Toward Connected Biological Digital Twins

What can be connected when biological datasets share nothing in common?

KEY IDEA

- Many biological datasets share few or no compounds.
- As a result, potentially useful evidence often remains isolated.
- This study explores whether disconnected datasets can be connected before prediction.

0 Shared Compounds

1,500 Potential Connections

WHAT WE FOUND

- Evidence bridges emerged despite zero direct overlap.
- Shared chemical neighbourhoods were identified across assay domains.
- Information from previously isolated datasets could be explored within a common framework.
- Evidence connectivity may help organise information before modelling.

Exploratory Analysis

Estimated reading time: 8–10 minutes

AvatarsBio

Toward Connected Biological Digital Twins

Exploring Evidence Bridges Between Biological Assay Domains

Chee Ping Ng, Founder
AvatarsBio, The Netherlands,

INTRODUCTION

Biological datasets are often fragmented. A toxicity study may contain thousands of compounds. A permeability study may contain hundreds more. A bioavailability dataset may provide additional context. Yet these resources frequently share few compounds and are therefore analysed independently. As a result, biological evidence often remains isolated within individual datasets.

Most modern AI approaches attempt to address this problem through predictive modelling. Large datasets are collected, models are trained, and predictions are generated for unseen compounds.

This study explores a different question:

> ***Can disconnected biological datasets be connected before prediction?***

This study does not introduce a new similarity algorithm. Rather, it explores similarity-based evidence bridging as a workflow for connecting disconnected biological datasets before prediction. The objective is not to establish biological transfer, predictive equivalence, or cross-assay validation, but to investigate whether fragmented evidence can be organised into interpretable connectivity structures across assay domains.

300	916	1,500
HepG2 compounds	Caco-2 compounds	bridge pairs

0	0.795	8.7%
direct compound overlap	strongest evidence bridge	PAINS flagged

Using HepG2 toxicity data and Caco-2 permeability data as a proof of concept, we investigated whether chemical similarity could create interpretable evidence bridges between assay domains despite the absence of direct compound overlap.

DATA SOURCES AND WORKFLOW

HepG2 viability data from ChEMBL3507681 and a curated Caco-2 permeability dataset assembled from public ChEMBL records were used. A total of 300 HepG2 compounds were selected as bridge seeds and compared against 916 Caco-2 compounds. Morgan fingerprints and Tanimoto similarity were used to identify the five nearest Caco-2 neighbours for each HepG2 compound, generating 1,500 candidate bridge pairs for subsequent analysis.

THE OVERLAP PROBLEM

The datasets examined in this study represented related aspects of compound behaviour but contained no overlapping compounds.

Exact compound overlap between the HepG2 and Caco-2 domains was zero.

Under a conventional integration workflow, this would largely prevent direct comparison or enrichment across datasets. Without shared compounds, the two datasets would typically be analysed independently, limiting opportunities to exchange context or evidence.

This raises a practical question:

Does the absence of overlap mean the datasets cannot inform one another?

BUILDING AN EVIDENCE BRIDGE

To explore this question, 300 HepG2 compounds were selected as bridge seeds.

For each compound, the five nearest neighbours within the Caco-2 dataset were identified using chemical similarity, producing 1,500 candidate bridge pairs [1]. The resulting bridge network can be viewed as a collection of local evidence neighbourhoods, where each HepG2 compound is connected to its nearest Caco-2 neighbours (Figure 1).

Additional biological and physicochemical information was then enriched across the resulting bridge network.

The objective was not to predict new values or infer biological equivalence between assays. Instead, the workflow sought to organise fragmented evidence into interpretable connectivity structures that could support exploration, prioritisation, and human interpretation.

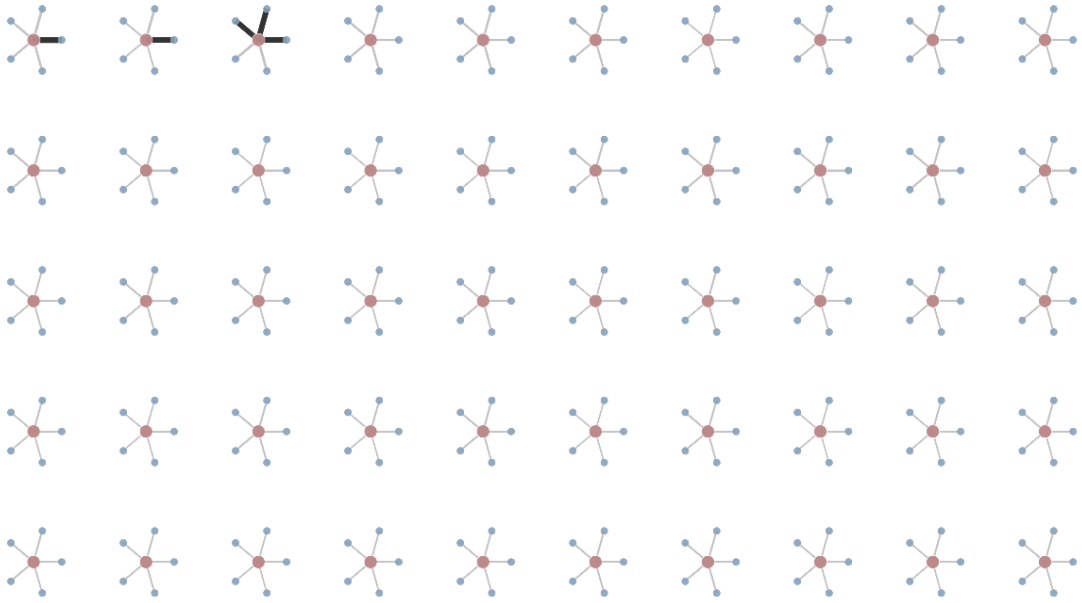

Figure 1. Conceptual representation of local evidence neighbourhoods generated from the HepG2–Caco-2 bridge workflow. Each HepG2 compound (red) is connected to its five nearest Caco-2 neighbours (blue) based on chemical similarity. The visualisation illustrates how local evidence neighbourhoods can be constructed despite the absence of direct compound overlap between datasets. Connections represent similarity-based evidence bridges rather than biological prediction.

KEY OBSERVATIONS

Several observations emerged.

1. Most bridges showed modest similarity

Most candidate bridges showed only modest similarity. The average Tanimoto similarity was 0.291, suggesting that most compounds occupied related but distinct regions of chemical space.

2. A small number of strong bridges emerged

Despite the complete absence of shared compounds, a small number of highly similar compound pairs were identified between the HepG2 and Caco-2 domains. Among 1,500 bridge pairs, five candidates exceeded a Tanimoto similarity of 0.60, including one pair with a similarity of 0.795. These compounds provided potential evidence bridges between previously isolated datasets.

3. Shared chemical space despite zero overlap

Projection of HepG2 and Caco-2 compounds into a shared chemical space revealed substantial overlap despite zero direct compound overlap (Figure 2). Rather than forming two completely separated chemical islands, compounds from both domains occupied many of the same regions of chemical space, suggesting the presence of shared chemical neighbourhoods.

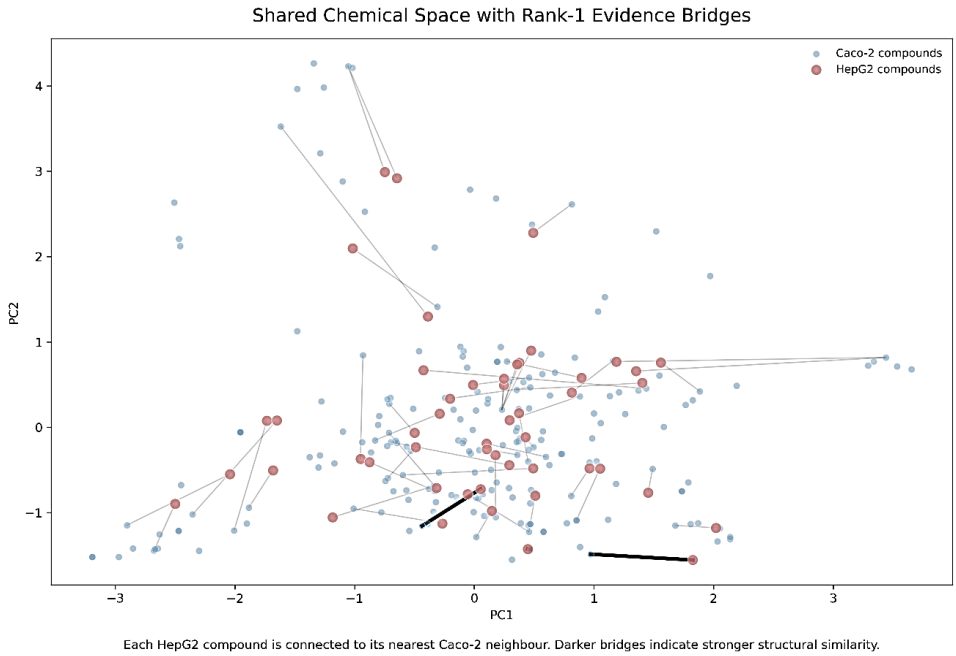

Figure 2. Shared chemical space with rank-1 evidence bridges. HepG2 compounds (red) and Caco-2 compounds (blue) were visualised within a shared chemical space using molecular similarity descriptors and principal component analysis (PCA). Each HepG2 compound was connected to its nearest Caco-2 neighbour using chemical similarity. Faint lines represent rank-1 evidence bridges, while darker lines highlight the strongest bridge candidates (Tanimoto ≥ 0.60). The PCA projection is intended as an illustrative visualisation of shared chemical space rather than a quantitative measure of chemical similarity or coverage.

4. The strongest bridges are chemically coherent

The strongest bridge candidates appeared chemically coherent. Several converged on the same local chemical neighbourhood, suggesting that similarity-based workflows may identify interpretable evidence clusters rather than isolated pairwise matches. Many of the strongest candidates belonged to related flavonoid-like or polyphenolic chemical families.

Interestingly, these compounds were not highly permeable molecules. Instead, they occupied a moderately permeable and relatively polar region of chemical space. This suggests that the observed bridges were driven primarily by structural similarity rather than a shared permeability profile.

5. PAINS sensitivity analysis

To assess whether the strongest bridge candidates were dominated by known assay-interference chemotypes, a PAINS (Pan-Assay Interference Compounds) sensitivity analysis was performed using RDKit filters. Among 1,500 bridge pairs, 130 pairs (8.7%) contained at least one PAINS alert. Among the five strongest bridge candidates, a single catechol-associated PAINS alert (catechol_A) was identified, while the remaining highlighted examples showed no PAINS flags. These findings suggest that the bridge network is not primarily driven by PAINS-associated chemotypes, although standard caution remains appropriate during downstream compound prioritisation.

6. Biological evidence could be explored across domains

Information that previously existed in separate toxicity and permeability datasets could be viewed within a common framework, allowing related observations to be examined together.

DISCUSSION

The purpose of this study was not to establish drug repositioning, cross-tissue prediction, or biological equivalence between assay systems.

Instead, the study explores whether chemical similarity can serve as a practical mechanism for connecting fragmented biological evidence across disconnected assay domains. Although direct compound overlap was absent, structurally related compounds may still provide useful context, a principle that has long been used in medicinal chemistry and similarity-based discovery workflows [2].

As illustrated in Figure 3, many modelling workflows follow a familiar pattern in which information flows directly from a dataset to a predictive model.

The present work explores an alternative pathway. Rather than moving immediately to predictive modelling, the workflow first seeks to establish evidence connectivity between otherwise disconnected biological datasets. These connections can then be explored and interpreted before any prediction is made.

The objective is not to replace modelling, but to introduce an intermediate step in which potentially relevant biological context can be identified, connected, and examined.

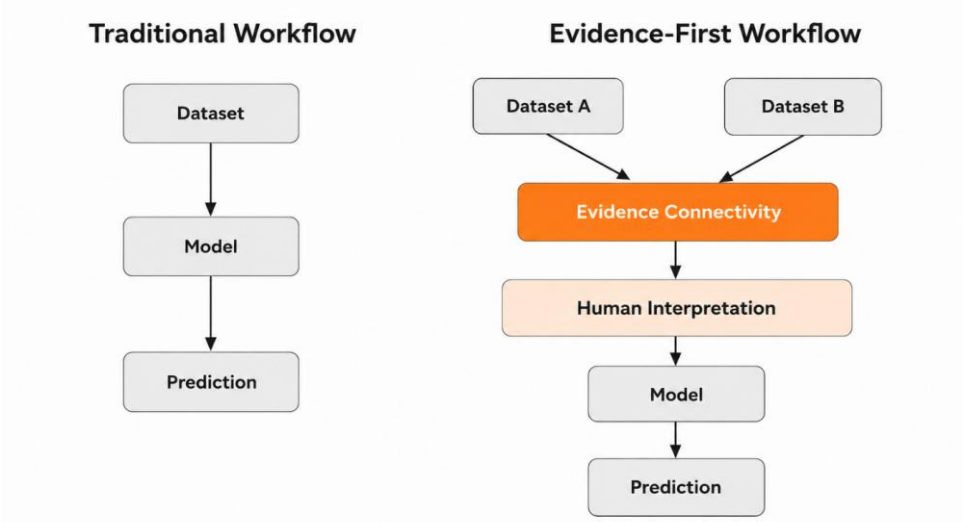

Figure 3. Traditional prediction workflow versus evidence-first workflow. The proposed workflow introduces an evidence bridge that allows fragmented biological information to be connected and interpreted before predictive modelling.

An alternative interpretation should also be considered. Structural similarity may reflect shared chemotypes without implying biological transfer between assay systems. Consequently, evidence bridges should be viewed as hypothesis-generating connections rather than demonstrations of biological equivalence. Additional validation would be required to determine whether structural connectivity consistently translates into biological relevance.

Visualisation of the bridge network provided an additional perspective. Although the datasets contained no overlapping compounds, HepG2 and Caco-2 compounds occupied broadly overlapping regions of chemical space. In this context, similarity-based bridges did not necessarily

connect distant chemical islands. Instead, they highlighted local evidence relationships within a shared landscape, allowing potentially relevant biological context to be identified and explored.

More broadly, the workflow raises an interesting question beyond the specific HepG2–Caco-2 example. Biological assays are often treated as isolated sources of evidence, particularly when direct compound overlap is absent. However, structurally related compounds may provide opportunities for context to be explored across assay boundaries. Whether different biological assays can systematically exchange useful context through such evidence bridges remains an open question for future investigation.

The concept may also have implications for future biological digital twins. Rather than viewing digital twins as increasingly sophisticated models of individual biological systems, an alternative possibility is that future biological twins may function as connected networks of specialised models.

One possible interpretation is that evidence connectivity represents only the first step in a broader progression. Structural connectivity may eventually support biological coherence, robustness across datasets, and ultimately multi-domain integration. In this view, future biological digital twins may function not as isolated models, but as connected networks capable of exchanging context, evidence, and hypotheses.

The present study addresses only the first step. It does not demonstrate connected biological digital twins. Rather, it explores one possible mechanism through which future biological models may exchange context, evidence, and hypotheses.

CONCLUSION

This work forms part of a broader exploration of workflows that seek to reduce uncertainty before modelling, complementing earlier work on experimental design and data curation [3].

This proof-of-concept study explored whether chemical similarity could connect biological evidence across disconnected assay domains. Despite zero direct compound overlap, a small number of chemically coherent bridge candidates emerged, and information from previously isolated toxicity and permeability datasets could be explored within a common framework.

The findings suggest that disconnected biological datasets may contain more opportunities for connection than direct overlap alone would imply. Importantly, the value of such connections may not lie solely in prediction. Evidence connectivity may also help organise information, generate hypotheses, and support scientific decision-making before predictive models are applied.

The present study addresses only the first step toward connected biological digital twins. It explores one possible mechanism through which future biological models may exchange context, evidence, and hypotheses. More broadly, it proposes that prediction may not be the only way to extract value from scientific data.

> ***Before predicting, it may first be useful to connect.***

REFERENCES

1. Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling. 2010. 50(5), 742–754. https://doi.org/10.1021/ci100050t

2. Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular Similarity in Medicinal Chemistry. Journal of Medicinal Chemistry. 2014. 57(8), 3186–3204. https://doi.org/10.1021/jm401411z

3. Ng CP. Design Before Prediction: Reducing Complexity in Biological Models. AvatarsBio White Paper. 2026. https://avatarsbio.com/research/pdfs/design-before-prediction-doe-ml.pdf