Sugi Predict: An Open System for Target Prediction across Patent Chemical Space

Tamer Gür

Rottweil, Germany

tamer.gur07@gmail.com

July 1, 2026

Abstract

Two questions recur in drug discovery and chemical biology: what a given compound is likely to do, and what chemistry has already been explored around a given target. Much of the chemistry that bears on them sits in the patent record, which discloses tens of millions of compounds and is often where new chemical matter appears first. We present *Sugi Predict*, an open, reproducible system that predicts human protein targets across patent chemical space. Every drug-like compound in SureChEMBL (roughly 30 million) is assigned its predicted human protein targets by chemical k-nearest-neighbour transfer from an uncapped reference of 1.25 million ChEMBL ligand–target pairs. The transfer follows the established similarity principle, but each prediction is reported as two separate signals rather than one score, a confidence calibrated from chemical similarity and a count of supporting reference neighbours, with a known-compound flag, so that a single close analogue is not read like a broad consensus of weaker matches. Sugi Predict answers two questions from one index: for a molecule, its predicted targets; and, in the reverse direction, for a target the patented chemistry predicted against it, joined to patent number, assignee, date, and claim status. We validate the prediction across four test conditions of increasing difficulty: a leave-one-out interpolation upper bound (83.0 % recall@1), a scaffold split (76.5 %), a 1,556-drug named-target panel (54 % within the top 5), and a temporal split on chemistry disclosed after the reference was frozen (40.8 %, a prospective test). Accuracy throughout is proportional to chemical similarity to known chemistry, and we drop predictive routes that fail the same test. Sugi Predict is browsable as a web application at https://sugi.bio/predict.

1 Introduction

Much of drug discovery and chemical biology turns on two recurring questions about the molecules and targets at hand: what is a given compound likely to do, and what has already been explored around a given target. These questions are put not only by researchers but also by the large language models and autonomous agents now used to triage and reason over biomedical knowledge [1, 2]. A large share of the chemistry that bears on them sits in the patent record, which discloses tens of millions of compounds and is often where new chemical matter appears first. The targets those compounds may act on, together with the clinical trials and proteins connected to them, speak directly to both of these questions.

A number of well-developed tools address parts of this. Ligand-based target prediction, inferring a molecule's protein targets from its structure, is mature: methods such as the Similarity Ensemble Approach [3, 4], SwissTargetPrediction [5, 6], the Polypharmacology Browser [7], and SuperPred [8] are openly available and widely used, transferring activity from known ligands to a query by chemical similarity. They are, however, built on curated bioactivity databases

such as ChEMBL [9], rather than the patent record. The patent chemistry itself is well served on the structure side: SureChEMBL [10] makes tens of millions of patent compounds openly structure-searchable, and commercial platforms such as SciFinder, Reaxys, and PatSnap offer extensive patent search. Each of these tools covers part of the problem; what they do not provide is a combined view: a predicted-target view over chemistry at patent scale, open and reproducible. That is the gap this work addresses.

We present *Sugi Predict*, an open, reproducible system that predicts the human protein targets of a small molecule and serves those predictions at the scale of the patent record. Targets are assigned by an established principle, that chemically similar molecules tend to share targets: a query compound is matched by its chemical fingerprint to a reference of known ChEMBL ligands, and the targets of its nearest neighbours are carried over with a confidence calibrated to the accuracy observed at that similarity, not the bare similarity itself. Applied across the roughly thirty million drug-like compounds in SureChEMBL, it yields a predicted-target view of patent chemical space; and from a single index it answers both questions posed above, from a compound to its predicted targets and, inverting the same search, from a target to the patented chemistry predicted against it, each surfaced compound carrying its provenance: the patents that claim it, their assignees, and their dates. Alongside the chemistry it holds supporting context, clinical trials, proteins, and patent text used for retrieval rather than prediction, and it resolves every result to a stable gene, protein, or disease identifier through BioBTree [11]. Its predictions are stored in a vector database and served as a web application (Figure 1).

We evaluate these predictions on held-out chemistry of increasing difficulty and report how accuracy tracks a query's chemical similarity to known compounds; held-out recovery is an interpolation upper bound on accuracy.

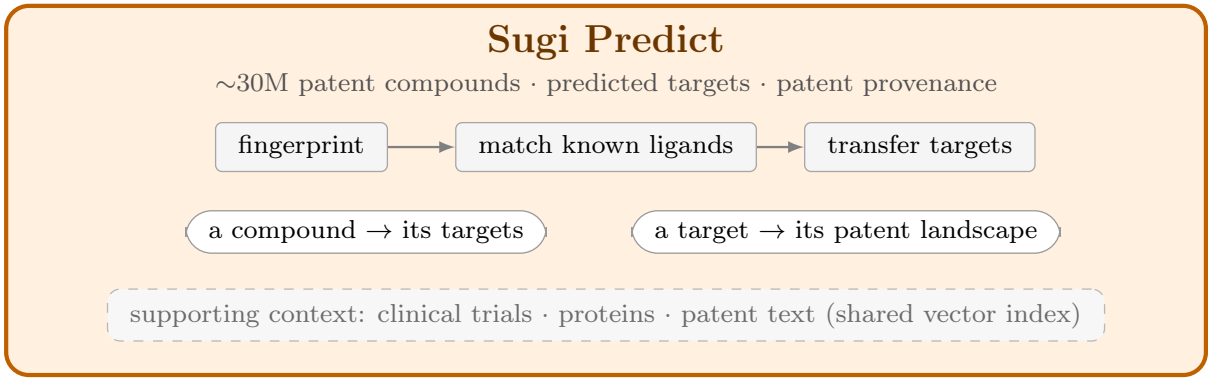

Figure 1: **Sugi Predict at a glance.** Two inputs, roughly thirty million patent compounds from SureChEMBL and a reference of known ChEMBL ligand–target pairs, feed a single chemical-similarity prediction: each compound is represented as a chemical fingerprint, matched to its nearest known ligands, and assigned their targets with a confidence set by how close the match is and a count of supporting neighbours. Applied at scale it yields an open, predicted-target view of patent chemical space, in which every compound carries its predicted targets and patent provenance, answering both query directions (compound → targets and target → patent landscape). Predictions are stored in a vector database, grounded to stable gene, protein, and disease identifiers through BioBTree, and backed by supporting context (clinical trials, proteins, and patent text). Figures 2 and 3 detail the prediction method and the data engine.

2 Methods

We first describe the prediction method and how it is evaluated (Sections 2.1 and 2.2), then the data engine that applies it across the patent corpus and serves the predictions (Section 2.3).

2.1 Sugi Predict

Sugi Predict applies an established principle, that structurally similar molecules tend to bind similar targets [12], to patent chemistry at scale. A query molecule is represented as a chemical fingerprint and compared against a reference of known ligand–target pairs drawn from ChEMBL; the targets of its nearest reference neighbours are transferred to it and graded by how close the match is, so a prediction states that the molecule lies in the chemical neighbourhood of a target's known ligands rather than asserting activity. Run once over the full SureChEMBL corpus, the same procedure annotates every drug-like patent compound with its predicted targets, and inverting the same index turns a target into the patented chemistry predicted against it; predicted targets are resolved to genes and proteins through BioBTree [11], and the predictor is checked on held-out chemistry. The same vector index also returns the patent compounds most chemically similar to a query, with clinical trials and proteins held as supporting context that complements the predictions. The individual steps follow established similarity-based target-prediction work, including SEA [3, 4], SwissTargetPrediction [5, 6], and PPB2 [7], which we adopt and run openly and reproducibly over the patent record. The reference set, the prediction step, and the patent-scale annotation are described in turn below. Figure 2 summarises the prediction method.

Reference ligand–target set. The reference is a set of known ligand–target pairs from ChEMBL [9]. It comprises 1,248,456 distinct compounds and 2,815,184 ligand–target associations across 7,929 protein targets (uncapped: all reported ligands per target are kept, since the exact-Tanimoto search scales to the full reference and capping would discard the chemical diversity of well-studied targets). Of these, 4,884 are human targets; the rest are non-human orthologs, each resolvable to its human gene where one exists. Each compound is represented by a Morgan (ECFP4) fingerprint (radius 2, 2048 bits, RDKit [13]), and the similarity between two compounds is the exact Tanimoto coefficient of their fingerprints. Each target carries its UniProt and gene identifier, looked up once through BioBTree [11] so that a transferred target resolves to a named protein.

Prediction by chemical k-NN. Prediction uses a single modality, small-molecule chemical similarity; two other predictive routes we tested did not validate and are not used (Section 2.2.6). For a query molecule, the nearest reference compounds by Tanimoto similarity are retrieved and their targets transferred, ranked by a similarity-derived confidence. The confidence is the nearest-neighbour Tanimoto; a separate support count gives how many of the 20 nearest neighbours share the predicted target, reported alongside it rather than folded in. Predicted targets are reported at the human gene level: non-human orthologs in the reference are collapsed to their human counterpart, which removes spurious non-human off-targets from the ranking. A prediction is thus an explicit statement that a molecule lies in the chemical neighbourhood of known ligands of a target, not an assertion of activity. Patent membership, chemical proximity, and confirmed activity are distinct, and we claim only the first two. Below a nearest-neighbour Tanimoto of 0.3 a query has no close known analogue and is labelled *novel* (chemical whitespace), to be read with caution rather than as a confident prediction.

Patent-scale annotation. The same procedure annotates every drug-like patent compound (those with at least ten heavy atoms, which excludes solvents and fragments) among the 30,937,359 patent compounds extracted from SureChEMBL [10]. The neighbour search is run once on GPU with an exact-Tanimoto kernel, storing each compound's top-100 raw neighbours; ranking and confidence are then computed locally and cheaply, so the scoring can be re-tuned without re-searching. Because the annotation is a deterministic function of the reference set and the patent corpus, it is regenerated whenever either is refreshed. This yields target annotations for the 30,046,840 drug-like compounds (of 30,937,359): 39.8 % fall in the high-confidence band

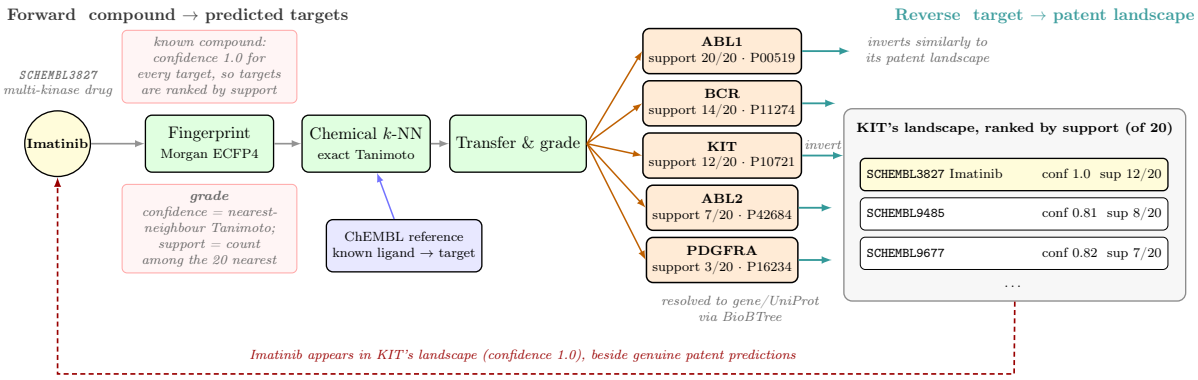

Figure 2: The prediction method, worked on a multi-target example. A query molecule (here Imatinib, SureChEMBL `SCHEMBL3827`) is represented as a Morgan (ECFP4) fingerprint, matched to its nearest known ligands in the ChEMBL reference by exact Tanimoto, and their targets are transferred and graded: confidence is the nearest-neighbour Tanimoto and support is the count among the 20 nearest neighbours. The compound fans out to its predicted targets, ranked by support (ABL1 20/20, BCR 14/20, KIT 12/20, ABL2 7/20, PDGFRA 3/20); because Imatinib is itself in the reference, its confidence saturates at 1.0, so support orders its targets. Predicted targets are resolved to gene and UniProt identifiers through BioBTree. Inverting the same annotated index turns a target into its patent landscape: KIT's landscape, ranked by support, contains Imatinib (confidence 1.0) alongside genuine patent-compound predictions such as `SCHEMBL9485` (0.81) and `SCHEMBL9677` (0.82). The same compound thus appears in both directions, as a query and within the landscape of a target it is predicted against.

(nearest known ligand at Tanimoto ≥0.5), 68.4 % at ≥0.4, and only 6.5 % are novel chemical whitespace with no known analogue at ≥0.3 (median nearest-ligand Tanimoto 0.457). The uncapped reference thus covers a large fraction of drug-like patent chemical space, though confidence (and, as the validation shows, reliability) falls with chemical similarity. Inverting the same index gives the second query direction, from a target to the patented chemical matter predicted against it (Section 3). Because each compound carries its SureChEMBL identifier, this view joins directly to patent provenance (the patent number(s), assignee, publication date, and whether the compound is *claimed* as opposed to merely disclosed) from the same owned snapshot, linking each compound in the landscape to its patent provenance and thereby to who is patenting which chemistry against a target. (Locally held dates are publication dates; priority and filing dates and a normalized applicant are fetched on demand from EPO Open Patent Services for the patents shown.)

2.2 Validation

A prediction is only useful if it comes with a sense of when to trust it. We therefore measure accuracy directly, on compounds set aside from the reference (held out, so the method has not seen them). We test the forward direction (compound to target) under four conditions of increasing difficulty, from a leave-one-out bound through a scaffold split and the recovery of known drugs' targets to a temporal split that mimics predicting on newly disclosed chemistry; we compare against the established methods Sugi Predict descends from, and report how its confidence score calibrates to accuracy. We then validate the reverse direction (from a target to the chemistry predicted against it) against known annotations, and corroborate it with an independent signal from the compounds' own patent text. Finally we note the predictive routes we tested and dropped. Unless stated otherwise, figures are means over five random samples of 2,000 test compounds, each scored against all 7,929 targets and ranked by Sugi Predict's confidence (Table 1; ± is 1 SD across samples).

Table 1: Validation across four conditions of increasing difficulty (Sugi Predict's max-Tanimoto ranking; predicting among all 7,929 targets; splits are 5-seed means). Accuracy declines toward the prospective test, staying above the popularity ($\leq 6.2\,\%$ recall@1) and random ($\leq 0.1\,\%$) baselines throughout. The drug panel is mechanism-target gene recovery over 1,556 drugs (Table 2).

Condition	Recall@1	Recall@5	Recall@10	What it measures
Leave-one-out	83.0 %	92.6 %	95.6 %	interpolation upper bound
Scaffold split	76.5 %	87.7 %	92.1 %	generalization across chemotypes
Drug panel*	34 %	54 %	60 %	recovery of a named target
Temporal split (>2018)	40.8 %	55.7 %	60.9 %	generalization over time ($\approx$ patents)

* 1,556 approved drugs, mechanism-target gene recovery, each with its own Bemis–Murcko scaffold withheld (Table 2).

2.2.1 Accuracy on unseen chemistry

A leave-one-out evaluation, removing each test compound and its near-duplicates from the reference before predicting, yields recall@1 of 83.0 % and recall@5 of 92.6 %. This is an *interpolation upper bound*: the reference is chemically redundant, so leave-one-out largely measures retrieval within known chemotype space rather than generalization to the novel chemistry typical of patents. Generalization is tested instead by a *scaffold split*, in which every reference ligand sharing the query's Bemis–Murcko scaffold is withheld, holding out the query's analogue series (acyclic queries, for which the split is undefined, are excluded; 0.4 % of the sample). Under this stricter condition recall@1 is 76.5 % and recall@5 87.7 %, a drop of 6.5 points from the upper bound, indicating the ranking is not driven solely by same-series leakage; a stricter generic-framework scaffold, which also collapses atom identity, gives 74.9 %, so the result is robust to the scaffold definition. Both exceed the popularity baseline (6.2 % recall@1) and the random baseline (0.1 %).

The two splits above sample random reference compounds; we next ran the same test on a panel of 1,556 recognizable approved drugs (every ChEMBL small-molecule drug with an annotated mechanism-of-action target gene present in the reference), predicting each with its own Bemis–Murcko scaffold withheld and scoring whether the drug's mechanism gene is recovered. The gene is recovered within the top 5 for 54.0 % of the panel (95% CI 51.6–56.5 %) and within the top 10 for 60.5 % (58.0–62.9 %); at strict rank 1 it is 34.1 % (31.7–36.5 %), the harder bar of placing the single primary target first among 7,929. Recovery is strongly target-class dependent and tracks how densely each class is represented in the reference (Table 2): well-studied classes recover at top-5 rates comparable to the splits above (proteases 88 %, kinases 85 %), while sparsely-liganded classes are lower (ion channels 20 %). This is the coverage-dependence seen throughout, the prediction reliable where known chemistry is dense and declining where it is thin.

Finally, a *temporal split*, training the reference only on ligands first disclosed up to 2018 (668,055 ligands) and testing on 2,000 compounds disclosed afterwards, with all post-cutoff chemistry excluded from the neighbour pool, is the closest proxy to the real task of predicting targets for newly disclosed (e.g. patent) chemistry. It is the hardest of the splits, as expected: recall@1 of 40.8 % and recall@5 of 55.7 %, against a popularity baseline of 0.4 %. Accuracy again tracks similarity to known chemistry: 76 % recall@1 where a pre-cutoff analogue at Tanimoto $\geq$0.7 exists, falling to 23 % in the 0.3–0.5 band where new chemistry typically lands. Taken together, the four conditions (Table 1) decline in order from the interpolation upper bound to this most realistic test, all above the popularity and random baselines, with performance throughout proportional to chemical similarity to known ligands. The substantive comparators are the established methods compared next, not these baselines.

Table 2: Mechanism-target recovery on the 1,556-drug panel, by target class. Each drug is predicted with its own Bemis–Murcko scaffold withheld; recovery means the drug's mechanism-of-action gene appears in the predicted top-k of 7,929. Recovery tracks how densely each class is liganded in the reference; the major pharmacological classes are shown and "All" is the full panel.

Target class	n	Recall@1	Recall@5	Recall@10
Protease	43	74 %	88 %	91 %
Kinase	121	67 %	85 %	87 %
Nuclear receptor	185	44 %	67 %	72 %
GPCR	472	35 %	61 %	72 %
Other enzyme	259	32 %	52 %	57 %
Transporter	90	27 %	53 %	62 %
Ion channel	195	14 %	20 %	23 %
All	1,556	34 %	54 %	60 %

Table 3: Comparison with established baselines on the temporal split (the prospective regime; 2,000 post-2018 test compounds, identical no-leakage retrieval, predicting among all pre-cutoff targets). Sugi Predict is highest at recall@1; its margin over plain single-nearest-neighbour transfer is modest at rank 1 and widens in the candidate list.

Method	Recall@1	Recall@5	Recall@10
Sugi Predict (confidence + support)	40.8 %	55.7 %	61.0 %
Single nearest neighbour	37.5 %	46.7 %	48.5 %
SEA-style (set-size correction)	32.2 %	46.9 %	50.5 %
Naive Bayes (Morgan bits)	20.8 %	46.4 %	54.7 %
Popularity (floor)	0.4 %	3.9 %	6.7 %

2.2.2 Comparison with established methods

A SEA-style set-size correction, the refinement that distinguishes the Similarity Ensemble Approach [3, 4] from plain nearest-neighbour transfer, did not improve retrieval here: on the scaffold split it reached recall@1 of 62 % (best 66 % across thresholds) against Sugi Predict's max-Tanimoto ranking, because the set-size penalty demotes a true target precisely when it is itself well annotated. We therefore retain the simpler ranking, noting that the correction serves a different objective, controlling false-positive significance across a whole target panel, than the top-k retrieval evaluated here.

On the prospective (temporal) split we compare Sugi Predict's ranking against the methods it descends from and the popularity and random baselines, all on the identical test compounds with no post-cutoff chemistry in the neighbour pool (Table 3). Sugi Predict's confidence+support ranking is highest at recall@1 (40.8 %), ahead of a single-nearest-neighbour transfer (37.5 %), a SEA-style set-size correction (32.2 %, again trailing plain transfer, consistent with the scaffold split above), a Bernoulli naive-Bayes over fingerprint bits (20.8 %), and the popularity floor (0.4 %). The reading is twofold: single-nearest-neighbour transfer accounts for most of the recall@1 accuracy (most of the top-1 accuracy comes from the single closest prior ligand), and the k-neighbour aggregation contributes further down the list, where it lifts recall@5/@10 to 55.7/61.0 % against single-NN's 46.7/48.5 % as one ligand's targets are exhausted.

2.2.3 Confidence calibration

Confidence is calibrated to the nearest-neighbour Tanimoto so that a reported score reflects observed accuracy in that band (Table 4); for exact-match queries many targets tie at the top, so we present a support count and a "known" flag rather than a bare maximal score.

Table 4: Confidence calibration: recall@1 by nearest-neighbour Tanimoto band, under the scaffold-split evaluation. The reported confidence (the nearest-neighbour Tanimoto) maps to an observed accuracy, so a prediction can be trusted in proportion to it; below 0.3 a query has no close known analogue and is treated as out of domain.

Nearest-neighbour Tanimoto	Recall@1	n
≥ 0.7	89 %	1,153
0.5–0.7	64 %	761
0.3–0.5	26 %	85
< 0.3 (novel)	out of domain	—

2.2.4 The reverse direction

The two directions share one index, so the forward accuracy above bounds the reverse; we also measured the reverse directly, as a precision question on the scaffold-split test compounds: when a compound is surfaced in a target's landscape at per-target confidence $\geq t$, how often is that target already annotated for the compound? Precision rises with confidence, from 20 % at t=0.5 to 52 % at t=0.7 (26 % and 56 % macro-averaged over targets), while 79 % of a compound's annotated targets are surfaced at t=0.5. Because ChEMBL annotations are sparse and incomplete (a compound is typically tested against only its primary targets), a high-confidence assignment to an as-yet-untested target is scored as a miss, so these are lower bounds. The reading is that the confidence ordering is informative for the landscape as well: tightening the cut to $\geq$0.7 roughly doubles precision at the cost of recall, exactly the trade-off a user browsing the top of a target's landscape controls.

2.2.5 An independent check from patent text

The reverse-direction precision above is scored against ChEMBL's own annotations, the same source the predictor draws on. As a check *independent of the chemistry*, we ask whether a landscape compound's own patent text concerns the predicted target. For the landscape compounds whose patent carries machine-readable text (about one in seven of the patents linked to the patent compounds), we score the patent text against every human target with a biomedical sentence encoder (MedCPT [14]), correcting for a background of generically text-proximal targets, and count the predicted target as supported when it ranks among the text's top ten of 4,885. Across twelve diverse targets this independent signal agrees with the support-ranked landscape for 47 % of full-text compounds, against 0.2 % for a random target, and the agreement rises with neighbour support, from 5.7 % at a single supporting neighbour to 38.7 % at eight or more, indicating that support, more than raw confidence, tracks this text-agreement signal. The agreement is strongly target-class dependent: high for proteins that are the *subject* of patents (nuclear receptors near 98 %, aminergic GPCRs near 80 %, EGFR 67 %) and low for those that appear mainly as selectivity *counter-screens* rather than the claimed target (the hERG and SCN5A channels at 2–4 %). To test independence, we permute compounds only among patents of the same technical classification; the agreement persists at about 14 times the permuted baseline, so the signal is weakly independent of patent class. It reflects what a patent is *about*, the consistency of the prediction with the inventor's stated purpose, rather than an orthogonal binding measurement; the available text is typically an abstract rather than a full specification. We therefore read it as a corroborating signal for the high-support, primary-target portion of a landscape, not a confirmatory one, and available only for the minority of compounds with patent text.

Table 5: The engine's collections (measured on the production stack: Qdrant, single 32-core / 125 GB server).

Modality	Embedding	Vectors	Dim.
Patent compounds	Morgan ECFP4 (bin.)	30,937,359	2048
Patent text	MedCPT	39,614,358	768
Clinical trials	MedCPT	4,476,465	768
Proteins	ESM-2	574,648	1280
Total		75,602,830	

2.2.6 Routes that did not validate

This harness is also a filter on what is treated as a predictor at all. Two candidate routes were tested on it and *dropped* because they did not beat baseline: peptide sequence-embedding similarity performed at chance, and protein-homology-based repurposing fell below the random baseline. Only ligand–target chemical similarity survived. We report this because it is the basis on which the surviving predictions should be trusted: each is retained because it exceeded a held-out baseline, not on the basis of selected examples.

2.3 The data engine

Sugi Predict is built on a single, reproducible data engine that turns each biomedical corpus into a collection of dense or binary vectors, indexes them for approximate nearest-neighbour search, and serves the result interactively (Figure 3). Although the prediction uses only the patent-compound modality, the engine is general, and we describe it in full because it is the part that makes patent-scale annotation and interactive serving practical. Data processing and serving are deliberately separated: building a collection's vectors uses GPU acceleration and a per-modality embedding model, while serving the result is handled by a vector database on standard CPU servers, with no GPU at query time.

Modalities and embeddings. The engine currently spans four modalities, each embedded with a model suited to it: patent compounds as Morgan (ECFP4) fingerprints [13], patent text [15] and clinical-trial sections with MedCPT [14], and UniProt [16] proteins with ESM-2 [17]. Clinical trials [18] are chunked by section so that retrieval is fine-grained rather than per-document. Together the collections hold 75,602,830 vectors (Table 5).

Indexing and serving. Each collection is built from a FAISS source and served from Qdrant [19] using an HNSW index. Memory is the binding constraint at this scale, and the pipeline is engineered around it: the binary fingerprints are kept in RAM as bit vectors (Qdrant's binary quantization, scored bitwise to approximate Tanimoto), the large dense collections keep their full vectors on disk while the HNSW graphs stay in RAM, and the whole store runs inside a memory-capped container. The result is that all 75,602,830 vectors are searched interactively on a single server. Each collection is rebuilt from a declarative per-collection profile by a single command, so a build is reproducible, and updates are applied incrementally by deterministic-identifier upsert into the existing index rather than by full rebuild.

The web application. Sugi Predict is served as a web application over the same engine (Figure 3). Four read functions back the pages: `query` (retrieve by payload filter or by similarity, embedding a free-text, SMILES, or protein query server-side), the compound page (a compound's baked predicted targets with its patent provenance), similar-compound retrieval (its nearest patent compounds by fingerprint), and `predict` (rank a molecule's targets). The site presents one page per target, carrying its patented-chemistry landscape, and one per compound; the

compound page also shows an in-place reverse view of the chemistry predicted against each of its targets. Pages are rendered on demand and cached.

Identifier lookup and what is validated. Where a retrieved record must be tied to a known entity (a predicted target to its gene symbol and UniProt accession, a compound to its registry identifiers) the engine resolves it through BioBTree [11], which maps the reference's ChEMBL target identifiers to those stable keys and so lets a predicted target join the protein and clinical-trial records sharing them. This lookup keeps results actionable and lets hits in different modalities be related through shared identifiers, but it is a convenience layer over the vector pipeline, not the subject of this paper. The engine retrieves by nearest-neighbour similarity within each modality; of those similarities, only chemical similarity is used here as a *predictor* of an unobserved property, a likely target, and it is the only one we validate as such (Section 2.2). The other modalities are retrieval assets, not predictors.

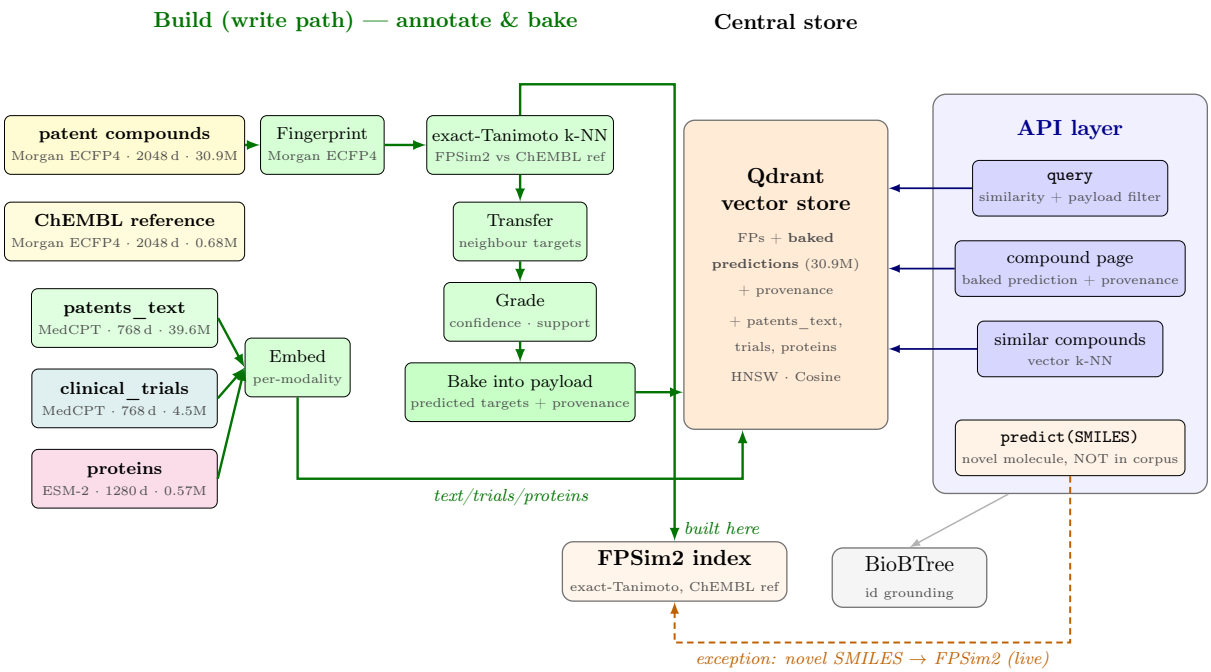

Figure 3: The data engine. Five modality collections are embedded with a model suited to each: patent compounds and the ChEMBL reference as Morgan ECFP4 fingerprints (2048 d; 30.9M and 0.68M vectors), patent text and clinical-trial sections with MedCPT (768 d; 39.6M and 4.5M), and UniProt proteins with ESM-2 (1280 d; 0.57M). At build time (write path, green) each compound is fingerprinted, matched to its nearest ChEMBL reference ligands by an exact-Tanimoto k-nearest-neighbour search, and its neighbours' targets are transferred and graded (confidence, the nearest-neighbour Tanimoto; support, the count among the nearest neighbours); the predicted targets and patent provenance are then baked into the payload and upserted into Qdrant, alongside the embeddings of the other modalities. The exact-Tanimoto index (FPSim2) built in this step is retained for serving. Qdrant is the served vector store, holding the fingerprints, the baked predictions for all 30.9M compounds, their provenance, and the patent-text, trial, and protein collections. The API layer's functions read Qdrant for almost everything: `query` (similarity and payload filter), the compound page (its baked prediction and provenance), and similar-compound retrieval (vector k-nearest-neighbour). The one live exception is `predict(SMILES)` for an arbitrary molecule that is not in the corpus, which is scored against the FPSim2 exact-Tanimoto index rather than read from Qdrant. A lightweight BioBTree lookup resolves retrieved records to gene and UniProt identifiers.

3 Use cases

Sugi Predict answers two complementary questions from one index: for a protein target, what patented chemical matter is predicted against it; and for a molecule, what target it is predicted to hit. We work each on real, recognizable examples, leading with the direction the owned patent corpus supports. Every number below is read from the served collection (Section 5).

3.1 From a target to its patent landscape

A competitive-intelligence analyst or discovery lead asks: *what chemical matter has been patented against my target, and by whom?* Sugi Predict answers by filtering its predicted annotations: a predicted-target layer over the patent corpus, queryable in either direction.

Taking the epidermal growth factor receptor (EGFR, UniProt P00533) as the target, 841,765 patent compounds carry EGFR in their predicted target profile. That set is not used flat: the per-target confidence (the Tanimoto to the nearest known EGFR ligand) narrows it to the compounds repeatedly placed in EGFR's chemical neighbourhood (Table 6), 331,991 at confidence $\geq$0.5 and 70,973 at $\geq$0.7. The narrowing has a measured effect: on held-out compounds the precision of the surfaced landscape, how often a surfaced target is already an annotated target of the compound, rises from 20 % at confidence 0.5 to 52 % at 0.7, while 79 % of a compound's annotated targets are still surfaced at 0.5 (Section 2.2). An independent check on the compounds' own patent text agrees with the high-support landscape for 47 % of full-text compounds against 0.2 % for a random target (Section 2.2). The confidence cut therefore trades breadth for reliability with a measured, not asserted, response.

Each surfaced compound is *attributable*. Because every entry carries its SureChEMBL identifier, the landscape joins to patent provenance from the same owned snapshot: patent number, country, assignee, publication date, and whether the compound appears in the patent's *claims* rather than only its description (22 % of the surfaced EGFR patents are claims; a claim marks the specific matter a patent protects, not the broad Markush enumeration of a description). Grouping the high-confidence set by assignee and date turns the landscape into an activity view, recognizable pharmaceutical filers (AstraZeneca, Boehringer Ingelheim, and Novartis among them) over a publication span from the early 1990s to the present, rather than a single count. A known-origin spot check confirms the join resolves correctly: a sildenafil-class compound (`SCHEMBL735`) traces to Pfizer's foundational PDE5 patents (EP-0463756, US-5250534), and losartan to DuPont's foundational angiotensin-II patent family. The result is a queryable, attributable view of who is patenting which chemistry against a target, assembled from openly available data. (Locally held dates are publication dates and assignees are the raw SureChEMBL strings; priority and filing dates and a normalized applicant are fetched on demand from EPO Open Patent Services for the patents shown.)

3.2 From a compound to its predicted target

A medicinal chemist holding a patent compound asks the inverse question: *what target does its chemistry most resemble?* Sugi Predict returns the prediction for any of the roughly thirty

Table 6: The EGFR patent landscape narrows as the confidence cut tightens. Confidence is the Tanimoto to the nearest known EGFR ligand; precision is the held-out reverse-direction precision at that cut (Section 2.2). Counts are from the served collection (30,937,359 patent compounds).

EGFR confidence	Patent compounds	Held-out precision
$\geq$0.7	70,973	52 %
$\geq$0.5	331,991	20 %
$\geq$0.3 (all)	841,765	—

Table 7: Worked compound→target predictions from the served collection, each with its confidence (nearest-ligand Tanimoto), support (neighbours of 20 carrying the target), and the held-out recall@1 of its calibration band (Table 4). The last row is an abstention: no reference ligand within Tanimoto 0.3, so no target is assigned.

Patent compound	Predicted target	Confidence	Support	Band recall@1
`SCHEMBL8383`	EGFR (P00533)	0.83	16/20	89 %
`SCHEMBL735`	PDE5A (O76074)	0.79	14/20	89 %
`SCHEMBL40470`	BRAF (P15056)	0.59	11/20	64 %
`SCHEMBL3669`	none (novel)	<0.3	—	out of domain

million patent compounds, pre-computed and grounded to a stable identifier, with its patent provenance attached. Table 7 works three patent compounds and one abstention.

Consider the patent quinazoline `SCHEMBL8383`, not itself a marketed drug. Sugi Predict predicts EGFR at confidence 0.83, supported by 16 of its 20 nearest known ligands, which are themselves EGFR-inhibitor quinazolines, and grounded to gene *EGFR* (UniProt P00533). The confidence is not a bare similarity: 0.83 falls in the calibration band where held-out recall@1 was 89 % (Table 4), so the score carries its own reliability. The prediction is an explicit, traceable statement that the molecule lies in the chemical neighbourhood of known EGFR ligands, not an assertion of measured activity. And because it is a patent compound, the same record carries its provenance, the patents that disclose it with their assignees and dates.

The behaviour holds across target classes and includes a principled abstention (Table 7): a sildenafil analogue against PDE5A at confidence 0.79 (the 89 % band), a diaryl amide against BRAF at 0.59 (the 64 % band, a materially weaker call the score makes explicit), and a complex polycyclic compound (`SCHEMBL3669`) with no reference ligand within Tanimoto 0.3, returned with no target rather than forced into a low-quality guess. On molecules whose targets are externally established, recovery follows the held-out drug panel (Section 2.2): gefitinib and imatinib, both kinases (the panel's best-covered class), recover EGFR and ABL1, while recovery is lower for the sparsely represented classes, as the panel reports.

The chemical neighbourhood behind a prediction is also exposed directly: each compound view lists its nearest patented compounds, with their own predicted targets and those shared with the query, and a molecule outside the corpus can be entered as a structure and matched the same way, returning the nearest patented chemistry and the targets predicted for it. This is retrieval over the shared chemical index, offered for exploration rather than as a separately validated prediction.

3.3 Across modalities

Because a predicted target resolves to a stable gene and protein identifier through BioBTree, a compound query extends in one step to the rest of the substrate. From the EGFR prediction for `SCHEMBL8383`, the shared *EGFR*/P00533 identifier links to the EGFR protein record and the EGFR clinical-trial records held in the other engine collections, so a single patent compound lands the user at the trial and protein context for its predicted target. This cross-modal assembly is a convenience of the shared identifier layer, not a separate prediction (chemical similarity is the only relationship we validate as predictive), but it connects one answer to its related clinical and protein records without leaving the index.

4 Discussion

Sugi Predict applies established similarity-based target prediction to the patent record at scale, and shows that it holds there with measured, calibrated reliability. Its contribution is not a new

predictor, the method is the similarity-transfer family, but the system: an open, reproducible, predicted-target view over roughly thirty million SureChEMBL compounds, validated on held-out chemistry and served so that it answers two questions from one index. For a target it returns the patented chemistry predicted against it and who has claimed it; for a molecule, the target its chemistry most resembles. The first direction, which the owned patent corpus uniquely supports, needs both a patent corpus and a predicted-target lens over it. Identifier grounding through BioBTree keeps each result actionable, joining a compound, its predicted target, and that target's trials and protein record in one place.

What the validation shows is consistent across the four conditions: accuracy declines from the leave-one-out interpolation upper bound, through the scaffold split that breaks chemotype overlap, to the temporal split. That last split is the most realistic of the four: it withholds chemistry disclosed after the reference was frozen, the regime a patent-scale tool actually operates in, so past annotations cannot leak into the answer. There it gives a prospective 40.8 % recall@1 (55.7 % at recall@5), ahead of the single-nearest-neighbour, SEA-style, and naive-Bayes baselines the method descends from. The recurring reading is that accuracy is proportional to a query's chemical similarity to known ligands and to how densely the target's class is represented: confidence, the nearest-ligand Tanimoto, is calibrated to observed accuracy (Table 4), strong for densely studied families such as kinases and proteases and weak for sparse ones such as ion channels, and below a similarity of 0.3 a query is out of domain. A prediction is therefore a hypothesis weighted by its own confidence and supporting-neighbour count, not a measurement. The same held-out bar is also a filter: two further routes we tried, peptide sequence-embedding similarity and protein-homology repurposing, did not clear it and are not used, so the surviving predictor is the one that earned its place.

These results support the two use cases worked in Section 3. A medicinal chemist gets a molecule's likely target with a calibrated confidence and the patents that disclose it; a competitive-intelligence analyst gets the patented chemistry predicted against a target, grouped by assignee and date, with a confidence cut that trades breadth for precision. For the landscape direction, a weakly-independent check against the compounds' own patent text corroborates the high-support portion of a target's landscape, with neighbour support, more than raw confidence, tracking that agreement.

4.1 Limitations

Several limits should be read alongside every prediction, and the per-prediction outputs are designed to surface them. Accuracy tracks reference density: it is highest for well-studied classes (kinases, GPCRs, proteases) and lower for sparsely annotated ones (ion channels, novel families), which is why each call carries a similarity band, a supporting-neighbour count, and a target class rather than a bare score. The held-out accuracies are best read as an *upper* estimate for chemistry near the reference, since even the scaffold and temporal splits draw queries from ChEMBL-like space rather than the more novel matter typical of patents. Predictions use *2D fingerprints*, so they miss activity cliffs that turn on 3D or shape, and they inherit ChEMBL's assay and publication bias. A molecule already in the reference saturates confidence at 1.0 with many ties, which the supporting-neighbour count and known-compound flag disambiguate. Patent chemistry also carries structural noise (Markush enumerations, which list many theoretical, often un-synthesized variants of a generic scaffold, along with salts and fragments), and the novel-chemistry flag marks out-of-domain matter, which includes non-drug-like structures as well as genuinely new chemotypes.

The two query directions carry different weight. The reverse (target→compounds) direction is bounded by the forward accuracy, and its measured precision (rising from 20 % at confidence 0.5 to 52 % at 0.7) is a lower bound, because a confident assignment to an as-yet-untested target scores as a miss against sparse annotation. The patent-text check is corroborating rather than confirmatory: it covers only compounds whose patent carries machine-readable text and

reflects what a patent is *about*, not a binding measurement. The landscape is bounded by what SureChEMBL extracted.

4.2 Future directions

Several extensions follow from these limits. Three-dimensional or shape-based descriptors would address the activity cliffs that 2D fingerprints miss. A larger or more balanced reference would lift the sparsely liganded classes where coverage is currently thin. The engine already holds other modalities (clinical trials, proteins, and patent text) as supporting context, used for retrieval rather than prediction; with more data and validation they could be developed into predictive routes of their own. More broadly, because every prediction resolves to a stable identifier and is served through a fast index, the substrate can act as a grounded source for the language-model and autonomous-agent workflows noted at the outset, returning identifier-resolved answers rather than free text.

5 Availability

Sugi Predict and the underlying substrate are open and accessible three ways. Because the prediction is a deterministic function of the reference set and the patent corpus, it is regenerated as those are refreshed, and it is grounded throughout to BioBTree [11].

- **Browse:** *Sugi Predict* is published per target and per compound at https://sugi.bio/predict.
- **Code:** the data-engine pipeline and substrate build code (https://github.com/tamerh/sugi-predict) is open source.

References

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. "Large language models encode clinical knowledge". In: *Nature* 620.7972 (2023), pp. 172–180.

[2] M. Kuehl, D. P. Schaub, F. Carli, L. Heumos, M. Hellmig, C. Fernández-Zapata, et al. "BioContextAI is a community hub for agentic biomedical systems". In: *Nature Biotechnology* 43.11 (2025), pp. 1755–1757.

[3] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet. "Relating protein pharmacology by ligand chemistry". In: *Nature Biotechnology* 25.2 (2007), pp. 197–206. DOI: 10.1038/nbt1284.

[4] E. Lounkine et al. "Large-scale prediction and testing of drug activity on side-effect targets". In: *Nature* 486.7403 (2012), pp. 361–367. DOI: 10.1038/nature11159.

[5] D. Gfeller, A. Grosdidier, M. Wirth, A. Daina, O. Michielin, and V. Zoete. "SwissTargetPrediction: a web server for target prediction of bioactive small molecules". In: *Nucleic Acids Research* 42.W1 (2014), W32–W38. DOI: 10.1093/nar/gku293.

[6] A. Daina, O. Michielin, and V. Zoete. "SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules". In: *Nucleic Acids Research* 47.W1 (2019), W357–W364. DOI: 10.1093/nar/gkz382.

[7] M. Awale and J.-L. Reymond. "Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning". In: *Journal of Chemical Information and Modeling* 59.1 (2019), pp. 10–17. DOI: 10.1021/acs.jcim.8b00524.

[8] K. Gallo, A. Goede, R. Preissner, and B.-O. Gohlke. "SuperPred 3.0: drug classification and target prediction—a machine learning approach". In: *Nucleic Acids Research* 50.W1 (2022), W726–W731. DOI: 10.1093/nar/gkac297.

[9] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, et al. "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods". In: *Nucleic Acids Research* 52.D1 (2024), pp. D1180–D1192.

[10] G. Papadatos et al. "SureChEMBL: a large-scale, chemically annotated patent document database". In: *Nucleic Acids Research* 44.D1 (2016), pp. D1220–D1228. DOI: 10.1093/nar/gkv1253.

[11] T. Gür. *BioBTree v2: Grounding LLM Responses with Large-Scale Structured Biomedical Data*. 2026. DOI: 10.5281/zenodo.18962899.

[12] M. A. Johnson and G. M. Maggiora, eds. *Concepts and Applications of Molecular Similarity*. Wiley, 1990.

[13] RDKit. *RDKit: Open-source cheminformatics*. https://www.rdkit.org. 2024.

[14] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, and Z. Lu. "MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval". In: *Bioinformatics* 39.11 (2023), btad651. DOI: 10.1093/bioinformatics/btad651.

[15] E. Felix. *uspto-chem: chemical structures and abstracts from USPTO patents*. https://github.com/eloyfelix/uspto-chem. 2024.

[16] The UniProt Consortium. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1 (2023), pp. D523–D531.

[17] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130. DOI: 10.1126/science.ade2574.

[18] U.S. National Library of Medicine. *ClinicalTrials.gov*. https://clinicaltrials.gov. 2026.

[19] Qdrant. *Qdrant: vector similarity search engine and database*. https://qdrant.tech. 2024.