Pharmacogn. Res., 2027; 19(2):856-869.
https://www.phcogres.com

Original Article

Chemical-Space Characterization and *Target* Assessment of Reported *Citrullus colocynthis* Phytochemicals in Colorectal Cancer

Khalid Alshaghdali*

Department of Medical Laboratory Science, College of Applied Medical Sciences, University of Ha'il, Hail, SAUDI ARABIA.

ABSTRACT

Background: Reported constituents of *Citrullus colocynthis* occupy diverse chemical space, but evidence connecting individual structures to Colorectal Cancer (CRC) targets remains incomplete. **Objectives:** This study characterized 45 curated structures using RDKit descriptors, Morgan fingerprints, scaffolds and clustering, then examined ChEMBL target transfer, CRC associations and target-specific activity models. **Materials and Methods:** The library comprised 19 scaffold categories and ten clusters. Seven targets met the initial CRC intersection; omitting the aggregate PubChem source record reduced this set to five. A broader review supported a detailed assessment of PTPN1 and TDP1. Leakage-controlled, nested scaffold model selection for PTPN1 achieved ROC-AUC 0.910 at pChEMBL ≥ 6.0. Twenty non-exact library structures met the operational applicability criterion. Their maximum ensemble probability was 0.116 at ≥ 6.0, whereas probabilities at ≥ 5.0 ranged from 0.212 to 0.760. In-domain regression estimates spanned pChEMBL 4.60–5.34 against a scaffold-validation RMSE of 0.545. **Results:** Held-out PTPN1 ensemble sensitivity declined from 0.667 to 0.069 across the highest- to lowest-similarity bands, supporting caution with remote chemistry. No TDP1 library structure met the applicability criterion. Saved PTP1B reference poses were recovered at receptor-frame RMSDs of 0.840 and 0.719 Å; coordinate-derived contact maps supplied structural context without establishing affinity. **Conclusion:** Source-ablation and applicability diagnostics strengthen the rationale for retaining PTPN1 as a secondary biochemical hypothesis while withholding prospective direct-hit claims. The resulting compound, assay, and structural records support focused experimental testing rather than a claim of validated CRC inhibitors.

Keywords: Applicability domain; Bioactivity provenance; Chemoinformatics; Citrullus colocynthis; Colorectal cancer; Molecular docking; PTPN1; Target prioritization

Correspondence:
Dr. Khalid Alshaghdali
Department of Medical Laboratory Science, College of Applied Medical Sciences, University of Ha'il, Hail-55476, SAUDI ARABIA.
Email: K.alshaghdaly@gmail.com

Received: 13-10-2026;
Revised: 02-11-2026;
Accepted: 21-11-2026.

INTRODUCTION

Colorectal cancer accounted for approximately 9.9% of new cancer cases and 9.4% of cancer deaths worldwide in the 2024 GLOBOCAN estimates (Sung *et al.*, 2026), while incidence continues to increase among adults younger than 50 years (Siegel *et al.*, 2026). Treatment response depends on disease stage, molecular subtype and acquired resistance. These challenges motivate the search for chemically distinct starting points and computational approaches that can help define testable hypotheses before experimental evaluation (Sadybekov and Katritch, 2023).

DOI: 10.5530/pres.20270026

Copyright Information :
Copyright Author (s) 2027 Distributed under Creative Commons CC-BY 4.0

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

Natural products provide structurally complex frameworks that complement conventional synthetic screening collections (Atanasov *et al.*, 2021). Their computational assessment still requires careful attention to provenance and chemical coverage. Occurrence databases combine reports with varying levels of analytical support; stereochemical assignments may be incomplete, and natural-product structures can differ substantially from the chemistry in predictive-model training sets. Structure curation and applicability assessment are therefore central to interpreting a prediction, rather than preliminary formalities (Sadybekov and Katritch, 2023; Simoben *et al.*, 2023).

Citrullus colocynthis (L.) Schrad., commonly known as colocynth or bitter apple, is a desert-adapted member of the Cucurbitaceae. Reported constituents include cucurbitacins, cucurbitane glycosides, flavonoids, and phenolic compounds (Hussain *et al.*, 2014; Li *et al.*, 2022). Cucurbitacins have been isolated from its leaves (Chawech *et al.*, 2015), and its potential relevance to CRC has been reviewed (Abdulridha *et al.*, 2020). Separate preclinical studies examined cucurbitacins B and I in relation

to Notch signaling in colon cancer and JAK2/STAT3-associated macrophage polarization in CRC (Dandawate *et al.*, 2020; Zhang *et al.*, 2022). A recent fruit-profiling study included EGFR docking, although its cellular experiments used pancreatic and epidermoid models rather than colorectal models (Mandour *et al.*, 2023). These findings provide a biological rationale for examining the reported chemistry, but they do not establish direct protein targets for all database-associated constituents.

Similarity-based target inference, quantitative activity modeling, and docking address different aspects of this problem. Similarity links a query structure to proteins measured for related ligands; a target-specific model estimates an activity label or continuous endpoint within its chemical coverage; docking evaluates compatibility with a prepared protein conformation. Agreement can support experimental prioritization, but shared source data and model assumptions can also produce correlated conclusions (Eberhardt *et al.*, 2021; Simoben *et al.*, 2023; Tropsha, 2010).

A transferred association also inherits the limitations of its underlying measurement. Endpoint labels, assay context, source concentration, and the activity of reference ligands therefore require separate scrutiny. Multiple document identifiers do not necessarily indicate independent replication, and a reference-ligand value cannot be transferred numerically to a related phytochemical. Examining these distinctions helps determine which target hypotheses merit focused assessment.

The study first characterized 45 reported structures and intersected similarity-derived target hypotheses with CRC associations. Broader target review then supported quantitative assessment of PTPN1 and TDP1, with a limited structural examination of PTP1B. A retrospective provenance audit described the endpoint and source composition of the transferred evidence. Additional source-ablation and held-out similarity diagnostics examined how much the conclusions depended on source representation and training-chemistry proximity.

MATERIALS AND METHODS

Phytochemical dataset and structure curation

The starting library was a LOTUS export for *Citrullus colocynthis* (Rutz *et al.*, 2022). After structure-level curation, 45 structures formed the authoritative screening set. All were parsed and sanitized with RDKit 2026.03.5 under Python 3.12.14. Canonical representations and InChIKeys were retained, and unassigned stereochemistry was never filled in by inference. The input structures, the curation record and compound-level annotations are provided as Supplementary Table S1.

LOTUS occurrence and DOI fields were treated as source annotations, not as proof that a structure had been characterized from a particular specimen. PubChem identities were resolved only on exact InChIKey match, and failure to resolve a database identity did not invalidate the supplied structure. One record, LTS0248750, retained a single unassigned stereocenter and was flagged throughout.

Structures were assigned to descriptive groups using ordered substructure rules: a flavone core, pyranose-like rings with at least two hydroxyl substituents, total ring count minus the identified sugar-ring count, carbon count and presence of a ring carbonyl. These rules identify broad chemical features; they do not establish a biosynthetic scaffold such as cucurbitane or confirm metabolite occurrence. The carbonyl-bearing polycyclic group includes named cucurbitacins and related structures. Per-record labels and the original rule outputs are supplied in Supplementary Table S1.

Physicochemical descriptors, drug-likeness annotations and structural alerts

The RDKit descriptor panel comprised *Molecular* Weight (MW), cLogP, Topological Polar Surface Area (TPSA), hydrogen-bond donor and acceptor counts (HBD, HBA), Rotatable Bonds (RotB), fraction Csp3, total and aromatic ring counts, and the Quantitative Estimate of Drug-Likeness (QED). Lipinski-style and Veber-style criteria were recorded as descriptive annotations (Bickerton *et al.*, 2012; Lipinski *et al.*, 2001; Veber *et al.*, 2002). No structure was removed for falling outside conventional oral drug-likeness ranges.

PAINS and Brenk structural alerts were obtained with the RDKit FilterCatalog and retained as flags for assay-specific review rather than as evidence of interference or inactivity (Baell and Holloway, 2010).

Molecular fingerprints, similarity, scaffolds and clustering

Morgan fingerprints used radius 2 and 2,048 bits, giving an ECFP4-like circular representation (Rogers and Hahn, 2010). Pairwise similarity used the Tanimoto coefficient. Bemis-Murcko scaffolds summarise molecular frameworks (Bemis & Murcko, 1996); scaffold frequencies are shown in Figure 1A. The full similarity matrix and scaffold assignments are provided as Supplementary Tables S2 and S3.

Butina clustering was examined at Tanimoto-distance cutoffs of 0.40, 0.50, 0.60 and 0.70 (Butina, 1999), with 0.50 used as the primary descriptive partition. Clusters were not treated as biological classes. Stereochemically distinct records were kept separate, so two records can share an identical fingerprint because the representation used here does not encode chirality.

Principal component analysis of chemical space

The ten descriptors above were standardized to zero mean and unit variance before principal component analysis. PCA was used to describe physicochemical variation and not to infer activity (Figure 1B). Coordinates and loadings accompany Supplementary Table S1 to S3.

Ligand-based target inference, CRC convergence and provenance audit

Target inference used ChEMBL release 37 (Zdrazil *et al.*, 2024). Neighbours were retrieved at an initial similarity cutoff of 0.40 and rescored locally with the same Morgan representation applied to the plant library. Evidence was retained for human single-protein targets with an interpretable activity type and an available pChEMBL value. Transfer cutoffs of 0.40, 0.50 and 0.60 were examined, and 0.50 was used for the primary set. Transferred associations were treated as hypotheses throughout, including where the source neighbour carried an experimental measurement. We retained up to 25 locally rescored neighbors per library structure after retrieval; the captured associations are therefore not exhaustive.

CRC associations came from Open Targets release 26.06 using colorectal cancer, colon carcinoma and rectal carcinoma queries (Ochoa *et al.*, 2021). Up to 500 associations per disease entity were retained at association score 0.20 or above, then intersected with the primary transferred set to define the initial CRC-convergent targets (Figure 2A).

The provenance audit used all 473 retained transfer rows, including neighbors below the 0.50 cutoff used for the initial CRC intersection. It counted reference ligands, assays, document identifiers and reference-ligand/assay/endpoint/value tuples to distinguish transfer-row multiplicity from distinct source evidence. Within the 14 reviewed targets in Table 1, descriptive stratum A required at least 50% IC_{50}/Ki/Kd/EC_{50}-labeled rows and at least two document identifiers; B required some such rows without meeting A; C comprised Potency-labeled rows only. These strata summarize endpoint labels and source composition, not validated evidence quality or independent replication. Potency can itself represent quantitative concentration-response activity (Zdrazil *et al.*, 2024). Assay descriptions and source metadata were reviewed separately; current metadata checks did not replace the frozen release-37 activities.

Protein interaction and enrichment analyses used the STRING database (Szklarczyk *et al.*, 2025), run against STRING v12.0 with the service background and reported false discovery rates retained without further local adjustment. The seven input genes were selected partly based on disease annotations, so the enrichment analysis provided biological context (Figure 2B). Structural review used PDBe UniProt-to-author-chain mappings and RCSB entry and ligand metadata. The source records and the retrospective audit are supplied in Supplementary Table S4.

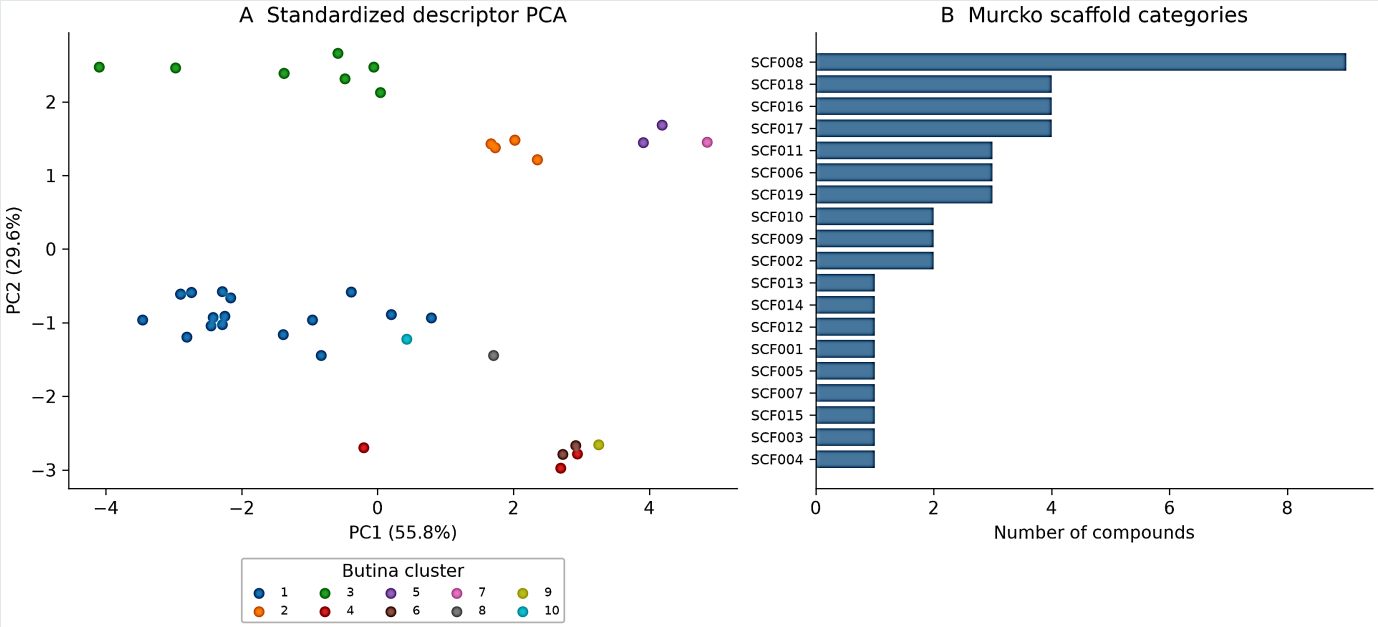

Figure 1: Chemical-space characterization of the 45 curated structures. (A) Principal component analysis of standardized descriptors, colored by Butina cluster; PC1 and PC2 explain 55.8% and 29.6% of variance. (B) Frequency of Bemis–Murcko scaffold categories, including the most frequent category and multiple singleton frameworks.

PTPN1 target-specific machine learning

PTPN1/PTP1B (UniProt P18031; ChEMBL CHEMBL335) was selected for detailed assessment because it combined transferred chemical support, human ligand-bound structures and experimental CRC relevance (Hoekstra *et al.*, 2016). The retrospective provenance audit placed its transfer records in stratum A; this annotation was not used as an independently validated selection criterion. The activity models assessed compatibility with known PTP1B ligand chemistry rather than establishing a mechanism of action for the plant.

Primary ChEMBL records required human assays, confidence score 9, exact relations, an available pChEMBL value, no

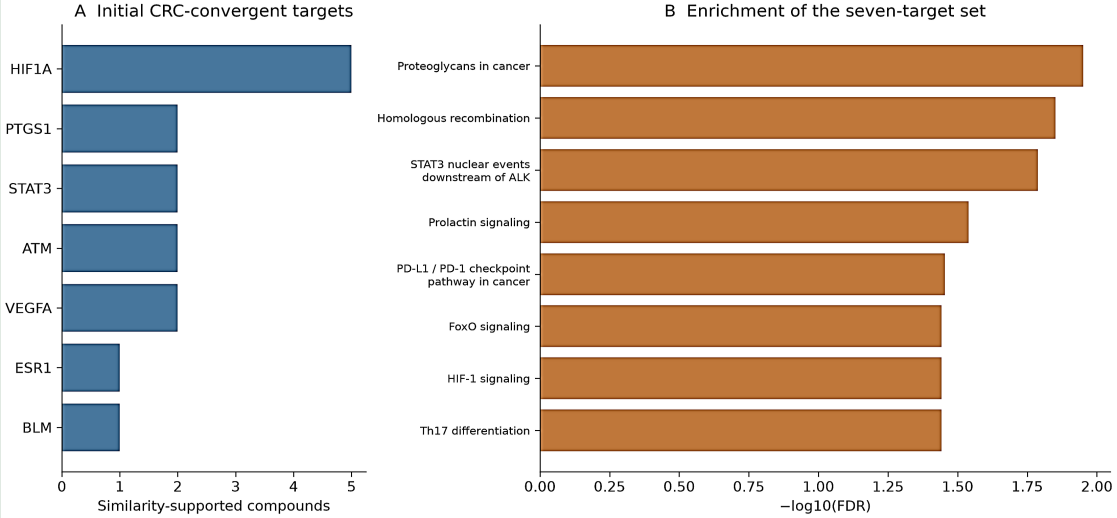

Figure 2: *Target* evidence and pathway context for the initial colorectal cancer intersection. (A) Number of similarity-supported compounds for each of the seven convergent targets. The axis shows counts, not a composite score or a binding probability. (B) Selected enrichment terms for the same set, plotted as the negative logarithm of the false discovery rate. Because the input genes were selected partly on disease annotation, these terms give biological context rather than independent support.

Table 1: Endpoint and source composition of transferred target evidence. The audit uses the full retained transfer population rather than only the initial CRC intersection. Endpoint counts are transferred rows; one source measurement may contribute to multiple library structures. Ligands, assays and sources denote distinct reference-ligand, assay and document identifiers. A/B/C are descriptive strata defined in the ligand-based target inference methods, not validated quality grades.

Target	Initial CRC set	Library (scaffolds)	Ligands / assays / sources	Endpoint labels (transfer rows)	Ref. pChEMBL max (median)	Stratum
PTPN1	No	17 (6)	7 / 6 / 6	IC50 40; Ki 3	5.60 (5.37)	A
TDP1	No	11 (7)	6 / 1 / 1	Potency 22	8.74 (4.58)	C
STAT3	Yes	2 (2)	1 / 1 / 1	IC50 2	5.52 (5.52)	B
ATM	Yes	2 (2)	1 / 1 / 1	Potency 2	4.70 (4.70)	C
HIF1A	Yes	6 (4)	2 / 3 / 2	IC50 2; Potency 10	6.57 (6.10)	B
VEGFA	Yes	2 (1)	1 / 1 / 1	Kd 2	8.77 (8.77)	B
BLM	Yes	1 (1)	1 / 1 / 1	Potency 1	4.60 (4.60)	C
PTGS1	Yes	2 (1)	1 / 1 / 1	IC50 2	4.14 (4.14)	B
ESR1	Yes	1 (1)	4 / 5 / 4	EC50 5; IC50 1	7.12 (5.95)	A
AKR1B1	No	7 (4)	1 / 1 / 1	IC50 7	5.72 (5.72)	B
ALDH1A1	No	11 (8)	5 / 1 / 1	Potency 15	4.55 (4.55)	C
APEX1	No	9 (6)	5 / 1 / 1	Potency 21	5.90 (4.90)	C
IGF2BP1	No	2 (2)	1 / 2 / 2	Kd 4	7.18 (6.55)	A
NTSR1	No	8 (5)	1 / 1 / 1	EC50 8	5.57 (5.57)	B

validity comment, and IC_{50} or Ki endpoints. Structures were cleaned, salt-reduced with FragmentParent, uncharged, sanitized, canonicalized and deduplicated on standardized InChIKey. Measurements were aggregated as compound and endpoint medians, then as the median of the IC_{50} and Ki endpoint medians, which limits duplicate weighting without making the endpoints equivalent. Classification labels used aggregate pChEMBL thresholds of 5.0, 5.5 and 6.0, with lower values treated as inactive, and binary metrics used a fixed probability cutoff of 0.50.

Features combined Morgan radius-2, 2,048-bit fingerprints with the ten RDKit descriptors. Candidate classifiers were logistic regression, random forest, XGBoost and a radial-basis support vector machine (Breiman, 2001; Chen and Guestrin, 2016; Cortes and Vapnik, 1995; Pedregosa *et al*.,2011), with parameter presets fixed in advance and no outer-fold metric used for tuning. Evaluation used three outer and three inner scaffold folds. Outer folds were stratified by potency bin and held complete non-isomeric Murcko groups together, with acyclic structures sharing a single group. Inner folds selected the model family on an equal-weight mean of ROC-AUC, average precision, MCC and balanced accuracy, minus 0.1 times the Brier score. The selected procedure was then evaluated on the corresponding outer fold. Individual models and a predefined equal-weight four-model ensemble were evaluated alongside it. Headline cross-validation metrics report the nested selected procedure; prospective library probabilities report the four-model ensemble, and both are tabulated per compound.

Descriptor scaling was fitted separately within each training fold. Probability calibration used a sigmoid logistic fit to training-only, scaffold-grouped out-of-fold decision scores. The SVM was fitted with probability=False to avoid hidden calibration folds, and no random early-stopping subset was used. Final prospective models were calibrated through grouped cross-validation across all eligible training chemistry. This is internal nested validation, not validation on an independent experimental dataset.

Before any split, training structures with Morgan Tanimoto similarity of 0.90 or above to any library compound were excluded. Exact quantitative matches were retained as experimental references and received no prospective prediction. *Applicability* categories used maximum training similarity: in-domain at 0.50 or above, borderline from 0.35 to below 0.50, and out-of-domain below 0.35. Only in-domain, non-exact compounds received reportable prospective estimates, while raw outputs for more remote chemistry were retained for audit. Metrics included ROC-AUC, average precision, MCC, balanced accuracy, sensitivity, specificity, F1 and Brier score, with uncertainty from 500 scaffold-cluster bootstrap resamples of the outer predictions conditional on the fitted fold models. Continuous models were random forest regression, XGBoost regression and support vector regression, selected on inner-fold RMSE. We evaluated IC_{50}-only datasets separately at all three thresholds and by regression.

PTP1B reference-pose recovery and structural assessment

The structural stage focused on six ligand preparations covering five library records, including α-amyrin and both configurations of the unassigned LTS0248750 stereocenter. It was not a 45-compound docking screen. Saved AutoDock Vina 1.2.7 calculations were inspected (Eberhardt *et al.*, 2021). Human chain A from PDB 2F71 served as the primary receptor and 2QBS as a second reference-pose control, both mapping to UniProt P18031. Receptors were rigid, deposited ligands and waters were removed, and alternate conformations in 2F71 were resolved to conformer A. The primary box was centred at (39.945, 9.945, 14.680) Å with dimensions 28 × 28 × 28 Å.

Ligands were prepared using RDKit ETKDGv3 to generate 10 conformers, followed by MMFF minimization, lowest-energy-conformer selection, hydrogen addition, and Meeko PDBQT preparation. Library searches used exhaustiveness 16, 20 requested poses, an 8 kcal/mol energy range and seed 20260825; redocking used exhaustiveness 32 and 20 requested poses. We calculated reference-pose recovery as symmetry-aware heavy-atom RMSD in the fixed receptor frame using RDKit CalcRMS without ligand fitting. We redocked both deposited ligands as geometric controls. Their Vina scores are protocol-dependent outputs, not experimental potency anchors. The 2QBS/024 calculation used a separate receptor and was not treated as part of a common potency scale.

Residue contacts were quantified as minimum ligand–residue heavy-atom distances ≤ 4.0 Å. Figure S4 was reconstructed directly from the saved coordinates, and its two-dimensional maps link each residue to its closest ligand heavy atom; gray Cα traces provide local protein context. These distances do not assign hydrogen bonds, interaction energies or favorable/unfavorable effects. Figure 3 presents the same poses as rendered by Discovery Studio Visualizer, whose interaction categories are assigned by that program from a docked pose and were not used as quantitative evidence. Best-scoring and contact-selected poses are distinguished: the latter were chosen from the ten best-scoring poses using twice the core-contact count plus the active-site-contact count minus 0.25 times the score difference from the best pose. Core residues were Asp181 and residues 215–221. Source-coordinate hashes, atom identifiers and full contact listings are supplied. Figure S3 documents the second receptor control.

TDP1 target-specific modeling and applicability assessment

Coordinate-derived contact maps, source coordinates and atom-level provenance are supplied in Figure S4 and Supplementary Data S8. TDP1 (UniProt Q9NUW8; ChEMBL CHEMBL1075138) was evaluated because experimental CRC work links its depletion to irinotecan sensitization (Meisenberg *et al.*, 2015). Quantitative records required explicit human assays, IC_{50}, Ki, or Kd endpoints, exact relations, pChEMBL values, and valid standardized structures. Compound summaries were harmonized with the PTPN1 procedure. Large qHTS and cell-screen datasets were kept separate from the quantitative model. TDP1 used the same feature definition, candidate algorithms, nested scaffold validation, grouped calibration, leakage exclusion and applicability categories as PTPN1. No TDP1 training structure met the 0.90 exclusion threshold. Records, assignments and performance are in Supplementary Data S9, with the 45-compound applicability audit in S10.

Source-sensitivity and held-out applicability diagnostics

Post hoc source-sensitivity analyses held the retained similarity-transfer table and CRC gene universe fixed. Each of the 53 document identifiers was omitted in turn, and target support was recomputed as transferred rows, distinct library structures, scaffolds, reference ligands and assays. The primary CRC comparison retained the original Tanimoto cutoff of 0.50. A separate comparison used all retained similarities for the 14 reviewed targets. Omission of CHEMBL1201862 was highlighted because official metadata identify it as the aggregate PubChem BioAssay record. Removing a source identifier is a dependence diagnostic, not removal of a proven low-quality study or a test of experimental independence.

Applicability diagnostics used each outer-held-out compound's maximum similarity to the corresponding outer-training chemistry. The existing ≥ 0.50, 0.35–<0.50 and < 0.35 bands were applied without adjusting their boundaries to improve results. Counts, class prevalence, classification metrics and Brier scores were recomputed separately for the nested selected procedure and the predefined ensemble at all three activity thresholds, for mixed and IC_{50}-only datasets. Brier skill was calculated as one minus the ratio of model Brier score to the outer-training-prevalence baseline Brier score, within the same held-out stratum. At pChEMBL ≥ 6.0, 500 within-stratum scaffold-cluster resamples supplied conditional intervals. Models were not refitted, no new independent validation set was introduced, and no formal between-band significance tests were performed. These *post hoc* checks were not used to select models or revise applicability boundaries.

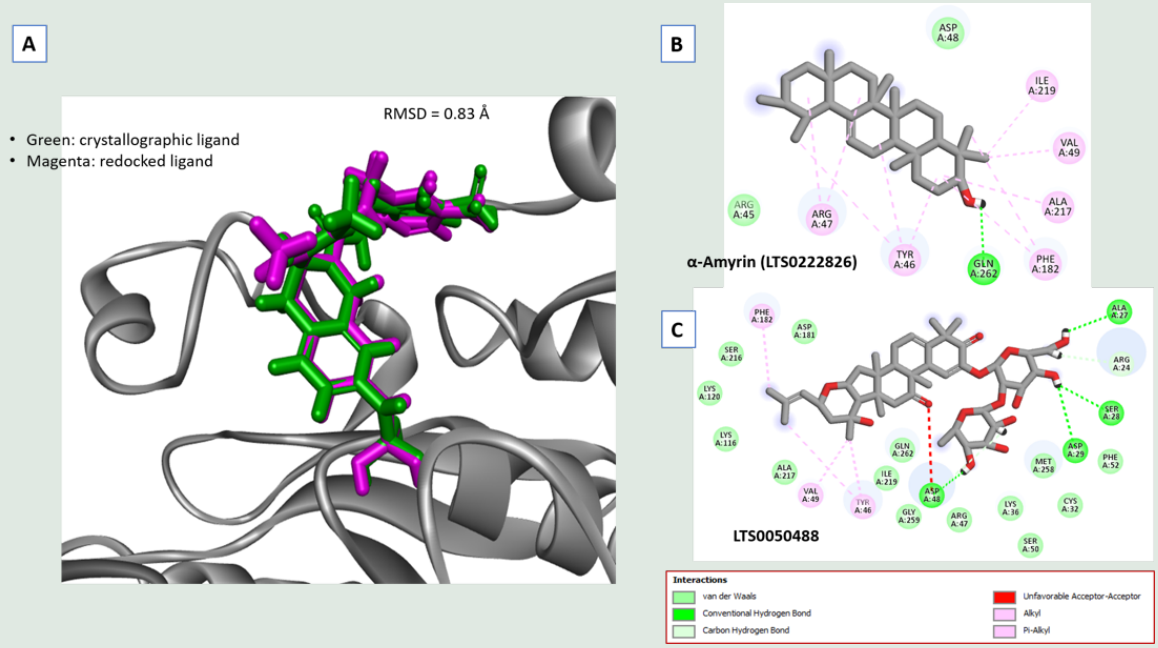

Figure 3: PTP1B pose recovery and ligand interaction diagrams. (A) Crystallographic ligand (green) and redocked pose 1 (magenta) in the 2F71 binding site. The in-panel label reports the historical ligand-superposed RMSD of 0.83 Å; the independently recomputed receptor-frame value used throughout the text is 0.840 Å. (B) α-Amyrin pose 1, Vina score −7.699 kcal/mol. (C) LTS0050488 contact-selected pose 4, score −7.357 kcal/mol; its distinct raw-score-best pose 1 scores −7.573.

RESULTS

Composition and physicochemical range of the curated library

All 45 curated structures were successfully parsed and sanitized. The rule-based grouping identified 20 carbonyl-bearing polycyclic terpenoid-like records, comprising 16 glycosides and four aglycones, together with eight flavonoid glycosides, eight other aromatic glycosides, six other polycyclic terpenoid-like records and three simple phenolics (Table 2). These groups describe the supplied structures only.

Descriptors spanned MW 138.17 to 806.94 g/mol, cLogP -2.44 to 8.02, TPSA 20.23 to 280.43 Å^2, HBD 1 to 11, HBA 1 to 16, RotB 0 to 9 and QED 0.102 to 0.704. Eleven compounds had no Lipinski-style violations, while 12, 6, and 16 had one, two, and three. Twenty had no Veber-style violations, and 25 had one. One compound matched a PAINS alert and 29 matched at least one Brenk alert, most commonly an isolated alkene. Exact PubChem InChIKey matching resolved 27 identities, leaving 18 unresolved.

Chemical-space diversity and scaffold organization

Mean and median pairwise Morgan-Tanimoto similarities were 0.267 and 0.207, across a range of 0.047 to 1.000. Nineteen Bemis-Murcko scaffold categories were present, nine of them singletons (Figure 1B), and Butina clustering at distance 0.50 produced ten clusters. A pairwise similarity of 1.000 arises between stereochemically distinct records because the fingerprint used for this comparison does not encode chirality.

PC1 and PC2 explained 55.8% and 29.6% of standardized descriptor variance, respectively, or 85.4% together (Figure 1A). The projection separated regions containing polar oxygen-rich structures, lipophilic polycyclic constituents and smaller aromatic compounds. LTS0222826, LTS0030927 and LTS0255445 were flagged as descriptor-distance outliers; this designation describes their position in chemical space rather than a demonstrated processing error.

Target inference, CRC convergence and provenance

All 45 structures contributed similarity-transfer evidence, producing 473 compound–target rows across 69 human single-protein target annotations and 53 source-document identifiers.

Table 2: Composition, physicochemical range and chemical-space metrics of the curated *Citrullus colocynthis* library.

Category	Metric	Result
Library	Curated / valid structures	45 / 45
Structural class	Carbonyl-bearing polycyclic terpenoid-like records (glycoside / aglycone)	20 (16 / 4)
Structural class	Flavonoid glycosides	8
Structural class	Other aromatic glycosides	8
Structural class	Other polycyclic terpenoid-like records (aglycone / glycoside)	6 (5 / 1)
Structural class	Simple phenolics	3
Identity	Exact PubChem InChIKey resolved / unresolved	27 / 18
Physicochemical space	MW	138.17-806.94 g/mol
Physicochemical space	cLogP	-2.44 to 8.02
Physicochemical space	TPSA	20.23-280.43 Å^2
Physicochemical space	HBD / HBA	1-11 / 1-16
Physicochemical space	Rotatable bonds	0-9
Physicochemical space	QED	0.102-0.704
Rule-based annotations	Lipinski violations (0/1/2/3)	11 / 12 / 6 / 16
Rule-based annotations	Veber violations (0/1)	20 / 25
Structural alerts	PAINS / Brenk	1 / 29
Chemical diversity	Mean / median pairwise Tanimoto	0.267 / 0.207
Chemical diversity	Murcko scaffold categories (singletons)	19 (9)
Chemical diversity	Butina clusters (distance 0.50)	10
Chemical space	PC1 / PC2 / cumulative variance	55.8% / 29.6% / 85.4%

Each row is a transferred association, not necessarily a unique experimental observation.

At the primary similarity cutoff of 0.50, intersection with CRC associations retained ATM, HIF1A, VEGFA, BLM, PTGS1, STAT3 and ESR1 (Figure 2A). Broader review of targets omitted by the disease-association threshold added candidates with substantial transferred support and usable structures, including PTPN1 and TDP1.

Across the full transfer table, 266 of 473 rows (56.2%) carried the endpoint label Potency, and 277 rows (58.6%) shared document identifier CHEMBL1201862. Official metadata identify this record as the PubChem BioAssay dataset; it aggregates multiple assays and is not one publication or screening campaign. PTPN1 had 43 transferred rows across 17 library structures, but these involved seven reference ligands, six assays and nine distinct reference-ligand/assay/endpoint/value tuples from six source records. Its transferred reference values ranged from pChEMBL 4.37 to 5.60. Those values characterize the selected neighbors rather than an upper bound on the activity of the phytochemicals.

The descriptive strata placed PTPN1, ESR1 and IGF2BP1 in A. For TDP1, the 22 transferred rows involved six reference ligands and one assay within the PubChem source collection, with Potency values ranging from pChEMBL 4.10 to 8.74 (Table 1). This source concentration warrants checking assay context, but neither the Potency label nor stratum C establishes poor assay quality, interference or inactivity. PTPN1 was assessed because its chemical support, CRC rationale and available structures made a focused analysis feasible. TDP1 provided a second biologically motivated hypothesis with a separately curated quantitative modeling dataset.

Cucurbitacin I (LTS0230723) illustrates the limits of the captured annotations. Its eight retained transfer rows mapped to ITGAL, SMN1/SMN2, HIF1A and CYP3A4, including fingerprint-identical neighbors, whereas no STAT3 row appeared in this retained subset. STAT3 entered the CRC intersection through other library–neighbor associations. This observation concerns the retrieval and filtering used here; it does not establish absence of an STAT3 record from all of ChEMBL or negate experimental signaling findings (Dandawate *et al.*, 2020; Zhang *et al.*, 2022).

Pathway context for the CRC-convergent set

Enrichment of the seven-target set identified proteoglycans in cancer (three of seven input genes; FDR 0.0112), homologous recombination (two of seven; 0.0141), STAT3 nuclear events downstream of ALK signaling (two of seven; 0.0163), prolactin signaling (two of seven; 0.0289) and the PD-L1 and PD-1 checkpoint pathway in cancer (two of seven; 0.0351). FoxO, Th17 and HIF-1 signaling each contained two input genes at FDR 0.0361 (Figure 2B). These terms align with CRC biology and help organize hypotheses, but the input was small, several terms share genes, and the gene set was selected partly based on disease annotation, so the analysis provides context rather than independent support.

The PTPN1 source contained 3,229 exact-relation IC_{50} and Ki records, of which 3,207 resolved to 2,662 standardized structures after 22 records without source SMILES were excluded. Removing the exact α-amyrin match left 2,661 modeling structures, 572 active and 2,089 inactive at pChEMBL 6.0. The IC_{50}-only sensitivity set contained 2,414 structures under the same leakage control.

Nested model selection at pChEMBL 6.0 gave ROC-AUC 0.910 (95% scaffold-bootstrap interval 0.882 to 0.933), average precision 0.760, MCC 0.615, balanced accuracy 0.795, sensitivity 0.656, specificity 0.934 and Brier score 0.090 (Figure 4A). The predefined ensemble gave an ROC-AUC of 0.914, and IC_{50}-only nested selection gave 0.897, with an MCC of 0.600. Nested regression returned R^2 0.621, RMSE 0.545 pChEMBL units, MAE 0.410 and Spearman correlation 0.719. Performance was threshold-dependent: ROC-AUC fell to 0.880 at pChEMBL 5.5 and to 0.820 (0.791 to 0.849) at 5.0, where specificity was 0.545 against an active prevalence of 65.2%. Model-specific and fold-level results are in Supplementary Data S7.

Of the 44 non-exact library structures, 20 met the operational in-domain criterion, 22 were borderline and two were out-of-domain. Among the 20 in-domain structures, maximum and median ensemble probabilities were 0.760/0.653 at pChEMBL ≥ 5.0, 0.403/0.207 at ≥ 5.5 and 0.116/0.034 at ≥ 6.0. At ≥ 5.0 the full probability range was 0.212–0.760, and 16 structures exceeded 0.50; none did so at ≥ 5.5 or ≥ 6.0. The maximum IC_{50}-only value at ≥ 6.0 was 0.103. In-domain regression predictions ranged from pChEMBL 4.60 to 5.34, with median 5.20.

Interpretation depends on the activity definition. At ≥ 5.0, the selected procedure retained ROC-AUC 0.820, but its sensitivity/specificity trade-off at the fixed 0.50 cutoff was 0.876/0.545. The corresponding ensemble values were 0.885/0.524. The median library probability of 0.653 was close to the source active fraction of 0.652, although the individual scores were not uniform, which falls short of demonstrating weak inhibition. At ≥ 6.0, prospective scores were low, but a subthreshold prediction is not an experimental negative. Regression errors of approximately 0.55 pChEMBL units and unverified calibration in plant-library chemistry also preclude assigning a precise potency range.

α-Amyrin (LTS0222826) matched a ChEMBL record with IC_{50}-derived pChEMBL 4.71 and was retained as a weak experimental comparator without a prospective prediction. Receptor-frame reference-pose RMSDs were 0.840 Å for 2F71/UN7 and 0.719 Å for 2QBS/024. The corresponding redocking scores were −12.396 and −8.619 kcal/mol. Across the six library preparations,

including the known comparator, best scores were −7.699 for α-amyrin, −7.573 for LTS0050488, −7.324/−6.308 for the two LTS0248750 configurations, −6.141 for LTS0102199 and −5.984 for LTS0036768. Figure 4B shows the 2F71 calculations together; the 2QBS score is a separate receptor result. No score difference was converted into a potency estimate, and the approximately 0.13 kcal/mol α-amyrin/LTS0050488 separation was not treated as reliable affinity discrimination.

Figure 3 summarizes this limited structural assessment, and Figure S4 maps the same poses from the deposited coordinates. α-Amyrin pose 1 had nine residue contacts within 4 Å, including Tyr46, Asp48, Phe182, Ala217, Ile219 and Gln262. LTS0050488 contact-selected pose 4 scored −7.357 kcal/mol and had 14 contacts, including Asp181 and the Ala27/Ser28/Asp29 region; its nearest Asp48 separation was 2.90 Å. Its raw-score-best pose 1 scored −7.573 kcal/mol, lacked Asp181 and Cys215 contacts, and included Gln21, His25, Arg254 and Gly259 (Figure S1). Neither of these two LTS0050488 poses contacted Cys215 under the same cutoff. LTS0248750 configuration 2 pose 6 scored −7.180 kcal/mol and contacted Cys215 and Arg221 among 16 residues, but the source stereocenter remains unresolved. These are geometric observations, not evidence of catalytic inhibition. The distances and pose identities quoted here are coordinate-derived and are plotted in Figure S4; the software-assigned interaction categories shown in Figure 3 were not used as independent quantitative evidence.

TDP1 applicability limits prospective interpretation

TDP1 retrieval returned 695 IC_{50}, Ki, Kd, and EC_{50} records; 407 qualified for the mixed quantitative set, representing 323 structures. Active and inactive counts were 182 and 141, 123 and 200, and 71 and 252 at thresholds 5.0, 5.5, and 6.0. The IC_{50}-only set contained the same 323 structures, with counts of 181 and 142, 124 and 199, and 76 and 247; the small differences reflect endpoint aggregation rather than changes in the compound universe.

Nested model selection gave ROC-AUC 0.884, 0.880 and 0.814 at the three thresholds. At 6.0 the 95% scaffold-bootstrap interval was 0.725 to 0.874, with average precision 0.574, MCC 0.496, balanced accuracy 0.744 and Brier score 0.127. IC_{50}-only nested ROC-AUC was 0.815, and nested regression gave R^2 0.545, RMSE

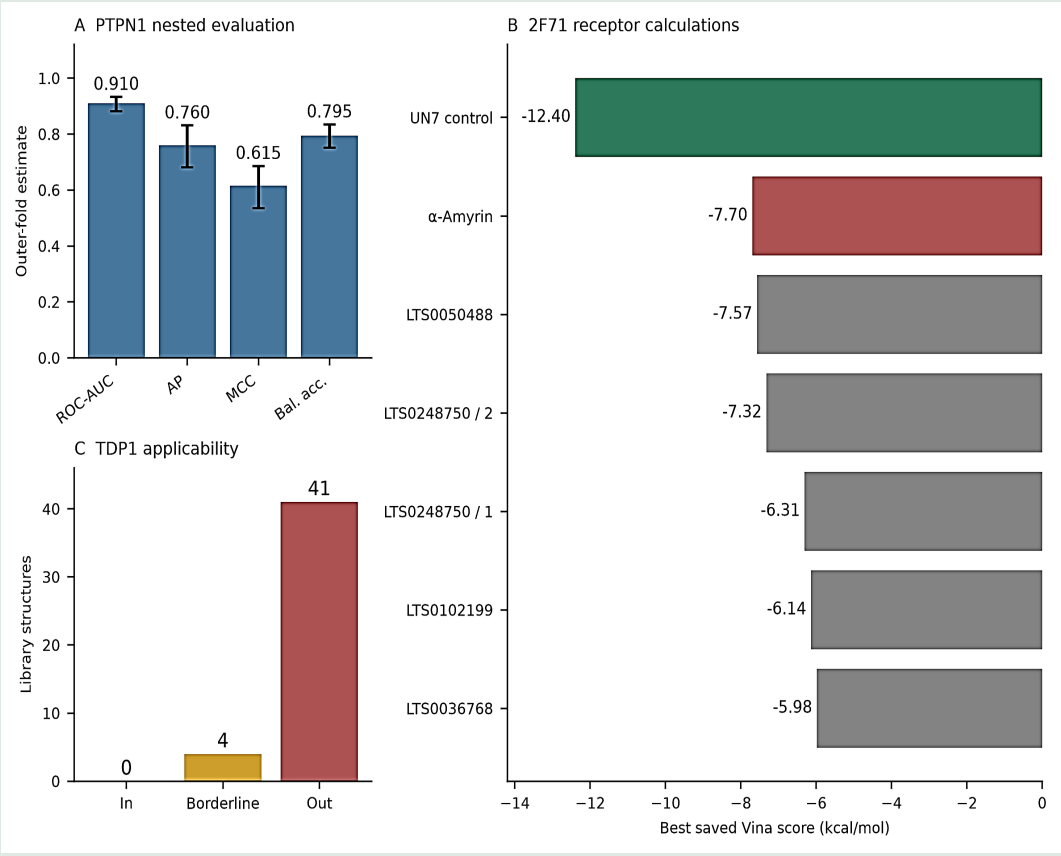

Figure 4: *Target* assessment and saved docking scores. (A) PTPN1 nested model-selection performance at pChEMBL ≥ 6.0. Error bars are 95% intervals from 500 scaffold-cluster resamples of fixed outer predictions, without model refitting; AP denotes average precision. (B) Best Vina scores from the 2F71 calculations for UN7 and six library preparations, including the experimentally weak α-amyrin comparator. The 2QBS/024 control scored −8.619 kcal/mol in a separate receptor run (Figure S3). These scores are not an experimental potency scale. (C) TDP1 applicability categories. Outer-fold model diagnostics are supplied in Figure S2.

0.564 pChEMBL units, MAE 0.425 and Spearman correlation 0.733.

No library structure met the in-domain criterion for TDP1: four were borderline and 41 were out-of-domain, and no exact quantitative match existed. Prospective ranking was therefore not accepted, and TDP1 was not advanced to docking (Table 3 and Figure 4C). The transfer audit and the quantitative-model domain describe different data populations, so the applicability result does not follow from the endpoint/source stratum, nor does it show that these compounds are inactive against TDP1.

Source dependence and empirical applicability diagnostics

Omitting the aggregate PubChem source removed ATM and BLM from the initial CRC intersection, leaving HIF1A, VEGFA, PTGS1, STAT3 and ESR1. HIF1A support decreased from five library structures across three scaffolds to two structures across two scaffolds; the supported compound counts for the other four retained targets were unchanged (Figure 5A and Table S4). Across all 53 individual omissions, HIF1A and ESR1 remained in the initial set throughout, whereas PTGS1, STAT3 and VEGFA each disappeared when their own single supporting source was removed. Retention after PubChem omission alone therefore did not establish source independence.

Within the broader 14-target review, PTPN1's 43 rows and 17 supported library structures were unchanged by PubChem omission. No single source-ID omission eliminated PTPN1, although removal of CHEMBL1921699 reduced support to six rows, two library structures and two scaffolds. All 22 transferred TDP1 rows originated from the PubChem collection and were lost in that specific sensitivity analysis. These transfer results leave the separately curated quantitative modeling datasets unchanged.

The PTPN1 ensemble showed declining performance across the prespecified held-out similarity bands at pChEMBL ≥ 6.0 (Figure 5B). ROC-AUC was 0.928, 0.834 and 0.786, and balanced accuracy was 0.800, 0.646 and 0.529 in the in-domain, borderline and out-of-domain strata, respectively. Sensitivity declined from 0.667 to 0.351 and 0.069. The strata contained 2,023, 326 and 312 compounds, including 486, 57 and 29 actives. Brier scores of 0.087, 0.112 and 0.079 must be read with these class distributions; the low remote-stratum Brier score did not imply good recovery of actives. Brier skill relative to the corresponding training-prevalence baseline was 0.525, 0.234 and 0.201.

TDP1 ensemble ROC-AUC was 0.837 in-domain and 0.309 out-of-domain, with balanced accuracy 0.713 and 0.477 (Figure 5C). Only five actives occurred among the 27 out-of-domain compounds. The borderline stratum contained one active among 23 compounds, making its ROC-AUC of 0.955 a particularly unstable summary; ensemble sensitivity there was zero. These small strata cannot establish a general performance gradient or validate a universal similarity cutoff. They do support the cautious decision to withhold TDP1 library rankings. Full selected-procedure, ensemble, threshold and endpoint diagnostics are supplied in S1 and S2.

DISCUSSION

The principal finding is a constrained target assessment rather than a direct-inhibitor nomination. The curated library is structurally varied, yet this diversity does not translate into uniform coverage by target-specific activity models. PTPN1 remained suitable for secondary biochemical investigation, whereas the TDP1 models could not support prospective ranking of this chemistry. The source-sensitivity analysis and held-out domain diagnostics make the reasons for these different decisions more explicit.

Chemical diversity is useful for selecting informative experiments, but the relevant diversity depends on the question. The scaffold and cluster summaries describe variation among the 45 supplied structures; they say nothing about how fully the set covers the metabolome, compound abundances or the constituents that reach CRC-relevant tissue. Likewise, physicochemical properties and structural alerts should inform solubility, assay design and interference controls without being interpreted as

Table 3: Target-assessment summary for PTPN1 and TDP1.

Target	Dataset and transfer context	Nested scaffold validation	Prospective result	Decision
PTPN1	Transfer stratum A; 2,661 leakage-controlled IC_{50}/Ki structures	At pChEMBL ≥ 6.0: ROC-AUC 0.910 (95% interval 0.882–0.933); AP 0.760; MCC 0.615; balanced accuracy 0.795	20/44 non-exact structures in-domain; max ensemble probability 0.116 at ≥ 6.0; regression 4.60–5.34; receptor-frame RMSDs 0.840/0.719 Å	Secondary biochemical hypothesis; no direct hit established
TDP1	Transfer stratum C; separate 323-structure quantitative model	At ≥ 6.0: ROC-AUC 0.814 (95% interval 0.725–0.874); continuous regression R^2 0.545	0/45 in-domain; 4 borderline and 41 out-of-domain; no exact quantitative match	Not advanced; no accepted prospective ranking

measured exposure, toxicity or developability (Atanasov *et al.*, 2021; Baell and Holloway, 2010; Simoben *et al.*, 2023). Polar glycosides and lipophilic polycyclic constituents may require different experimental handling, so a single unqualified ranking would obscure practical differences between them.

The provenance analysis shows why transfer-row counts require a source-level interpretation. Multiple phytochemicals can inherit the same reference observation, and an aggregate dataset can contain several assays under one document identifier. The PubChem omission analysis therefore asks whether the retained target association survives removal of that collection; it does not test whether the collection is valid. PTPN1 survived every individual source-ID omission, but its support decreased from 17 library structures to two when its most influential source was removed. Retention is consequently not equivalent to broad or independently replicated evidence. The 53 omissions also include many sources unrelated to a given target, so a high retention fraction must not be interpreted as a probability that the association is correct.

The cucurbitacin I example further illustrates the difference between a restricted annotation table and a mechanistic account. Preclinical signaling effects and database measurements of direct or assay-associated targets are different evidence types. Absence of STAT3 from the retained cucurbitacin I transfer subset neither contradicts the cited experiments nor establishes that STAT3 is absent from the complete database (Dandawate *et al.*, 2020; Zhang *et al.*, 2022). A full mechanism review would require a separate search of primary experiments, including assay conditions, concentrations and the distinction between direct engagement and downstream signaling.

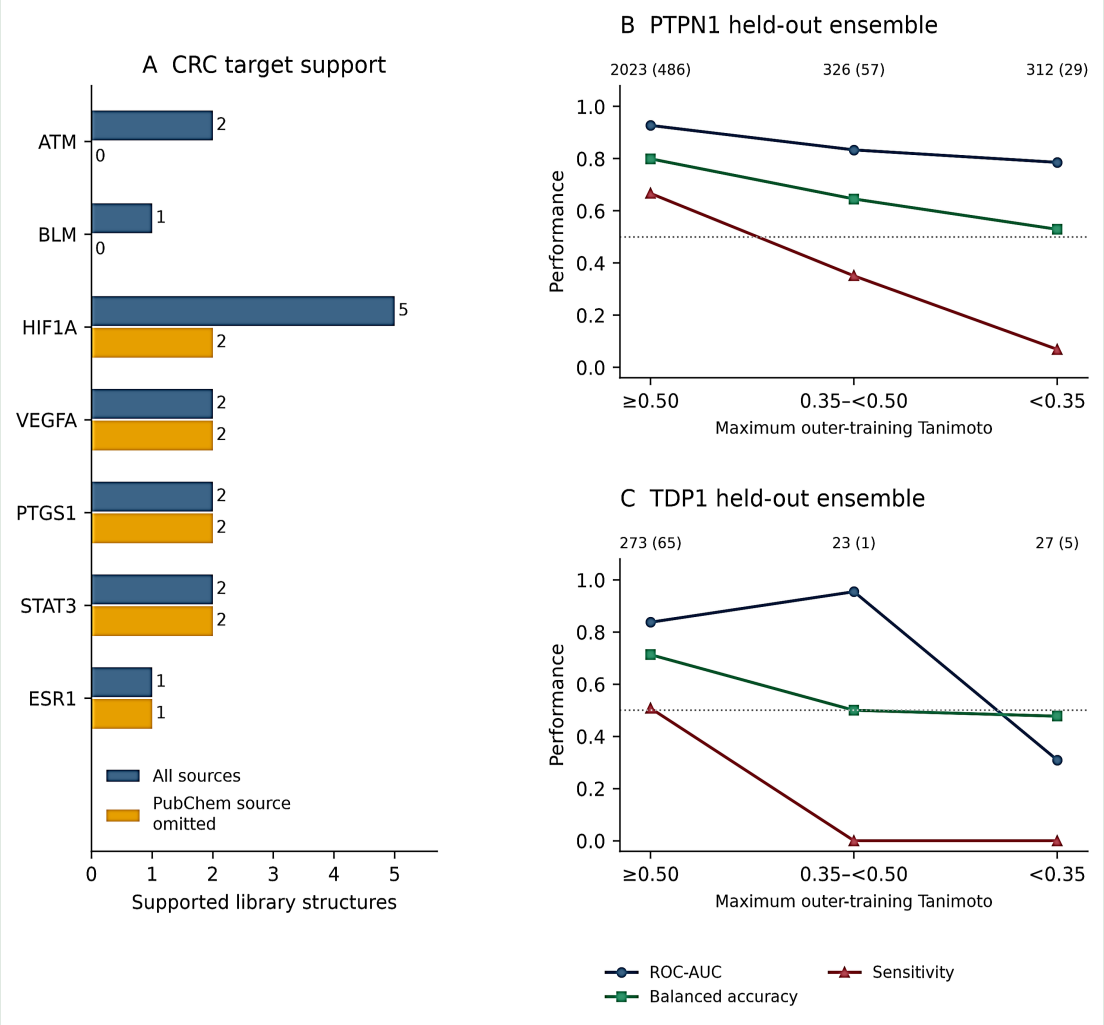

Figure 5: *Post* hoc robustness diagnostics. (A) Supported library structures in the initial CRC intersection before and after omission of the aggregate PubChem source identifier, with the original similarity cutoff and CRC universe fixed. This is a source-dependence analysis, not an assay-quality judgment. (B, C) Frozen outer-held-out ensemble performance at pChEMBL ≥ 6.0, stratified by maximum similarity to the corresponding outer-training chemistry. Numbers above each band give compounds (actives).

For PTPN1, the threshold analysis addresses an important alternative explanation for low prospective scores. Lowering the active definition to pChEMBL ≥ 5.0 increased probabilities for several in-domain structures, but the model also operated in a dataset where approximately two-thirds of compounds met that definition. At the fixed probability cutoff, specificity was modest. The more permissive scores therefore support, at most, prioritization for a low-potency biochemical question; they do not establish weak inhibition across the library. Separate classifiers at the three thresholds were not constrained to form a monotonic potency-response model, and their probabilities should not be treated as three measurements of a compound's potency.

The held-out domain analysis provides an empirical check on this interpretation. For PTPN1, the progressive loss of sensitivity away from training chemistry shows that acceptable aggregate ROC-AUC can coexist with poor identification of remote actives at a fixed decision threshold. Class prevalence also changes across the bands, so Brier score or precision alone cannot be used to rank domain reliability. The similarity criterion remains heuristic: performance within a ChEMBL stratum does not establish calibration for the plant library, where assay mechanisms and chemical distributions may differ. For TDP1, sparse remote-stratum actives add uncertainty to an already unfavorable coverage assessment.

The structural results answer a narrower question. Figure 3 illustrates reference-pose recovery and the selected ligand poses under the prepared receptor conditions, whereas Figure S4 provides coordinate-derived residue-contact maps and atom-level distances. The coordinate record is the auditable one, since pose identities and distances can be rechecked against the supplied files. Because α-amyrin is experimentally weak and several preparations scored within a narrow range, a favorable-looking pose cannot establish potency. The interaction categories drawn in Figure 3 are assigned by the rendering software and do not add an independent line of evidence. Experimental CRC work supplies a target rationale (Hoekstra *et al.*, 2016), rather than demonstrating that these phytochemicals engage PTP1B.

Target transfer and QSAR also share ChEMBL-derived information. Their agreement should therefore be interpreted as consistency within overlapping chemical evidence, not independent biological replication. Conversely, disagreement need not invalidate either calculation: transfer asks whether related ligands have annotations for a protein, whereas the quantitative models ask about a defined activity population after stricter curation and leakage exclusion. This distinction explains why broad transferred support can coexist with limited prospective coverage or low stringent-threshold probabilities.

Covalent reactivity is a separate mechanistic consideration. Primary experiments have demonstrated thiol modification by cucurbitacins D, E and I, including modification of Cofilin1 (Gabrielsen *et al.*, 2013). That evidence covers particular cucurbitacins and particular targets, so it cannot be extended to every derivative or to PTP1B. The docking protocol used here did not model covalent bond formation, and the QSAR labels were not stratified by reversible versus covalent mechanism, so these results can neither identify nor exclude a covalent mechanism.

Experimental follow-up should begin with compounds whose identity, purity and stereochemistry have been independently established, or with authenticated plant material accompanied by analytical profiling. Eighteen structures lacked an exact PubChem identity resolution in this workflow; that database result does not itself mean that 18 chemical identities are unknown. The unassigned stereocenter in LTS0248750 is a separate, experimentally relevant uncertainty. Neither a database name nor a favorable computational score substitutes for chemical characterization of the material used in an assay.

A focused biochemical panel would be more informative than expanding the docking list. It could include the experimentally weak α-amyrin comparator, structurally distinct in-domain compounds selected across the 5.0-threshold ranking, and a small set of lower-ranked controls. LTS0050488 and stereochemically resolved material corresponding to LTS0248750 would address the saved structural hypotheses, without being designated hits. Concentration–response measurements should be accompanied by solubility checks, compound-only signal controls and an orthogonal readout. Where relevant to the assay, aggregation and redox-sensitive or time-dependent effects should be examined before assigning a binding mechanism. These are proposed experiments, not results of the present study.

Biochemical activity would still need to be connected to CRC biology. Follow-up should distinguish general cytotoxicity from pathway-specific effects, assess exposure conditions, and use an appropriate target-engagement or genetic perturbation approach to test PTP1B dependence. A cellular response by itself would not establish direct engagement, and a purified-enzyme result would not establish cellular efficacy. This staged approach makes the secondary-target conclusion experimentally actionable while preserving the distinction between chemical prioritization and mechanism.

LIMITATIONS

The library consists of database-associated structures rather than an analytical profile of authenticated material. Occurrence, abundance and stereochemistry are incompletely established, and the structural-family rules are descriptive. *Target* transfer used bounded neighbor retrieval and depends on database coverage, assay context and filtering. Source-document counts do not establish independent replication, the endpoint/source strata were retrospective, and leave-one-source-ID-out results do not

estimate evidence quality. Enrichment used a small, partly disease-selected gene set and the service background.

Exact quantitative activity records represent a selected population: exclusion of censored observations and missing source graphs can change which measurements are available. IC_{50}-only sensitivity reduces endpoint mixing without removing assay heterogeneity. Nested validation remains internal validation on one three-fold outer scaffold assignment, and bootstrap intervals condition on fitted models. The added similarity-stratified checks are *post hoc* and do not establish causal effects of chemical distance or independently calibrate prospective probabilities. Very small TDP1 strata limit precision. Fingerprints omit chirality, and no experimental calibration in the plant library is available.

Docking covered six preparations with rigid receptors and limited state sampling. Reference-pose recovery evaluates geometry, not affinity ranking, screening enrichment or inhibition. *Molecular* dynamics simulations and free-energy calculations were not performed. The Discovery Studio composite used for Figure 3 has limited resolution and clipped source labels, and its interaction assignments could not be independently regenerated from an original session; the coordinate-derived maps in Figure S4 carry the auditable distances. No biochemical, cellular, animal or clinical experiments were performed.

CONCLUSION

The 45 reported *C. colocynthis* structures span diverse chemical space, but the available computational evidence does not establish a prospective direct CRC inhibitor. Source omission exposes dependence within the initial target shortlist, while held-out similarity diagnostics support caution when applying activity models to remote chemistry. PTPN1 remains a secondary hypothesis for focused biochemical testing; TDP1 rankings were withheld because no library structure met the operational applicability criterion. The curated compound and assay records, leakage-controlled validation and pose-specific structural evidence provide a reproducible basis for testing these bounded hypotheses.

ABBREVIATIONS

AP: Average Precision; **ChEMBL:** Chemical Biology Database; **cLogP:** Calculated Logarithm of the Partition Coefficient; **CRC:** Colorectal Cancer; **DOI:** Digital Object Identifier; **EC50:** Half-Maximal Effective Concentration; **ECFP4:** Extended-Connectivity Fingerprint (Diameter 4); **EGFR:** Epidermal Growth Factor Receptor; **FDR:** False Discovery Rate; **HBA:** Hydrogen-Bond Acceptors; **HBD:** Hydrogen-Bond Donors; **IC50:** Half-Maximal Inhibitory Concentration; **InChIKey:** International Chemical Identifier Key; **JAK2:** Janus Kinase 2; **Kd:** Dissociation Constant; **Ki:** Inhibition Constant; **MAE:** Mean Absolute Error; **MCC:** Matthews Correlation Coefficient; **MMFF:** Merck *Molecular* Force Field; **MW:** *Molecular* Weight; **PAINS:** Pan-Assay Interference Compounds; **PCA:** Principal Component Analysis; **PDB:** Protein Data Bank; **pChEMBL:** Negative Logarithm of Molar Activity Concentration; **PC1:** First Principal Component; **PC2:** Second Principal Component; **PTP1B:** Protein Tyrosine Phosphatase 1B; **PTPN1:** Protein Tyrosine Phosphatase Non-Receptor Type 1; **qHTS:** Quantitative High-Throughput Screening; **QED:** Quantitative Estimate of Drug-Likeness; **R^2:** Coefficient of Determination; **RDKit:** Open-Source Cheminformatics Toolkit; **RMSE:** Root Mean Square Error; **RMSD:** Root Mean Square Deviation; **ROC-AUC:** Receiver Operating Characteristic–Area Under the Curve; **RotB:** Rotatable Bonds; **SMILES:** Simplified *Molecular* Input Line Entry System; **STAT3:** Signal Transducer and Activator of Transcription 3; **TDP1:** Tyrosyl-DNA Phosphodiesterase 1; **TPSA:** Topological Polar Surface Area; **UniProt:** Universal Protein Resource; **Vina:** AutoDock Vina *Molecular* Docking Software.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest.

DATA AVAILABILITY

Curated structures, descriptors, structural groups, activity and assay records, target evidence, scaffold assignments, model outputs and saved docking coordinates are supplied in Supplementary Data S1-S4 includes source-ablation table include held-out similarity-stratified diagnostics. The accompanying reproducibility archive preserves source code, frozen input hashes, analysis settings and file-level checksums.

REFERENCES

Abdulridha, M. K., Al-Marzoqi, A.-H., & Ghasemian, A. (2020). The anticancer efficiency of *Citrullus colocynthis* toward the colorectal cancer therapy. *Journal of Gastrointestinal Cancer*, *51*(2), 439–444. https://doi.org/10.1007/s12029-019-00299-6

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., Orhan, I. E., Banach, M., et al. (2021). Natural products in drug discovery: Advances and opportunities. *Nature Reviews Drug Discovery*, *20*(3), 200–216. https://doi.org/10.1038/s41573-020-00114-z

Baell, J. B., & 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. https://doi.org/10.1021/jm901137j

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, *39*(15), 2887–2893. https://doi.org/10.1021/jm9602928

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, *4*(2), 90–98. https://doi.org/10.1038/nchem.1243

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Butina, D. (1999). Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, *39*(4), 747–750. https://doi.org/10.1021/ci9803381

Chawech, R., Jarraya, R., Girardi, C., Vansteelandt, M., Marti, G., et al. (2015). Cucurbitacins from the leaves of *Citrullus colocynthis* (L.) Schrad. *Molecules*, *20*(10), 18001–18015. https://doi.org/10.3390/molecules201018001

Chen, T., & Guestrin, C. (2016, August 13). *XGBoost: A scalable tree boosting system* [Paper presentation]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Dandawate, P., Subramaniam, D., Panovich, P., Standing, D., Krishnamachary, B., et al. (2020). Cucurbitacin B and I inhibits colon cancer growth by targeting the Notch signaling pathway. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-57940-9

Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*, *61*(8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

Gabrielsen, M., Schuldt, M., Munro, J., Borucka, D., Cameron, J., et al. (2013). Cucurbitacin covalent bonding to cysteine thiols: The filamentous-actin severing protein Cofilin1 as an exemplary target. *Cell Communication and Signaling*, *11*(1). https://doi.org/10.1186/1478-811x-11-58

Hoekstra, E., Das, A. M., Swets, M., Cao, W., van der Woude, C. J., et al. (2016). Increased PTP1B expression and phosphatase activity in colorectal cancer results in a more invasive phenotype and worse patient outcome. *Oncotarget*, *7*(16), 21922–21938. https://doi.org/10.18632/oncotarget.7829

Hussain, A. I., Rathore, H. A., Sattar, M. Z. A., Chatha, S. A. S., Sarker, S. D., et al. (2014). *Citrullus colocynthis* (L.) Schrad (bitter apple fruit): A review of its phytochemistry, pharmacology, traditional uses and nutritional potential. *Journal of Ethnopharmacology*, *155*(1), 54–66. https://doi.org/10.1016/j.jep.2014.06.011

Li, Q.-Y., Munawar, M., Saeed, M., Shen, J.-Q., Khan, M. S., et al. (2022). *Citrullus colocynthis*. In *Frontiers in Pharmacology* (Vol. 12). Schrad : Promising traditional uses, pharmacological effects, aspects, and potential applications. Frontiers in Pharmacology, 12. https://doi.org/10.3389/fphar.2021.791049

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1. *Advanced Drug Delivery Reviews*, *46*(1–3), 3–26. https://doi.org/10.1016/s0169-409x(00)00129-0

Mandour, Y. M., Refaat, E., & Hassanein, H. D. (2023). Anticancer activity, phytochemical investigation and molecular docking insights of *Citrullus colocynthis*. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-46867-6

Meisenberg, C., Gilbert, D. C., Chalmers, A., Haley, V., Gollins, S., et al. (2015). Clinical and cellular roles for TDP1 and TOP1 in modulating colorectal cancer response to irinotecan. *Molecular Cancer Therapeutics*, *14*(2), 575–585. https://doi.org/10.1158/1535-7163.mct-14-0762

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., et al. (2021). Open Targets Platform: Supporting systematic drug–target identification and prioritisation. In *Nucleic Acids Research* (Vol. 49, Issue D1, pp. D1302–D1310). Nucleic Acids Research, 49 , D1302–D1310. https://doi.org/10.1093/nar/gkaa1027

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

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

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., et al. (2022). The LOTUS initiative for open knowledge management in natural products research. In *eLife* (Vol. 11). eLife, 11. https://doi.org/10.7554/elife.70780

Sadybekov, A. V., & Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature*, *616*(7958), 673–685. https://doi.org/10.1038/s41586-023-05905-z

Siegel, R. L., Wagle, N. S., Star, J., Kratzer, T. B., Smith, R. A., et al. (2026). Colorectal cancer statistics, 2026. *CA: A Cancer Journal for Clinicians, 76*(2), *76*(2). https://doi.org/10.3322/caac.70067

Simoben, C. V., Babiaka, S. B., Moumbock, A. F. A., Namba-Nzanguim, C. T., Eni, D. B., et al. (2023). Challenges in natural product-based drug discovery assisted with in silico-based methods. *RSC Advances*, *13*(45), 31578–31594. https://doi.org/10.1039/d3ra06831e

Sung, H., Filho, A. M., Laversanne, M., Ferlay, J., Siegel, R. L., et al. (2026). Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA: A Cancer Journal for Clinicians, 76*(4), *76*(4). https://doi.org/10.3322/caac.70090

Szklarczyk, D., Nastou, K., Koutrouli, M., Kirsch, R., Mehryary, F., et al. (2025). The STRING database in 2025: Protein networks with directionality of regulation. In *Nucleic Acids Research* (Vol. 53, Issue D1, pp. D730–D737). Nucleic Acids Research, 53 , D730–D737. https://doi.org/10.1093/nar/gkae1113

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics*, *29*(6–7), 476–488. https://doi.org/10.1002/minf.201000061

Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., et al. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry*, *45*(12), 2615–2623. https://doi.org/10.1021/jm020017n

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., et al. (2024). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. In *Nucleic Acids Research* (Vol. 52, Issue D1, pp. D1180–D1192). Nucleic Acids Research, 52 , D1180–D1192. https://doi.org/10.1093/nar/gkad1004

Zhang, H., Zhao, B., Wei, H., Zeng, H., Sheng, D., et al. (2022). Cucurbitacin B controls M2 macrophage polarization to suppresses metastasis via targeting JAK-2/STAT3 signalling pathway in colorectal cancer. *Journal of Ethnopharmacology*, *287*, 114915. https://doi.org/10.1016/j.jep.2021.114915

Cite this article: Alshaghdali K. Chemical-Space Characterization and *Target* Assessment of Reported *Citrullus colocynthis* Phytochemicals in Colorectal Cancer. Pharmacog Res. 2027;19(2):856-69.

SUPPLEMENTARY MATERIAL

Table S1: PTPN1 model and docking de-risking summary.

Metric	Value
Primary modeling dataset	3229 exact-relation IC50/Ki records; 2662 standardized unique structures
Primary activity threshold	pChEMBL >= 6.0; 572 active / 2090 inactive
Selected classifier	SVM
ROC-AUC	0.883
PR-AUC	0.643
MCC	0.587
Balanced accuracy	0.806
Sensitivity	0.722
Specificity	0.889
F1	0.680
Brier score	0.110
Applicability domain	21/45 in-domain in primary Stage-3 analysis
Alpha-amyrin experimental pChEMBL	4.71
2F71 redocking RMSD (Å)	0.826
2QBS redocking RMSD (Å)	0.354
Alpha-amyrin Vina score (kcal/mol)	-7.699
LTS0050488 Vina score (kcal/mol)	-7.57
LTS0248750 stereoisomer 2 Vina score (kcal/mol)	-7.32
LTS0102199 Vina score (kcal/mol)	-6.14
Final decision	No direct PTP1B hit supported

Table S2: PTP1B docking-score summary.

Compound	LOTUS ID	Role	Vina score (kcal/mol)	Interpretation
Alpha-amyrin	LTS0222826	experimentally characterized weak comparator	-7.699	Measured pChEMBL 4.71; best docking score
LTS0050488	LTS0050488	prospective	-7.57	Best prospective score; pose-degenerate
LTS0248750 stereoisomer 2	LTS0248750	conditional prospective	-7.32	Strong catalytic-contact pattern; native stereochemistry unresolved
LTS0102199	LTS0102199	prospective backup	-6.14	Weaker structural support

Table S3: TDP1 model summary.

Metric	Value
Retrieved quantitative records	695
Qualified primary records	407
Standardized unique structures	323
pChEMBL 5.0 active/inactive	181/142
pChEMBL 5.5 active/inactive	123/200
pChEMBL 6.0 active/inactive	71/252
ROC-AUC at 5.0/5.5/6.0	0.889/0.862/0.902
PR-AUC at 5.0/5.5/6.0	0.934/0.765/0.756
MCC at 5.0/5.5/6.0	0.582/0.573/0.631
Balanced accuracy at 5.0/5.5/6.0	0.791/0.765/0.780
XGBoost regression R2	0.768
XGBoost regression RMSE	0.429
XGBoost regression MAE	0.340
XGBoost regression Spearman rho	0.879
Applicability domain	0 in-domain / 4 borderline / 41 out-of-domain
Final decision	Rejected from direct-hit discovery arm

Table S4: TDP1 applicability-domain summary.

Applicability-domain class	Compound count	Definition
In-domain	0	Maximum training similarity >=0.50
Borderline	4	0.35 <= maximum training similarity <0.50
Out-of-domain	41	Maximum training similarity <0.35