging activity and behavior, Abnormal		Role change within caste, Abnormal		Colony, Weakened		Colony death/failure

Case Example

How to define the taxonomic relevance of an AOP?

Biologically Plausible tDOA: Decision Tree

Empirical Evidence tDOA

MIE

Is there evidence that the nAChR is conserved in the organism?

- No
- Yes

KE1

Is there evidence that the nAChR is conserved in the organism?

- No
- Yes

KE2

Is there evidence that the calcium-calmodulin signal transduction pathway is conserved in the organism?

- No
- Yes

KE3

Does the organism rely on learning and memory for survival?

- No
- Yes

KE4

Does species rely on foraging food sources to support a colony?
Does species use signals to relay information on food source location to other colony members?

- No
- Yes

KE5

Do inappropriate or untimely role changes of the organism lead to adverse impacts on the colony?

- No
- Yes

KE6

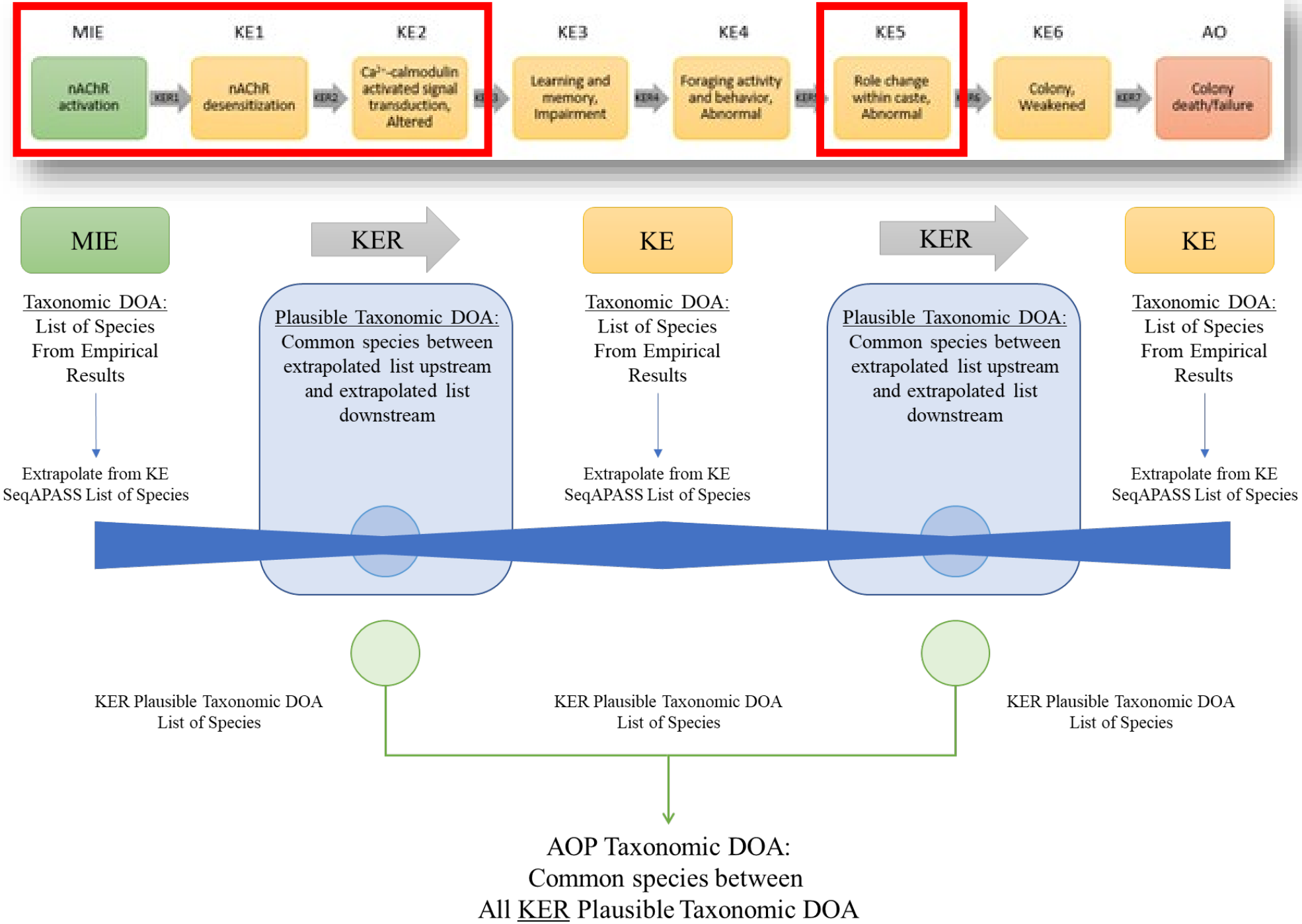
Does the organism depend on a colony structure for survival?

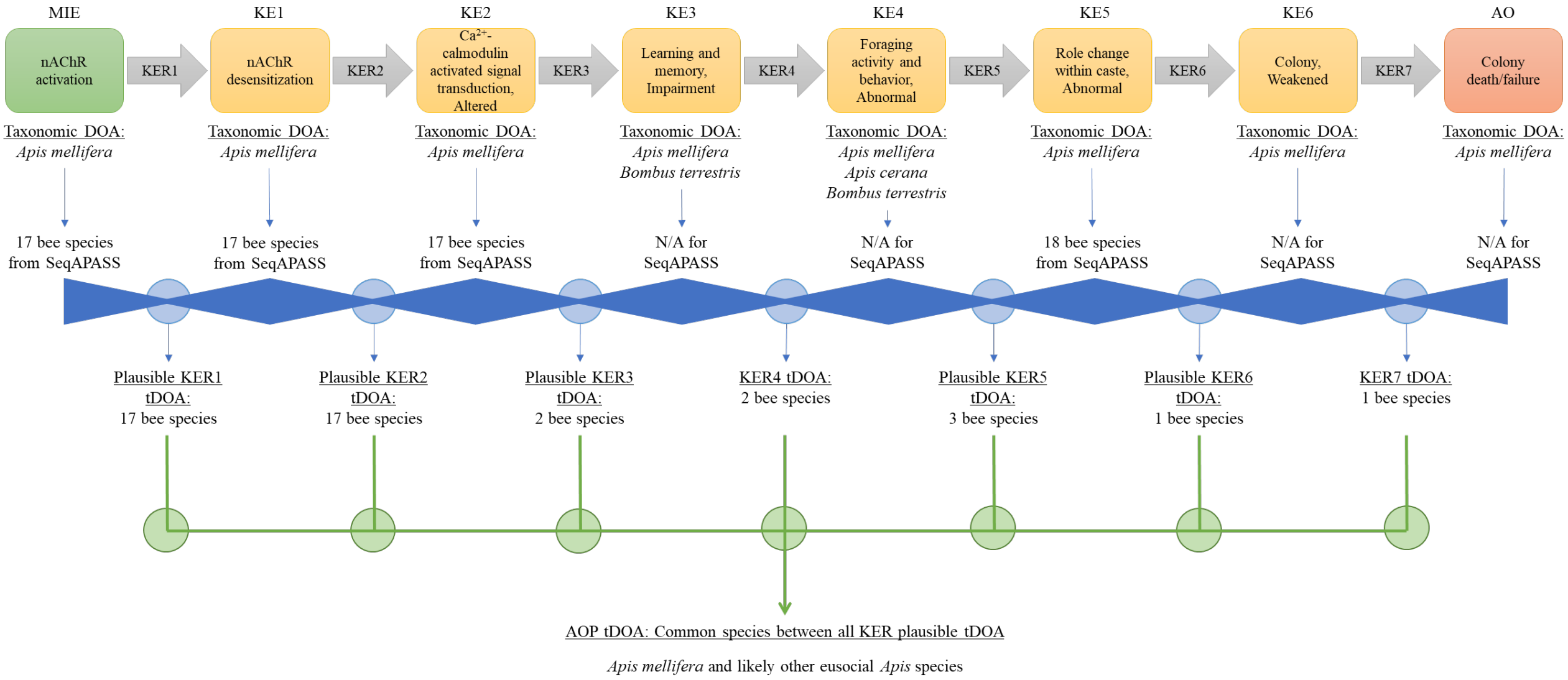
- No
- Yes

AO

Does the species live in a colony structure?

- No


MIE
KE1
KE2
KE3
KE4
KE5
KE6
AO
nAChR activation
nAChR desensitization
Ca2+-calmodulin activated signal transduction, Altered
Learning and memory, Impairment
Foraging activity and behavior, Abnormal
Role change within caste, Abnormal
Colony, Weakened
Colony death/failure
MIE
KER
KE
KER
KE
Taxonomic DOA: List of Species From Empirical Results
Plausible Taxonomic DOA: Common species between extrapolated list upstream and extrapolated list downstream
Taxonomic DOA: List of Species From Empirical Results
Plausible Taxonomic DOA: Common species between extrapolated list upstream and extrapolated list downstream
Taxonomic DOA: List of Species From Empirical Results
Extrapolate from KE SeqAPASS List of Species
Extrapolate from KE SeqAPASS List of Species
Extrapolate from KE SeqAPASS List of Species
KER Plausible Taxonomic DOA List of Species
KER Plausible Taxonomic DOA List of Species
KER Plausible Taxonomic DOA List of Species
AOP Taxonomic DOA: Common species between All KER Plausible Taxonomic DOA


MIE
KE1
KE2
KE3
KE4
KE5
KE6
AO
nAChR activation
KER1
nAChR desensitization
KER2
Ca2+-calmodulin activated signal transduction, Altered
KER3
Learning and memory, Impairment
KER4
Foraging activity and behavior, Abnormal
KER5
Role change within caste, Abnormal
KER6
Colony, Weakened
KER7
Colony death/failure
Taxonomic DOA: Apis mellifera
Taxonomic DOA: Apis mellifera
Taxonomic DOA: Apis mellifera
Taxonomic DOA: Apis mellifera Bombus terrestris
Taxonomic DOA: Apis mellifera Apis cerana Bombus terrestris
Taxonomic DOA: Apis mellifera
Taxonomic DOA: Apis mellifera
Taxonomic DOA: Apis mellifera
17 bee species from SeqAPASS
17 bee species from SeqAPASS
17 bee species from SeqAPASS
N/A for SeqAPASS
N/A for SeqAPASS
18 bee species from SeqAPASS
N/A for SeqAPASS
N/A for SeqAPASS
Plausible KER1 tDOA: 17 bee species
Plausible KER2 tDOA: 17 bee species
Plausible KER3 tDOA: 2 bee species
KER4 tDOA: 2 bee species
Plausible KER5 tDOA: 3 bee species
Plausible KER6 tDOA: 1 bee species
KER7 tDOA: 1 bee species
AOP tDOA: Common species between all KER plausible tDOA
Apis mellifera and likely other eusocial Apis species

Empirical Evidence tDOA

MIE

Is there evidence that the nAChR is conserved in the organism?

No

Yes

KE1

Is there evidence that the nAChR is conserved in the organism?

No

Yes

KE2

Is there evidence that the calcium-calmodulin signal transduction pathway is conserved in the organism?

No

Yes

KE3

Does the organism rely on learning and memory for survival?

No

Yes

KE4

Does species rely on foraging food sources to support a colony?
Does species use signals to relay information on food source location to other colony members?

No

Yes

KE5

Do inappropriate or untimely role changes of the organism lead to adverse impacts on the colony?

No

Yes

KE6

Does the organism depend on a colony structure for survival?

No

Yes

AO

Does the species live in a colony structure?

No

SeqAPASS Expands tDOA

Apis cerana	*Bombus terrestris*	*Habropoda laboriosa*
Apis dorsata	*Bombus vancouverensis nearcticus*	*Megachile rotundata*
Apis florea	*Bombus vosnesenskii*	*Nomia melanderi*
Apis mellifera	*Ceratina calcarata*	*Osmia bicornis bicornis*
Bombus biffarius	*Dufourea novaeangliae*	*Osmia lignaria*
Bombus impatiens	*Eurfriesea mexicana*	

Note: This is a rather focused example

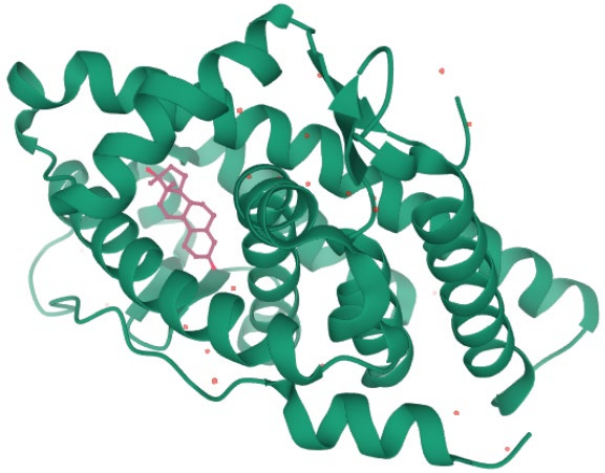
United States Environmental Protection Agency

Advances in Bioinformatics – Future of SeqAPASS

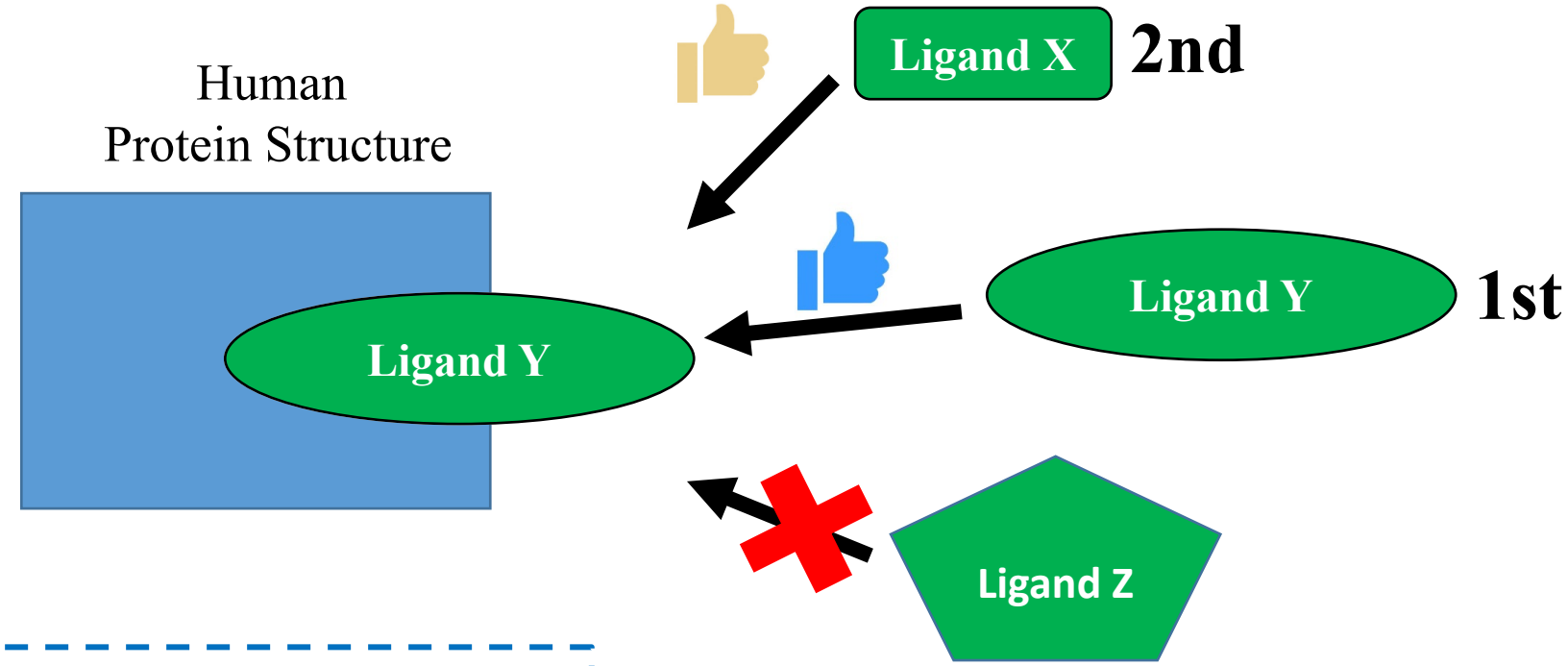
Always Look Several Steps Ahead

Advances in Drug Discovery/Development

(COVID-19 has led to advances)



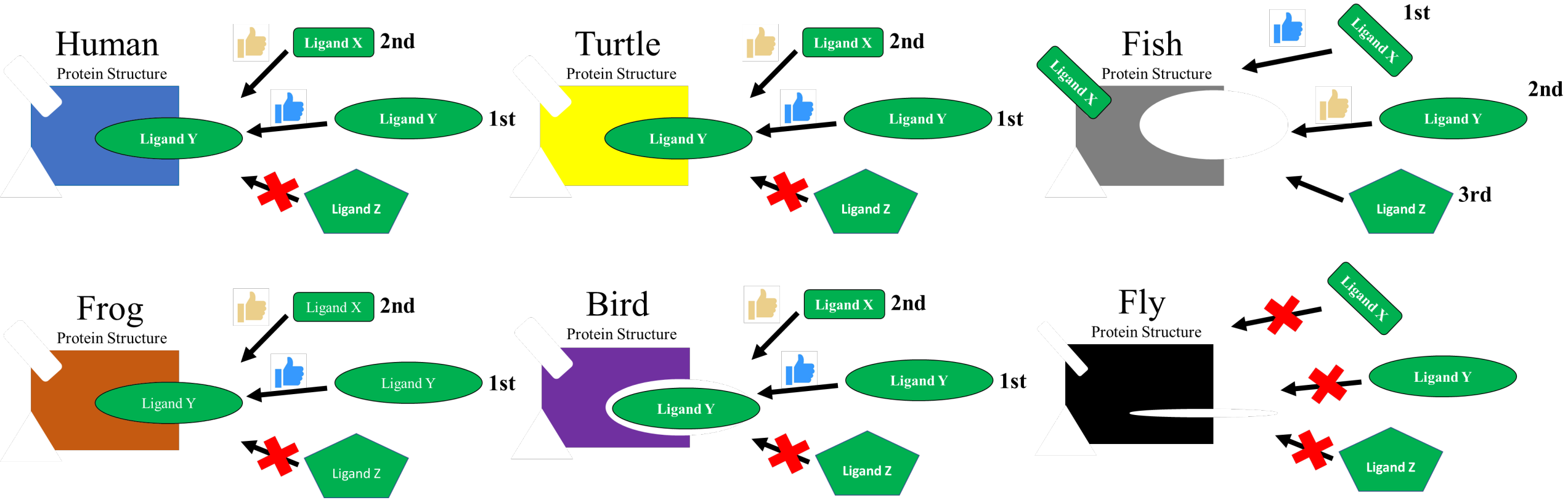
Structure derived from X-ray crystallography



Bioinformatics Toolbox:

Molecular modeling
Molecular docking
Virtual screening
Molecular dynamic simulations

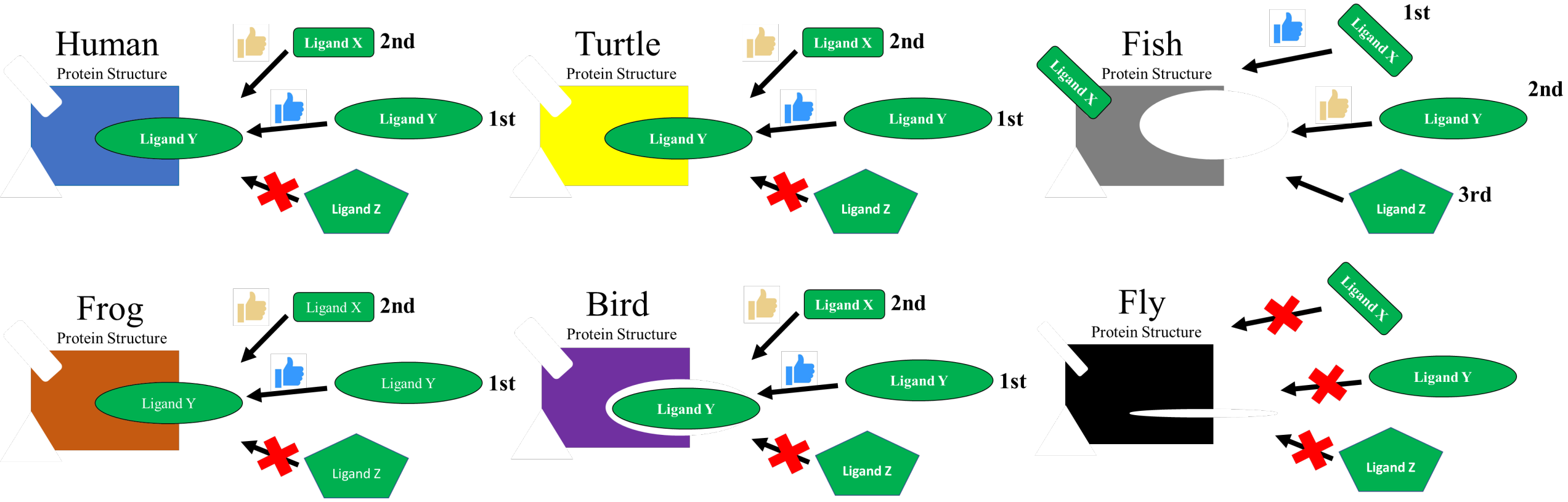
Application to Species Extrapolation



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

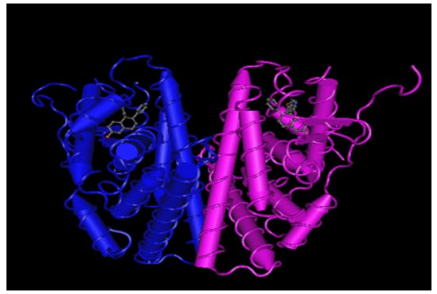
Thousands/Millions/Billions of Chemicals

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

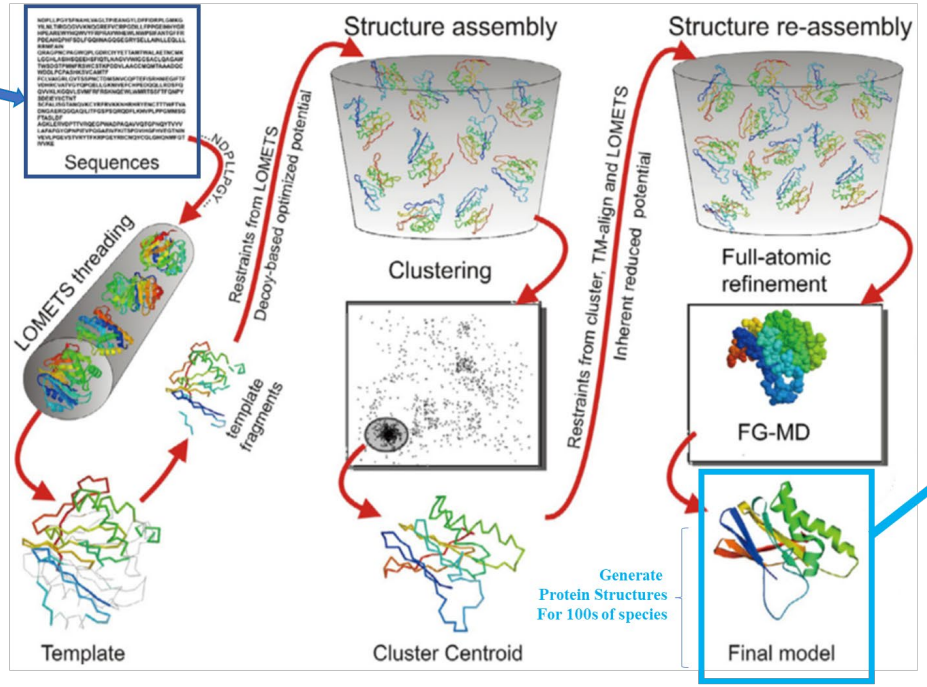
>KFQ76585.1 Protein X [Phoenicopterus ruber ruber]
MSFTGKYELQSQENFEPFMKALGLPDDQIQKGKD
IKSISEIVQDGKKFKVTVTTGSKVMQNEFTIGEECD
IEMLTGEKVKAVVQMEGNNRLVANLKGLKSVTEL
NGDIITHTMTMGDLTYKRISKRI

>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/)

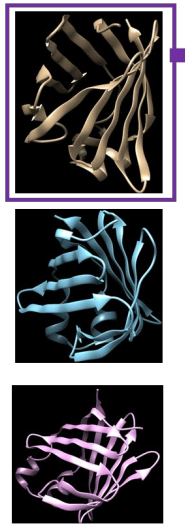


Graphic Modified from Zhang et al., 2019 I-TASSER gateway: A protein structure and function prediction server powered by XSEDE Figure 1

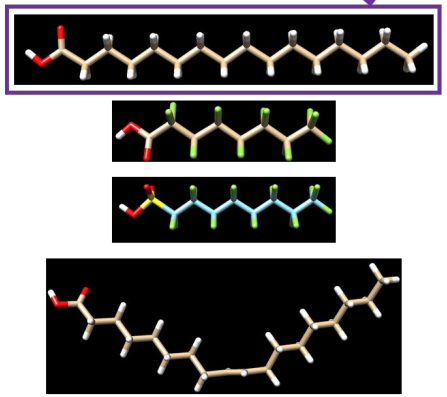
UCSF Chimera

DockPrep Structures and Minimize Ligands

Protein Structure Models From 100s of Species

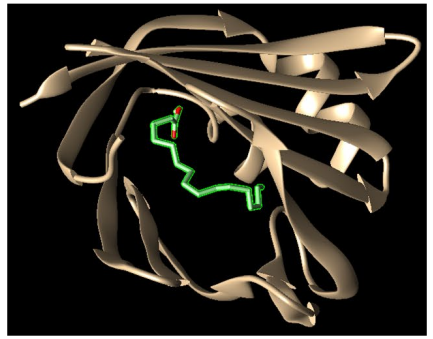


Ligands of Interest for Docking



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
```

Predicting Binding Affinity

Government

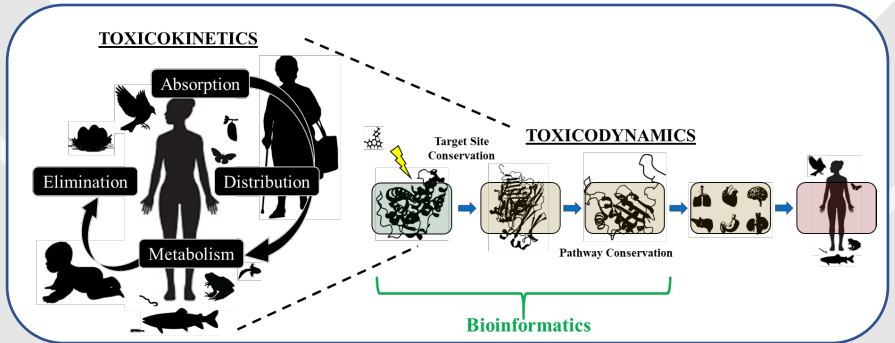
Industry

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)
Monique Hazemi (ORISE)
Sara Vliet (past ORISE 2021)
Jon Doering (past NRC 2018)
Colin Finnegan (past ORISE 2018)
Donovan Blatz (past ORISE 2021)

GDIT

Cody Simmons
Audrey Wilkinson
Wilson Menendez
Thomas Transue (past GDIT 2022)

SeqAPASS v6.0 (Released Sept. 2021)

LaLone.Carlie@epa.gov

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