



# A Reproducible Pipeline for Scaffold-Hopping Across Public Compound Libraries

Liam Larsen<sup>1\*</sup>, Tuan Jovanovic<sup>2</sup>, Kevin Ribeiro<sup>3</sup>

<sup>1</sup>Department of Life Sciences, Zambezi Institute of Life Sciences, Kyrgyzstan

<sup>2</sup>Department of Life Sciences, Balkan Institute of Life Sciences, Philippines

<sup>3</sup>Department of Life Sciences, Mekong Center for Life Sciences, Trinidad and Tobago

## Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijbc.2026.v1.i1.pp14-19](https://doi.org/10.62715/ijbc.2026.v1.i1.pp14-19)

Received: 24 February 2026

Accepted: 28 April 2026

Published: 22 June 2026

## Abstract

Scaffold hopping, the search for a different molecular framework that presents similar features to a biological target, is a routine task in medicinal chemistry and an awkward one to reproduce. Published searches differ in the library snapshot used, in how structures were standardised, in which similarity measure was applied and in where the cut-off was placed, and those choices are often described too loosely for another group to repeat the work. This article proposes a pipeline that fixes each of those choices explicitly and records them as machine-readable provenance alongside the result. The procedure has seven stages: define the query framework, assemble and version the public libraries, standardise structures with a published curation sequence, enumerate candidate frameworks, score candidates with three complementary measures of similarity, apply chemistry filters, and publish the ranked list together with everything needed to regenerate it. Each stage is described with the decision it requires, the open software that can perform it and the record it should leave. The pipeline is a proposal assembled from published methods and curation guidance. It has not been applied or validated here, and the article states what would have to be demonstrated before it could be recommended without reservation.

**Keywords:** Scaffold hopping, Reproducibility, Chemical structure curation, Virtual screening, Public compound libraries, Provenance

© 2026 Liam Larsen et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## 1. Introduction

A scaffold hop replaces the central framework of an active compound while keeping the arrangement of groups that the target recognises. The motive is usually practical: to escape a patent, to improve metabolic stability, to shed a liability carried by the original core. The term was given its current sense by [Schneider et al. \(1999\)](#), who searched for topologically distinct molecules with matching pharmacophore features, and the strategy was reviewed as a discipline of its own by [Bohm et al. \(2004\)](#).

Public compound collections have made the search cheap. PubChem ([Kim et al., 2023](#)), ChEMBL ([Gaulton et al., 2012](#); [Mendez et al., 2019](#)) and ZINC ([Irwin et al., 2020](#)) together offer hundreds of millions

\* Corresponding author: Liam Larsen, Department of Life Sciences, Zambezi Institute of Life Sciences, Kyrgyzstan. E-mail: [liam.larsen@zambezi.edu](mailto:liam.larsen@zambezi.edu)

© 2026 Liam Larsen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

of structures with no licence cost. What has not kept pace is the reproducibility of searches made in them. Two groups starting from the same active compound and the same nominal library routinely produce different hit lists, and the reasons are mundane rather than scientific: a different database release, a different treatment of salts and tautomers, a different fingerprint radius, an unstated threshold.

This article proposes a pipeline designed so that a scaffold-hopping search can be repeated exactly by a reader with no access to the original authors. Nothing in it is novel as chemistry; the contribution is the assembly, the explicit recording of every choice that affects the result, and the attention given to the points where published searches most often become irreproducible. The pipeline is a proposal. It has not been run on a real programme here, and no results are reported.

## 2. Why Scaffold-Hopping Searches Resist Repetition

Three properties of the task make it fragile. The first is that the object being searched for is defined by negation. A useful hit must be similar in the features that matter and dissimilar in the framework, so the search depends on two thresholds pulling in opposite directions, and small movements in either reshape the output. The second is that the frameworks themselves must be derived from the structures, usually by the rule of Bemis and Murcko (1996), which strips side chains to leave ring systems and the linkers between them. That rule is deterministic, but what it is applied to is not: a structure with an unresolved tautomer or an attached counter-ion yields a different framework.

The third is that public libraries are moving targets. Records are added, corrected and withdrawn, and a search described as having been run against a named database is not reproducible unless the release is identified. The curation problem was set out plainly by Fourches et al. (2010), who showed how often errors in structure representation propagate into modelling results, and an open implementation of a defensible curation sequence is described by Bento et al. (2020).

## 3. Choosing Measures of Similarity

No single measure captures what a medicinal chemist means by a scaffold hop, so the pipeline uses three and treats disagreement between them as informative rather than as noise. Table 1 lists the measures with what each one treats as similar and the role it plays. Circular fingerprints (Rogers and Hahn, 2010) supply a topological baseline; a pharmacophore-type description captures the feature arrangement that motivated the original work of Schneider et al. (1999); and a three-dimensional shape comparison of the kind evaluated by Hawkins et al. (2007) tests whether the candidate can occupy the same volume as the original.

**Table 1: Measures of similarity used in the pipeline and the role of each**

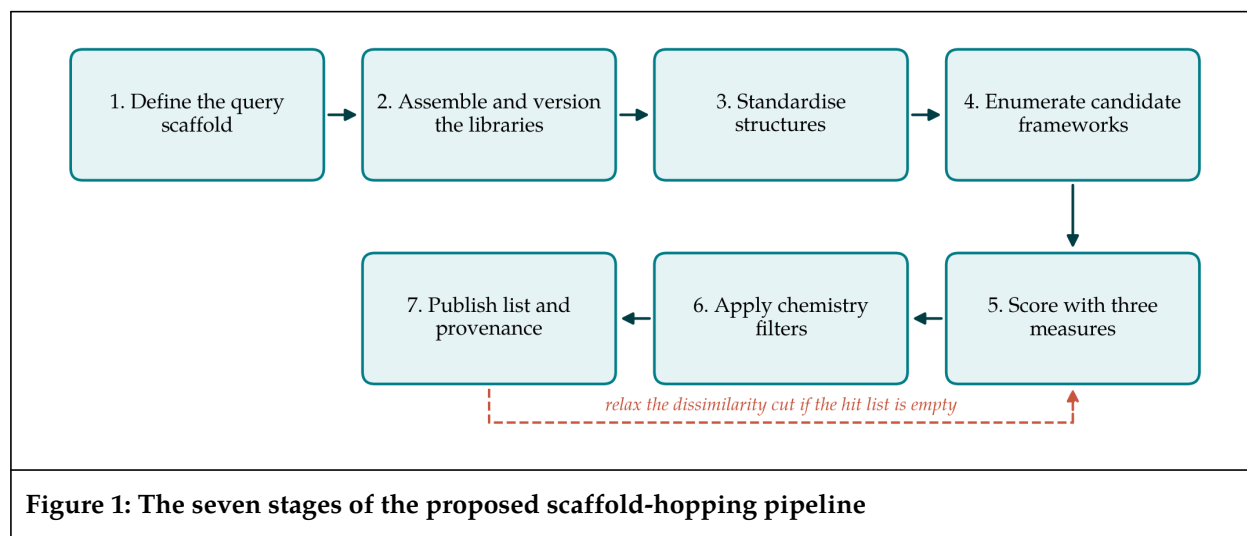
| Measure                        | What it treats as similar                        | Role in the pipeline                                                 |
|--------------------------------|--------------------------------------------------|----------------------------------------------------------------------|
| Circular fingerprint overlap   | Shared atom neighbourhoods up to a set radius    | Baseline; also used in reverse to enforce framework difference       |
| Murcko framework identity      | The same ring systems and linkers                | Defines what counts as a different scaffold (Bemis and Murcko, 1996) |
| Pharmacophore feature pairs    | Equivalent features at equivalent distances      | Carries the recognition hypothesis (Schneider et al., 1999)          |
| Shape and colour overlay       | Volume and feature overlap of aligned conformers | Three-dimensional check on the two-dimensional hits                  |
| Matched molecular pair context | Frameworks exchanged in known compound pairs     | Evidence that the replacement has precedent                          |
| Activity cliff awareness       | Close analogues with divergent potency           | Warns where similarity is a poor guide (Stumpfe and Bajorath, 2012)  |

The reverse use of the fingerprint deserves comment. A candidate is required to fall below a stated

topological similarity to the query, because a hit that is merely a close analogue is not a hop. Setting that lower bound is the single most consequential parameter in the whole procedure and the one most often left unstated in published work.

#### 4. The Proposed Pipeline

Seven stages follow, shown in Figure 1 and described in turn. Each stage writes a file, and the files together are the provenance record: given them, the search can be regenerated without reference to the people who ran it. Orchestration can use any workflow manager that records container images and parameter values, with a system such as the one described by [Di Tommaso et al. \(2017\)](#) a common choice, and software versions pinned through a package channel rather than described in prose.



##### 4.1. Stage 1: Define the query framework

The query is not a compound but a hypothesis about it. The chemist records the active structure, the framework obtained from it by the Bemis and Murcko rule, which atoms carry the features believed to be recognised by the target, and which part of the molecule is to be replaced. Writing this down separates the chemistry hypothesis from the computation, so that a disappointing hit list can be traced to the hypothesis rather than blamed on the software.

##### 4.2. Stage 2: Assemble and version the libraries

Each source is downloaded as a dated release, and the release identifier, the file checksum and the record count are stored. Where a subset is used, the selection query is stored with it rather than described in prose. Mixing sources is expected, since bioactivity-annotated collections and purchasable catalogues serve different purposes, but the origin of every record is kept so that a hit can be traced back to the database that supplied it.

##### 4.3. Stage 3: Standardise structures

Curation follows a fixed sequence: remove counter-ions and solvents, neutralise where the convention requires it, normalise functional group representations, choose a canonical tautomer, discard records that fail valence checks, and deduplicate on a canonical identifier. The standard InChI ([Heller et al., 2015](#)) is suitable as the deduplication key, and canonical SMILES ([Weininger, 1988](#)) as the working representation. The sequence itself matters: neutralise before tautomer selection and the outcome differs. The implementation of [Bento et al. \(2020\)](#) can be used as published, and whichever sequence is chosen it is recorded as an ordered list of operations rather than as the word curated.

##### 4.4. Stage 4: Enumerate candidate frameworks

Frameworks are extracted from every standardised structure and counted, which turns a library of

millions of compounds into a much smaller set of distinct cores. The search then runs over frameworks rather than over compounds, which is both faster and closer to the question being asked. Frameworks identical to the query are set aside at this point, as are those whose topological similarity to the query exceeds the stated upper bound.

#### **4.5. Stage 5: Score with three measures**

Each surviving framework, decorated with the substituents implied by the query hypothesis, is scored on the three measures of Table 1. Conformers for the shape comparison are generated with a stated method and a fixed random seed, since conformer generation is stochastic and an unseeded run cannot be repeated. The three scores are kept separate. Combining them into one number is convenient and discards the disagreement that is often the most interesting part of the output.

#### **4.6. Stage 6: Apply chemistry filters**

Filters remove what a chemist would reject on sight: reactive and unstable groups, structures outside the intended property range, and frequently flagged interference substructures of the kind catalogued by [Baell and Holloway \(2010\)](#). Those filters are heuristics and are applied as annotations rather than as silent deletions, so that a reviewer can see what was excluded and why. Commercial availability or synthetic accessibility can be annotated at the same stage without being used to rank.

#### **4.7. Stage 7: Publish the list with its provenance**

The deliverable is the ranked list together with the query definition, the library releases and checksums, the curation sequence, every threshold, the software versions and seeds, and the filter annotations. Deposited in a repository that issues a persistent identifier, this package satisfies the reuse expectations set out in the FAIR principles ([Wilkinson et al., 2016](#)) and makes the search repeatable by a stranger.

### **5. Fitting the Pipeline to Ordinary Practice**

The procedure is intended to run on a workstation rather than a cluster. Framework enumeration is the expensive stage and it is done once per library release, after which the framework table can be reused across projects. The chemistry filters and the shape comparison run only on the shortlist. Most of the stages can be built from RDKit ([Landrum, 2006](#)) with modest scripting, and groups already using a workflow manager will find the provenance requirements largely satisfied by their existing configuration files.

The discipline the pipeline imposes is mostly social rather than technical. It asks a group to fix thresholds before seeing the hit list, to keep every intermediate file, and to publish choices that are currently buried in a scientist's notebook. Groups that adopt it will find the first search slower than usual and later searches faster, because the standardised library and framework tables persist.

### **6. Limitations and What Remains to Be Shown**

This article proposes a procedure and does not evaluate it. Three claims would need testing. The first is retrospective recovery: whether, starting from one member of a pair of chemotypes known from the literature to bind the same target, the pipeline places the other high in the ranking more often than a single-measure search. The second is reproducibility in the strict sense, namely whether two groups given only the provenance package produce the same list. The third is the cost of the discipline, measured as the extra effort a group spends against the number of irreproducible searches avoided.

Two cautions apply regardless. Similarity is a weak guide to activity, and closely related structures with sharply different potency are common enough to have their own literature ([Stumpfe and Bajorath, 2012](#)), so a high-ranking framework is a hypothesis to be tested and never a result. And a pipeline can only be as good as the chemistry hypothesis entered at stage one; recording that hypothesis explicitly makes its weakness visible, which is a benefit rather than a defect.

## 7. Conclusion

Scaffold hopping across public libraries is cheap to attempt and difficult to repeat, and the difficulty lies in undocumented choices rather than in the chemistry. The seven-stage pipeline proposed here fixes the library release, the curation sequence, the similarity measures, the thresholds and the random seeds, keeps the three similarity scores separate, and publishes the provenance alongside the hit list. Whether it recovers known chemotype pairs better than simpler searches, and whether the provenance package is sufficient for independent repetition, are open questions that the groups running such searches are well placed to answer.

## Conflict of interest

To be declared by the author(s) before publication.

## Funding

To be declared by the author(s) before publication.

## References

- Baell, J.B. and Holloway, G.A. (2010). [New substructure filters for removal of pan assay interference compounds \(PAINS\) from screening libraries and for their exclusion in bioassays.](#) *Journal of Medicinal Chemistry*, 53(7): 2719-2740.
- Bemis, G.W. and Murcko, M.A. (1996). [The properties of known drugs. 1. Molecular frameworks.](#) *Journal of Medicinal Chemistry*, 39(15): 2887-2893.
- Bento, A.P., Hersey, A., Felix, E., Landrum, G., Gaulton, A., Atkinson, F., Bellis, L.J., De Veij, M. and Leach, A.R. (2020). [An open source chemical structure curation pipeline using RDKit.](#) *Journal of Cheminformatics*, 12(1): 51.
- Bohm, H.-J., Flohr, A. and Stahl, M. (2004). [Scaffold hopping.](#) *Drug Discovery Today: Technologies*, 1(3): 217-224.
- Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E. and Notredame, C. (2017). [Nextflow enables reproducible computational workflows.](#) *Nature Biotechnology*, 35(4): 316-319.
- Fourches, D., Muratov, E. and Tropsha, A. (2010). [Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research.](#) *Journal of Chemical Information and Modeling*, 50(7): 1189-1204.
- Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. and Overington, J.P. (2012). [ChEMBL: A large-scale bioactivity database for drug discovery.](#) *Nucleic Acids Research*, 40(D1): D1100-D1107.
- Hawkins, P.C.D., Skillman, A.G. and Nicholls, A. (2007). [Comparison of shape-matching and docking as virtual screening tools.](#) *Journal of Medicinal Chemistry*, 50(1): 74-82.
- Heller, S.R., McNaught, A., Pletnev, I., Stein, S. and Tchekhovskoi, D. (2015). [InChI, the IUPAC International Chemical Identifier.](#) *Journal of Cheminformatics*, 7: 23.
- Irwin, J.J., Tang, K.G., Young, J., Dandarchuluun, C., Wong, B.R., Khurelbaatar, M., Moroz, Y.S., Mayfield, J. and Sayle, R.A. (2020). [ZINC20: A free ultralarge-scale chemical database for ligand discovery.](#) *Journal of Chemical Information and Modeling*, 60(12): 6065-6073.
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J. and Bolton, E.E. (2023). [PubChem 2023 update.](#) *Nucleic Acids Research*, 51(D1): D1373-D1380.
- Landrum, G. (2006). [RDKit: Open-source cheminformatics software.](#) RDKit project.
- Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Felix, E., Magarinos, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., Gordillo-Maranon, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C.J., Segura-Cabrera, A., Hersey, A. and Leach,

- A.R. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1): D930-D940.
- Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5): 742-754. <https://doi.org/10.1021/ci100050t>
- Schneider, G., Neidhart, W., Giller, T. and Schmid, G. (1999). Scaffold-hopping by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie International Edition*, 38(19): 2894-2896.
- Stumpfe, D. and Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55(7): 2932-2942.
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1): 31-36.
- Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018.