ARNAUD Predict

Transparent and explainable permeability predictions

Has a compound similar to yours already been tested?

KEY IDEA

• Retrieve experimentally measured analogue compounds using molecular similarity.
• Estimate permeability from the experimental evidence.
• Present confidence and supporting evidence alongside every prediction.

Evidence by similarity, not black-box prediction.

WHAT WE FOUND

• Exact compounds and close analogues consistently retrieved chemically plausible experimental neighbours.
• Prediction confidence decreased as structural similarity declined.
• Every prediction could be traced directly to supporting experimental compounds.

Proof-of-Concept Study

Estimated reading time: 5 minutes

AvatarsBio

ARNAUD Predict

A Transparent Similarity-Based Approach to Caco-2 Permeability Estimation

Chee Ping Ng, Founder
AvatarsBio, The Netherlands,

Evidence by similarity, not black-box prediction.

INTRODUCTION

Computational approaches such as quantitative structure–activity relationship (QSAR) models and machine learning have become increasingly important for predicting molecular properties, including Caco-2 permeability [1,2]. These methods can provide valuable predictions, particularly when large datasets are available.

ARNAUD Predict explores a complementary question.

Rather than learning a predictive model from experimental data, can experimentally measured analogue compounds be used to provide a transparent, evidence-based estimate of permeability?

Given a single compound represented as a SMILES string, the prototype searches a curated experimental Caco-2 dataset for chemically similar compounds, estimates permeability from the closest experimental neighbours, and presents the supporting evidence used to generate the estimate.

The objective is not to replace QSAR or machine learning models, but to investigate an alternative workflow that is transparent, easy to inspect, and grounded in experimentally measured compounds.

ARNAUD Predict is currently implemented as an offline proof-of-concept research prototype intended to investigate this workflow.

CONCEPT

ARNAUD Predict follows a simple evidence-based workflow (Figure 1) consisting of six steps:

- **Compound Check** validates the input structure and calculates molecular descriptors using RDKit [3] before comparing them with the curated experimental dataset.

- **Similarity Search** identifies chemically similar compounds using established molecular similarity methods based on Morgan fingerprints and the Tanimoto similarity coefficient [4-6].
- **Experimental Neighbours** retrieves the closest matching compounds and their experimental measurements.
- **Permeability Estimate** calculates an estimate using either the closest neighbour or a similarity-weighted average of the top five neighbours.
- **Confidence Assessment** indicates how well the query compound is represented by the available experimental evidence.
- **Reasoning** presents the experimental compounds supporting the estimate so users can inspect the underlying evidence.

The emphasis of the workflow is on presenting experimental evidence rather than algorithmic complexity.

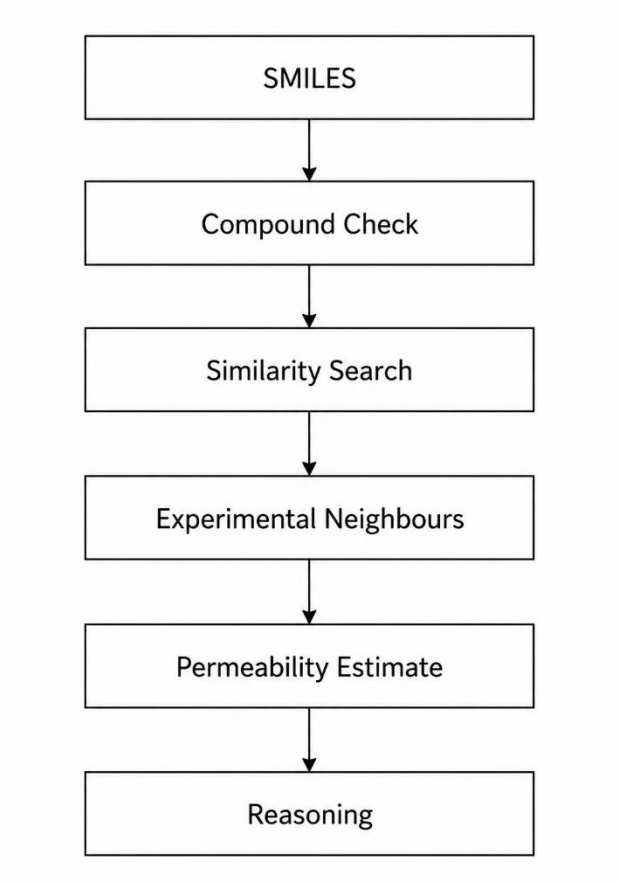

Figure 1. Overview of the ARNAUD Predict workflow.

EXAMPLE

Using aspirin as an example (Figure 2), the prototype (ARNAUD Predict v0.3) validates the input compound, retrieves the nearest experimentally measured analogue compounds, estimates permeability, and reports an associated confidence level. Because aspirin is present in the curated

dataset, it is identified as an exact match (similarity = 1.00). This illustrates how the workflow links permeability estimates directly to supporting experimental evidence.

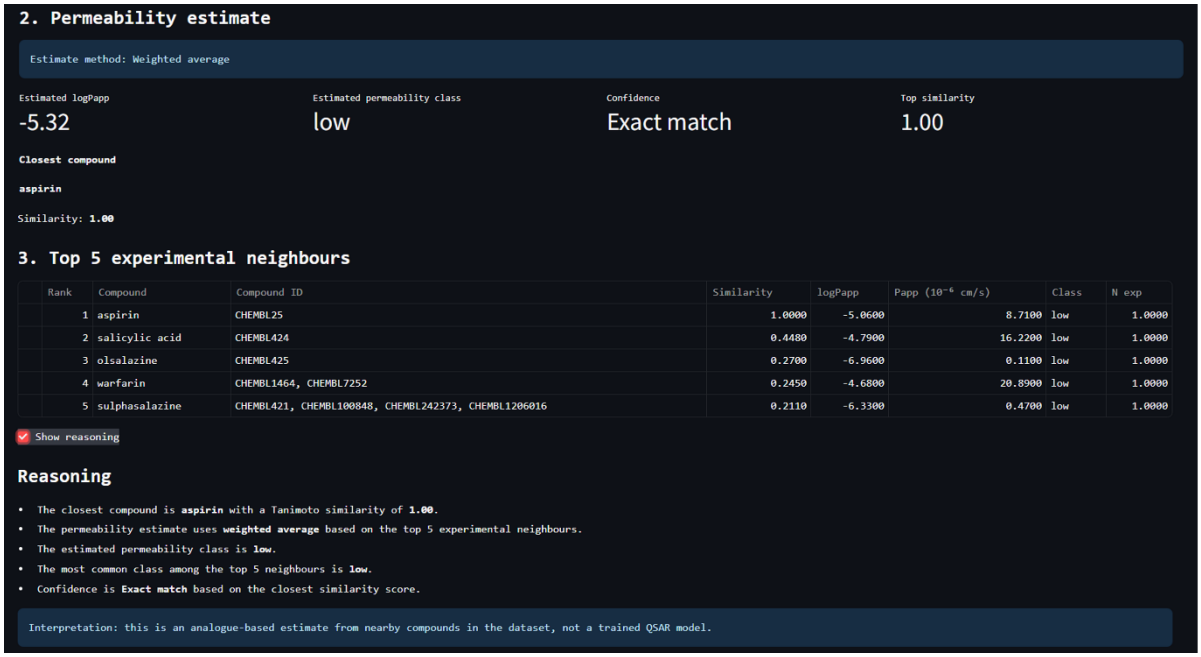

Figure 2. Example permeability estimate for aspirin generated by ARNAUD Predict v0.3, showing confidence assessment and supporting experimental analogue compounds.

DISCUSSION

QSAR models and machine learning methods remain powerful approaches for predicting molecular properties and have become widely used throughout drug discovery and development.

ARNAUD Predict is not intended to compete with or replace these approaches. Instead, it explores a complementary workflow centred on experimental analogue compounds.

Compared with predictive models, the prototype offers several practical advantages:

- Transparent evidence-based workflow
- Supporting compounds can be inspected directly
- Confidence communicated alongside the estimate
- No model training required
- Lightweight offline implementation
- Easy integration with curated internal datasets

The approach has important limitations.

During development, representative compounds—including exact matches, pharmaceutical salt forms, isotopically labelled compounds, metabolites, structural analogues, and unrelated molecules—were evaluated qualitatively.

The performance of similarity-based estimation depends on the quality and diversity of the underlying experimental dataset. Compounds that differ substantially from available experimental examples naturally result in lower confidence. The current proof of concept also uses a small, curated dataset and is intended to demonstrate the workflow rather than establish predictive superiority over validated QSAR or machine learning models.

Future work may focus on evaluating larger curated datasets, additional permeability assays, and hybrid workflows combining analogue evidence with predictive models.

Recent work in explainable AI has focused on improving the interpretation of predictive models. ARNAUD Predict takes a different approach to explainability. Rather than explaining how a trained model reached a prediction, it presents the experimentally measured analogue compounds that support an estimate, allowing users to inspect the underlying evidence directly.

CONCLUSION

ARNAUD Predict demonstrates a transparent similarity-based workflow for Caco-2 permeability estimation.

By combining compound validation, similarity search, experimental analogue retrieval, confidence assessment, and evidence-based reasoning, the prototype illustrates how permeability estimation can be presented as inspectable workflow rather than solely as a numerical prediction.

Although currently implemented as a proof of concept, the approach provides a foundation for future decision-support tools based on curated experimental evidence.

REFERENCES

1. Larregieu CA, Benet LZ. Drug discovery and regulatory considerations for improving in silico and in vitro predictions that use Caco-2 as a surrogate for human intestinal permeability measurements. AAPS J. 2013;15:483–497. https://doi.org/10.1208/s12248-013-9456-8

2. Falcón-Cano G, Molina C, Cabrera-Pérez MÁ. Reliable prediction of Caco-2 permeability by supervised recursive machine learning approaches. Pharmaceutics. 2022;14(10):1998. https://doi.org/10.3390/pharmaceutics14101998

3. RDKit: Open-source cheminformatics. https://www.rdkit.org. Accessed July 2026.

4. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–754. https://doi.org/10.1021/ci100050t

5. Tanimoto TT. An elementary mathematical theory of classification and prediction. New York: International Business Machines Corporation; 1958.

6. Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today. 2006;11(23–24):1046–1053. https://doi.org/10.1016/j.drudis.2006.10.005