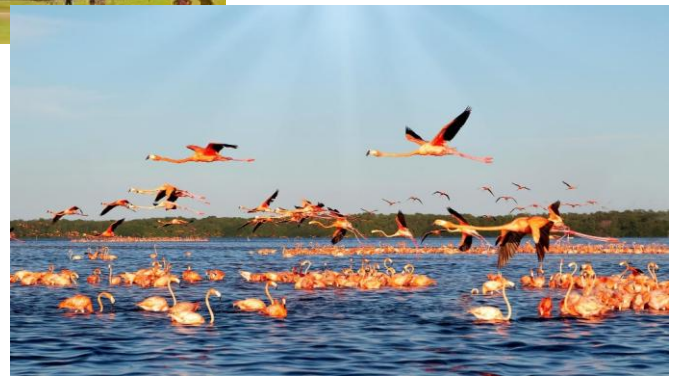
# Rethinking Bioactivity Modeling: Representation, Context, and Predictive Space

**Karina Martinez-Mayorga**
Institute of Chemistry, Campus Merida,
National Autonomous University of Mexico




# Outline of the talk

- What this talk is not about … *unfortunately*
- Main ideas
  - **Representation**: what models actually see
  - **Context**: when the same structure yields different outputs
  - **Predictive space**: intent matters
- Open questions for the community
- Take-home message

# What this talk is not about … *unfortunately*




Abraham Madariaga Mazón
Adriana Osnaya Hernández
Alejandro Hernández Cano
Andrés Felipe Marmolejo Valencia
Ariadna Irais Ramírez Hernández
Austin B. Yongye
Axel Adrián López Cedano
Axel Magdaleno Rodríguez
Camila Garibay Manriquez
César Raúl García Jacas
Coray Colina
David Ignacio Ramírez Palma
Diana Esmeralda Terrazas Álvarez
Elier Abreu Martínez
Erick Alberto Correa Padilla
Estibalis Arni Daniel Chávez Gómez
Fernando Daniel Prieto Martínez
Gabriela Gómez Jiménez
Jennifer Mendoza Cañas
José de Jesús Naveja Romero
José Guadalupe Rosas Jiménez
Jesús Cob
Julio Granados Montiel
Karla González Ponce
Karla Olivia Noriega Colima
Kendall G. Byler
Nancy de Jesús Reyes
Norberto Sánchez Cruz
Regnier Cano Blanco
Ricardo Bruno Hernández Alvarado
Roberto Rolón Cárdenas
Thomas R. Caulfield
Víctor Adrián Suárez Ortega
Víctor Alfonso Morado Rull

We work on cheminformatics, bioactivity modeling, and data-driven analysis of chemical–biological systems

# Why rethink bioactivity modeling?

## Twenty-one errors encounter when developing QSAR models

| Avoidable | Partially/hardly avoidable |
|---|---|
| Failure to take account of data heterogeneity | Use of incomprehensible descriptors |
| Use of inappropriate endpoint data | Error in descriptor values |
| Use of collinear descriptors | Lack of mechanistic interpretation |
| Poor transferability of QSAR/QSPR | Too narrow a range of endpoint values |
| Inadequate/undefined applicability domain | |
| Unacknowledged omission of data points/lack of activity cliffs analysis | |
| Use of inadequate data | |
| Replication of compounds in dataset | |
| Over-fitting of data | |
| Use of excessive numbers of descriptors | |
| Lack of/inadequate statistics | |
| Incorrect calculation | |
| Lack of descriptor auto-scaling | |
| Misuse/misinterpretation of statistics | |
| No consideration of distribution of residuals | |
| Inadequate training/test set selection | |
| Failure to validate a QSAR/QSPR correctly | |

*Chem. Sci*. **2024**, *15*, 1938-1952.
*J Med Chem*. **2014** 57(12):4977-5010.
*J Chem Inf Model*. **2012** *52* (8), 2044-2058.
*SAR QSAR Environ Res*. **2009**;20(3-4):241-66.

Oversimplifications and assumptions limits model performance

Beyond increasing data size or improving algorithms, context matters

***Predictive boundaries are not always technical***

Representation – Context – Predictive Space

# REPRESENTATION

- Endpoint definition introduces biological **simplifications** and may obscure complex mechanisms
- Representation (molecular descriptors) adapts to that simplified task
- What is represented determines what can be learned

# Molecular descriptors as a language

- Descriptors translate molecular structure into a numerical form
- Different descriptors emphasize different molecular properties
- Some aspects become explicit, while others remain invisible

# Quantum-derived descriptors

- Offer a complementary view of the molecules
- Make electronic and density-based features explicit
- Lead to a different chemical space
- Example: QTAIM-derived features

# Structural Interpretation of Activity Cliffs Revealed by Systematic Analysis of Structure−Activity Relationships in Analog Series

Mihiret T. Sisay,[†,‡,§] Lisa Peltason,[†,§] and Jürgen Bajorath*[,†]

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany, and Pharmazeutisches Institut, Pharmazeutische Chemie I, Rheinische Friedrich-Wilhelms-Universität, An der Immenburg 4, D-53121 Bonn, Germany

Structural interpretation helped to understand discontinuous SAR, in cases where ligand-based analysis was insufficient.

*J Chem Inf Model*. **2009** Oct;49(10):2179-89

## Inhibition of Prohormone Convertases PC1/3 and PC2 by 2,5-Dideoxystreptamine Derivatives

Mirella Vivoli, Thomas R. Caulfield, Karina Martínez-Mayorga, Alan T. Johnson, Guan-Sheng Jiao, and Iris Lindberg

*Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland (M.V., I.L.); Torrey Pines Institute for Molecular Studies, Port St. Lucie, Florida (T.R.C., K.M.-M.); and the Department of Chemistry, PanThera Biopharma LLC, Aiea, Hawaii (A.T.J., G.-S.J.)*

## Identification of a Small Molecule That Selectively Inhibits Mouse PC2 over Mouse PC1/3: A Computational and Experimental Study

**Austin B. Yongye[1], Mirella Vivoli[2], Iris Lindberg[2], Jon R. Appel[3], Richard A. Houghten[1,3], Karina Martinez-Mayorga[1,4]***

**1** Torrey Pines Institute for Molecular Studies, Port St Lucie, Florida, United States of America, **2** Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland, United States of America, **3** Torrey Pines Institute for Molecular Studies, San Diego, California, United States of America, **4** Instituto de Química, UNAM, Mexico City, Mexico

At low doses stimulate a response

and

High doses inhibit this response

*Mol Pharmacol* . **2012**, *81*, 440-54.
*PLOS One* **2013**, *8*, e56957

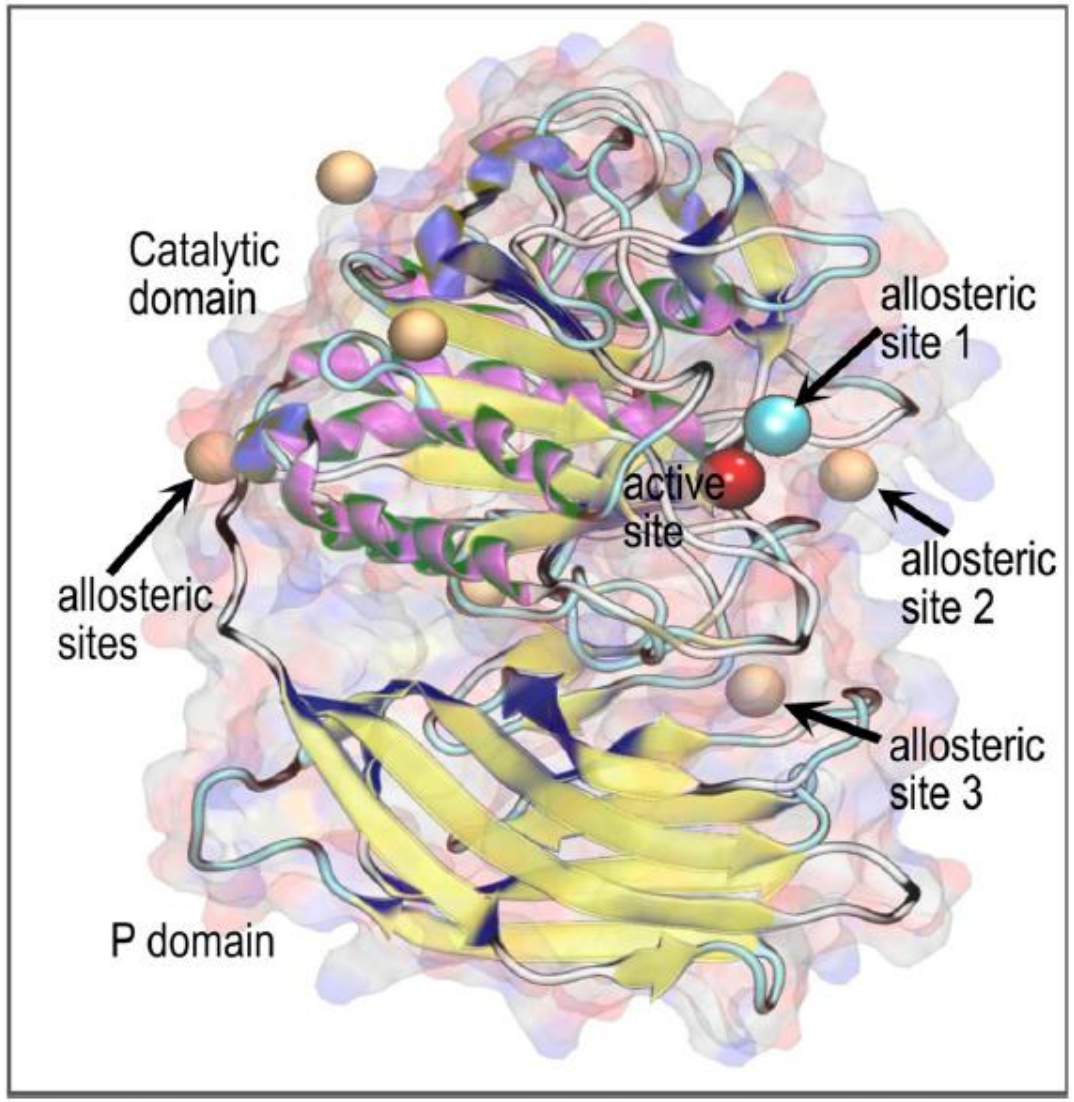



**Fig. 11.** Molecular representation of PC2. Overall view. Spheres represent the locations of binding sites: red, active site; orange, allosteric binding sites; and blue, a key allosteric binding site.

In loving memory of Dr. Richard Houghten

**Context**

*Relevant predictive limitations are introduced before modeling begins: context matters*

# Bioactivity as a function of structure

SAR

- QSAR and ML models are formulated in terms of structure–activity landscapes.
- Smoothness is assumed or hoped for, not guaranteed.
- **This view implicitly assumes a fixed experimental context.**

# Context matters: a conceptual blind spot

- The same molecules can behave differently under different contexts
- Ignoring context collapses distinct biological scenarios into a single readout

- Examples of **contextual variables**:
  - Dose
  - Assay type and experimental readout
  - Time, temperature, and medium composition

*Context is not noise – it is part of the signal*

# Predictive behavior is not uniform across predictive space

- Predictive models can be stable and well-behaved within a well-defined predictive space.
- Changes in experimental context induce a different predictive space.
- Inadvertently mixing readouts, by ignoring experimental contexts effectively merges distinct predictive spaces.
- Apparent performance breakdowns may reflect transitions between predictive spaces, due to contextual or mechanistic shifts, not technical flaws.

*Experimental context is critical*

The original formulation of these ideas is discussed in this work




*Drug Discov. Today* **2025**, *30*, 10

PNAS RESEARCH ARTICLE | IMMUNOLOGY AND INFLAMMATION

## Agonist concentration–dependent c
## conformation lead to biased signalir
## phagocyte functions

Junlin Wang[a] and Richard D. Ye[a,1]

Edited by Jason Cyster, HHMI, University of California, San Francisco, CA; received Janu

ACS FOOD SCIENCE & TECHNOLOGY

Contents lists available at ScienceDirect

Biochemical Engineering Journal

journal homepage: www.elsevier.com/locate/bej

ELSEVIER

Regular article

## Divergent effects of copper oxide nanoparticles on methanotrophs: Stimulation at low concentrations and inhibition at high concentrations

Ji Sun[a,1], Wei-Jia Yu[a,1], Xin Li[a], Xu-Yan Zhu[a], Jia-Chang Pi[a], Chen Di[a], Xin-Ya Tan[a], Na Li[c], Guang-Can Zhu[a,b], Yong-Ze Lu[a,b,*]

Article

## f Hesperetin and
## lant, and Anti-

Annals of Clinical Microbiology and Antimicrobials

Open Access

Check for updates

Journal of Fungi

Article

## In Vitro–In Vivo Correlation of Posaconazole–Amphotericin Combination against *Candida albicans*: In Vitro Interacting Concentrati

Biomass Conversion and Bi
https://doi.org/10.1007/s133

ORIGINAL ARTICLE

Contents lists available at ScienceDirect

International Immunopharmacology

journal homepage: www.elsevier.com/locate/intimp

ELSEVIER

## Potential antivirulence activity of sub-inhibitory concentrations of ciprofloxacin
## nst *Proteus mirabilis* isolates: an in-vitro
## n-vivo study

A. Elhosseini[1,2*], Tarek F. El Bassal, Fatma I. Sonbol and Maisra M. El Bouseary[1*]

eived: 4 July 2024 | Revised: 6 August 2024 | Accepted: 22 August 2024
: 10.1002/jat.4694

SEARCH ARTICLE

Journal of Applied Toxicology WILEY

## Endo-xylanase
## cobs, its struct
## antioxidant ac

Ishu Khangwal[1] · Sur

## Opposite effects of sevoflurane at high and low concentration on the airway inflammation in juvenile asthma rats: A TMT-based proteomics analysis study

Feng-Lin Wang[a,d], Yan-Nan Zhou[a], Guang-Ting Zhang[a], Yin Ran[a], Xin-Xin Yang[a], Lin Zhou[a], Zhao-Qiong Zhu[a], Yu Zhang[b,*], De-Xing Liu[a,c,*]

## oncentration-dependent effects of the nerve agents
## yclosarin and VX on cytochrome P450 in a HepaRG cell-based
## ver model

abriele Horn | Franziska Frielingsdorf | Tobias Demel | Simone Rothmiller | Franz Worek | Niko Amend

# Conceptual questions lead to modeling frameworks

## Conceptual observation

- We observed that the same molecule can produce different outcomes, suggesting the coexistence of distinct endpoints.
- This points to contextual effects beyond molecular descriptors.
- We hypothesize that apparent modeling “failures” may arise from implicitly mixing endpoints rather than from model limitations.

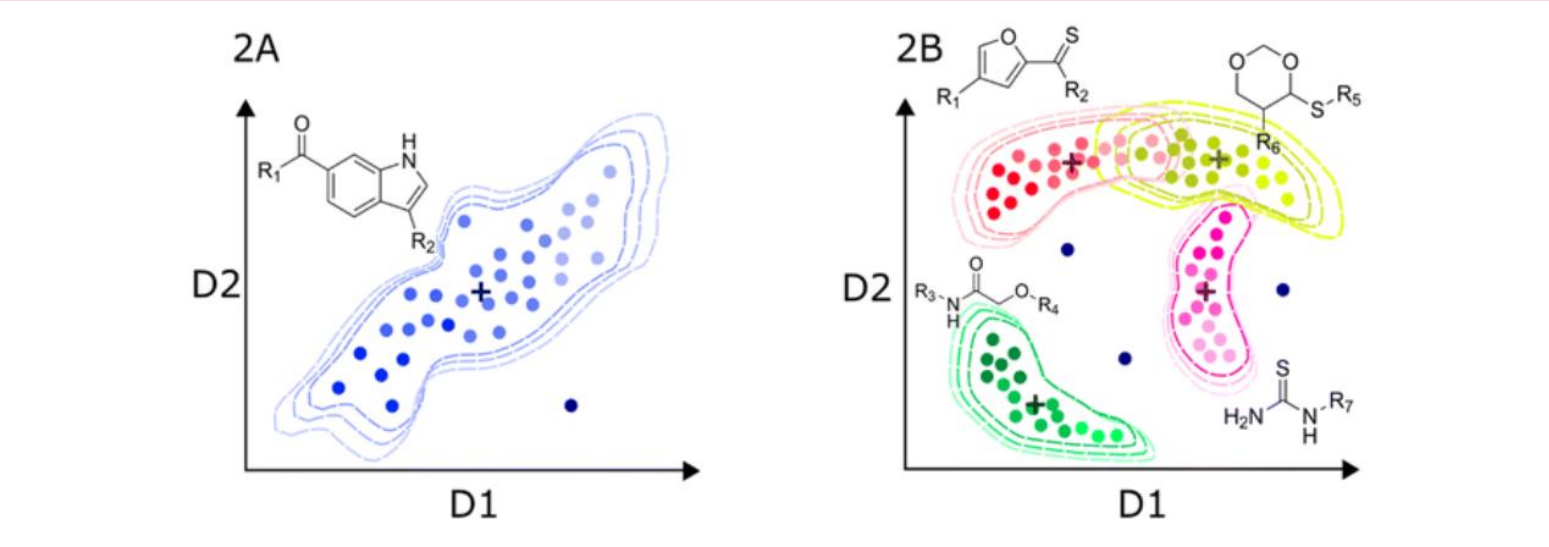
## Experimental evidence

- The literature reports a variety of examples consistent with endpoint-dependent behavior.
- These effects are often dose- or context-dependent.
- Such transitions can occur for the same molecule, leading to **bioactivity disruption**

## Mathematical framing

- What is the appropriate mathematical framework?
- In statistical terms, these effects are captured as contextual covariates.
- Here, we reframe them as elements defining predictive space.

# CHEMICAL SPACE vs PREDICTIVE SPACE




Organizes molecules based on general-uninformed representations

Supports inference about a specific property

*They are not equivalent*

*Predictive space is endpoint-dependent, therefore context-defined*

If you:

- change the dose range
- add cofactors
- or modify other contextual variable

…. you might be changing the endpoint, and therefore the predictive space must be redefined



Endpoint A



Predictive space A

Endpoint T



Predictive space T

Chemical space

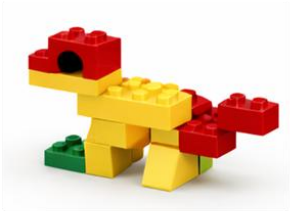


Endpoint A
Endpoint T



Context variable

Endpoint A



Predictive space A

Endpoint T



Predictive space T

Chemical space

***Context defines the endpoint and the predictive space***

# Practical next steps

1. Explicit endpoint documentation.
2. Context stratification before modeling.
3. Multi-level modeling (hierarchical).
4. Dose-aware modeling pipelines.

Rather than conclusions, I will delineate emerging issues

Many challenges persist despite technical advances

These challenges point to conceptual and semantic issues

*Some problems require reflection, not optimization*

# Do property cliffs exist?

- Property cliffs can exist
- They are rare
- They are often chemically interpretable

Properties behave as continuous, additive functions of structure governed by deterministic, incremental contributions of structural features.

- Structural activity cliffs are relatively frequent
- They are often difficult to rationalize
- They strongly impact predictivity

Emerge from complex, non-linear, and threshold-sensitive processes

Courtesy of Elier Abreu Martínez

***Not all cliffs have the same epistemic status***

Frontiers in Drug Discovery, 2026 in press

# Questioning *S* and R in *SAR*

- SPA formalizes the link between molecular representation and property behavior
- Properties emerge from organized patterns in chemical space, not isolated descriptors
- Predictive performance depends on how structure, representation, and context are aligned
- SPA highlights that different properties induce different effective spaces
- Apparent predictive issues often reflect misaligned representations, not model failure

## Structure–Property Associations (SPA)

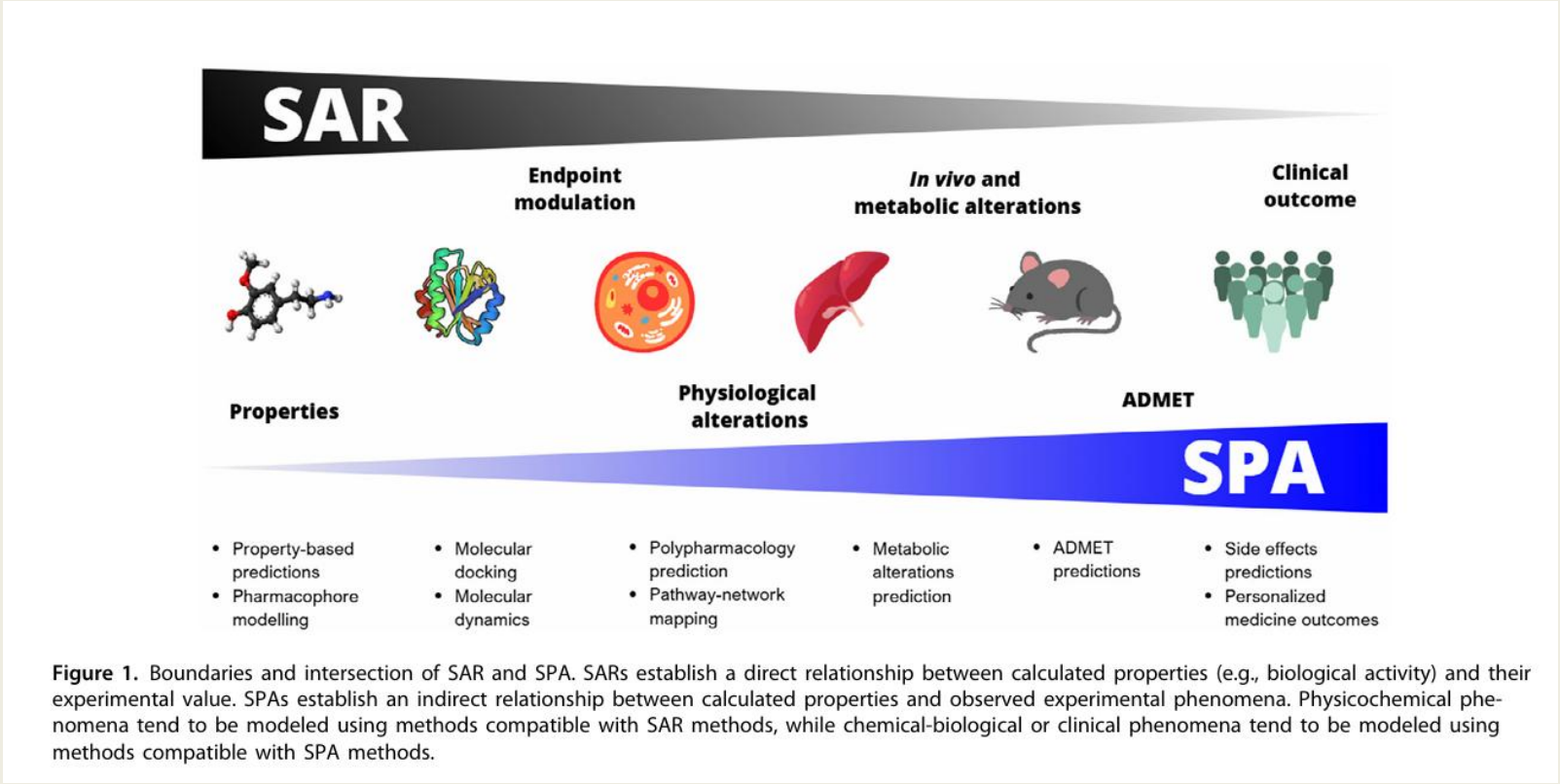
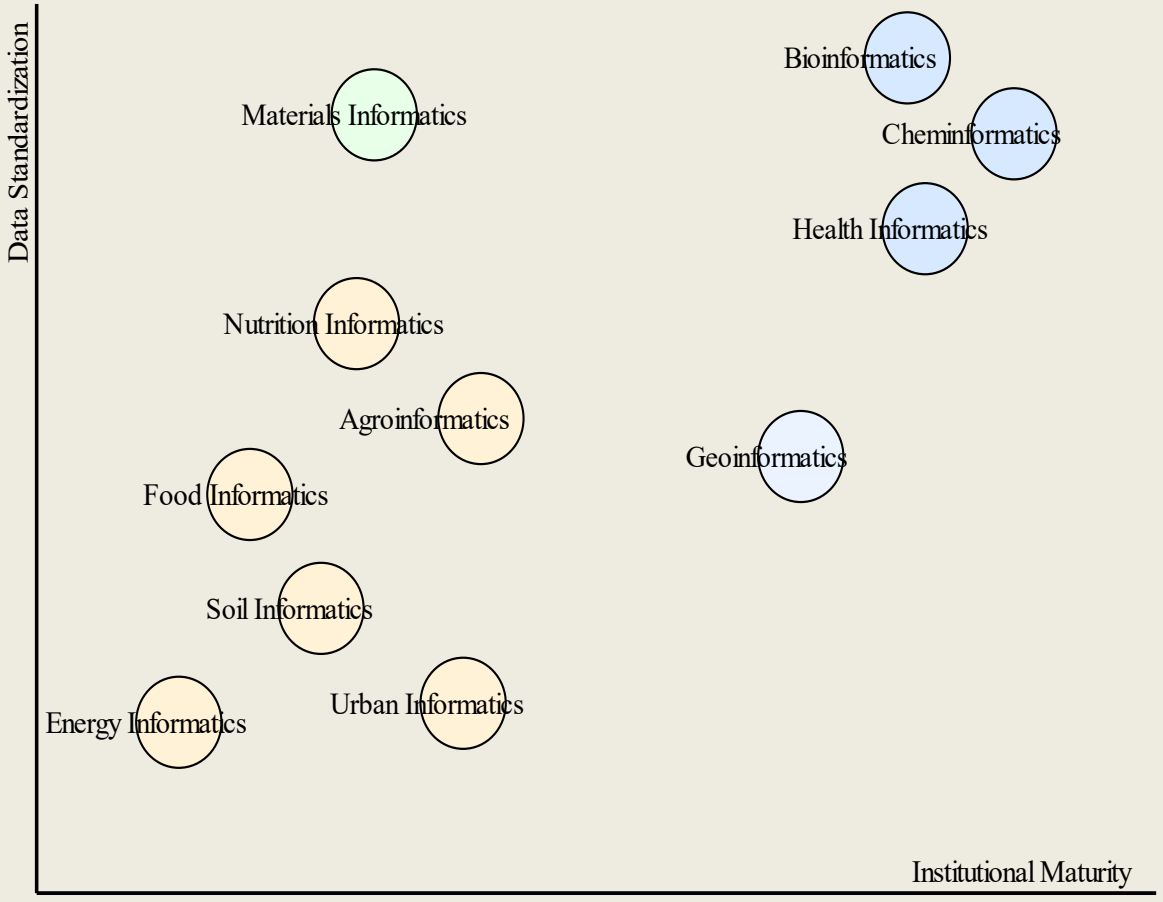



**Figure 1.** Boundaries and intersection of SAR and SPA. SARs establish a direct relationship between calculated properties (e.g., biological activity) and their experimental value. SPAs establish an indirect relationship between calculated properties and observed experimental phenomena. Physicochemical phenomena tend to be modeled using methods compatible with SAR methods, while chemical-biological or clinical phenomena tend to be modeled using methods compatible with SPA methods.

*ChemMedChem* **2026**, *21*, e202500847

# X-informatics: Conceptual fragmentation – terminology and semantics




- Proliferation of *X-informatics* labels
- Similar problems addressed under different names
- Risk of fragmentation and loss of shared language

- What remains to be done:
  - Build a comparative typology that reveals shared epistemic and infrastructural patterns across established and emerging informatics domains
  - Propose criteria and guidelines to assess the legitimacy, maturity, and interoperability potential of new X-informatics proposal

***Are we gaining clarity, or just new labels?***

Courtesy of Elier Abreu Martínez

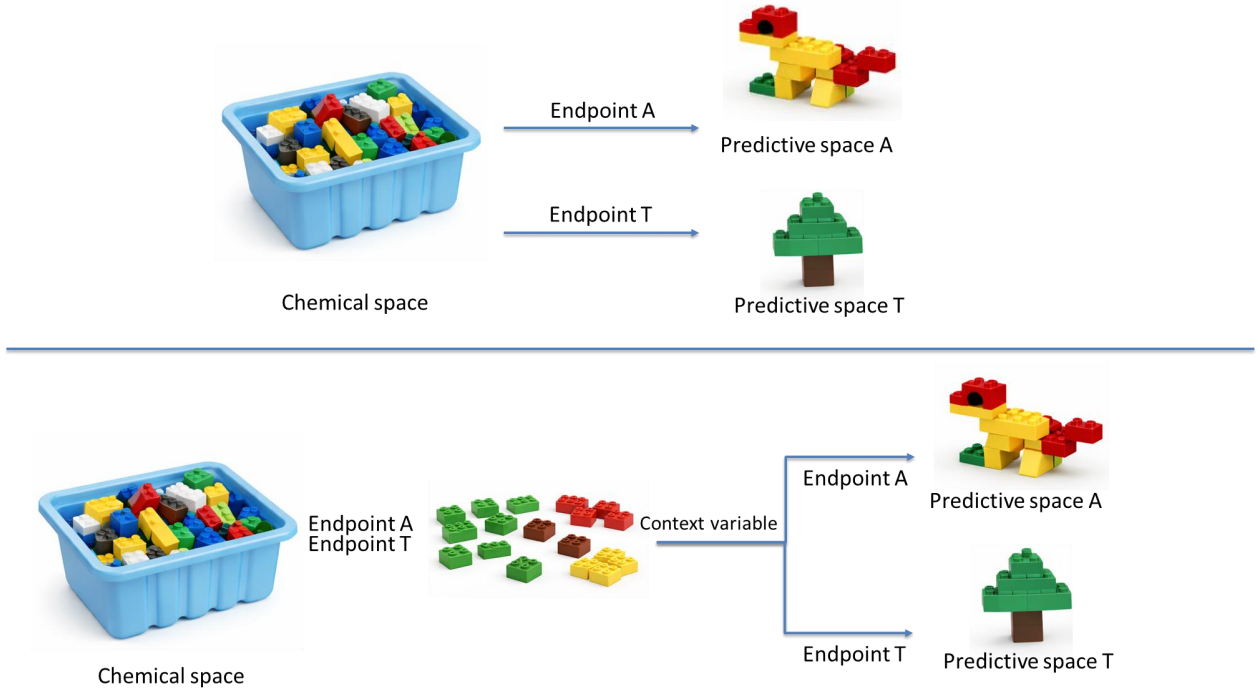
# What do we really mean by AI-*XYZ*?

- AI-driven, AI-based, and AI-supported are often used interchangeably
- Using AI tools does not imply the development of AI research

- Using a force field parameterized with quantum data does *not* mean doing quantum chemistry

***Precision in terminology matters***

# Take home message:
# Context is not noise – it defines the problem




Apparent failures may arise from

- how molecules are represented
- how endpoints are defined
- and what models are asked to explain

# Cheminformatics Network

https://sqm.org.mx/convocatorias-redes-especializadas-redes-existentes/

En fase de formación

Integra la química, la ciencia de datos, la inteligencia artificial y el modelado molecular para el análisis y diseño de sistemas químicos y biológicos, con aplicaciones en el descubrimiento de fármacos, materiales, agroquímicos y sostenibilidad. Fomenta la interacción entre investigadores, docentes, estudiantes y el sector industrial mediante cursos, talleres y la Escuela Latinoamericana de Quimioinformática.




Currently under development

It integrates chemistry, data science, artificial intelligence, and molecular modeling for the analysis and design of chemical and biological systems, with applications in the discovery of drugs, materials, and agrochemicals, as well as in sustainability. It fosters interaction among researchers, educators, students, and the industrial sector through courses, workshops, and the School of Cheminformatics Latin America.