PharmaDive: A Language Model Generates an Uncertainty-Graded Property Matrix for More Than 16,000 Compounds

Tyler M. Moore[1,*]

[1]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
*Corresponding author. Email: tymoore@pennmedicine.upenn.edu

Abstract

Structured, machine-readable knowledge about chemical compounds is distributed unevenly: expert-curated databases are deep but narrow, whereas automated resources are broad but confined to a single property class. I introduce PharmaDive, a compound-level property resource built by using a web-search-enabled large language model not as an extractor of facts stated in documents but as a structured annotator asked to render a graded judgment on a fixed battery of questions. A reasoning model with web search rated 16,116 substances—approved drugs, investigational agents, endogenous and natural substances, discontinued products, research chemicals, venoms, extracts, and fragrance ingredients—on 130 structural, physicochemical, pharmacological, and clinical attributes, using a five-point confidence scale with an explicit unknown option. Accuracy was evaluated on three tracks. Against cheminformatics ground truth computed from molecular structure, ratings were 95.7% accurate, or 99.0% excluding one item requiring enumeration rather than recognition. Against adjudication by an architecturally distinct model, pharmacological and clinical ratings reached approximately 92% weighted accuracy, roughly 15 percentage points above the majority-class baseline. Against curated external references for 2,079 matched compounds, median directional accuracy was 95.2% for structural attributes and 93.7% for mechanistically specific ones. Errors were strongly asymmetric: negative ratings were 98–99% accurate, “probably yes” was correct only 22% of the time, and 88% of errors were false-positive endorsements. Item factor analysis recovered 11 interpretable dimensions accounting for 53% of item variance, with compound-level scores showing strong face validity. PharmaDive is best used as a versioned exploratory overlay whose value is aggregate and comparative rather than cell-level and authoritative.

Keywords: large language models, chemical informatics, drug annotation, factor analysis, psychometrics, data quality

Introduction

Progress in pharmacology, medicinal chemistry, toxicology, and computational drug discovery depends on structured, machine-readable knowledge about chemical substances: what a molecule is, how it is built, what it does in a biological system, and how it behaves physicochemically. A single compound may be described across many information domains—structure and functional groups, physicochemical behavior, target and pathway interactions, pharmacokinetics, formulation, therapeutic uses, adverse effects, and preclinical or clinical evidence—yet these facts are rarely colocated. They are distributed instead across primary articles, regulatory records, product labels, curated databases, patents, and general web resources, often under multiple names and at different levels of certainty and granularity. For a researcher attempting to compare thousands of compounds, the central obstacle is therefore not information scarcity but the cost of converting heterogeneous, unevenly documented knowledge into a common matrix in which the same questions are asked of every compound and uncertainty is represented explicitly.

Two features of the problem make it hard. First, the properties of interest span radically different scales and epistemic kinds—from deterministic structural facts, such as whether a molecule contains a halogen,

to graded pharmacological judgments, such as whether a compound is an immunomodulator—so that no single measurement or computation suffices. Second, coverage is intrinsically uneven: a few thousand marketed drugs are exhaustively characterized, whereas most compounds appearing in the literature—investigational agents, endogenous metabolites, natural products, discontinued agents, reagents, and the long tail of research chemicals—are documented sparsely, in specialized venues, or only by inference from structurally related molecules. Any resource aspiring to breadth must therefore contend simultaneously with the cost of expert curation and the sparsity of the underlying evidence.

Structured Chemical Knowledge and Its Curation Bottleneck

The public infrastructure that addresses this problem is substantial and, within its domains, mature. PubChem aggregates chemical information from more than a thousand sources, reporting over 300 million substance records and more than 118 million unique compounds, but its model is that of an aggregator and archive—it organizes what others deposit—and dense, compound-level pharmacological characterization is not what it is designed to supply (Kim et al., 2025). Foundational archives such as the Protein Data Bank, the crystallographic databases, and the NIST Chemistry WebBook established the field's standards, coupling deposition with validation, stable identifiers, expert review, and the preservation of provenance and uncertainty (Berman et al., 2000; Groom et al., 2016; Linstrom & Mallard, 2001). Dense pharmacological characterization comes instead from manually curated knowledgebases—DrugBank, DrugCentral, ChEBI, BindingDB, KEGG, the Human Metabolome Database, the IUPHAR/BPS Guide to PHARMACOLOGY, and ChEMBL—each authoritative within a defined remit and maintained by curators working from formal standard operating procedures (Degtyarenko et al., 2008; Harding et al., 2024; Kanehisa & Goto, 2000; Knox et al., 2024; Liu et al., 2007; Ursu et al., 2017; Wishart et al., 2022; Zdrazil et al., 2024). A third strategy replaces curation with automation or first-principles computation—patent-scale structure mining, federated environmental toxicology, high-throughput density functional theory—buying scale at the cost of confinement to a single kind of property (Papadatos et al., 2016; Saal et al., 2013; Williams et al., 2017). Supplementary Text S1 reviews these resources in detail.

Viewed together, these resources map a trade-off rather than escape it. Quality and depth come from manual curation, at a cost that limits breadth and currency; breadth and currency come from automation or computation, at the cost of confinement to a single property kind. Each resource is optimized for a particular evidentiary object—an atomic structure, a docking-ready molecule, a target–ligand relation, a metabolite, an assay result—and integrating across these objects is difficult. Broad qualitative questions are moreover rarely encoded as searchable fields: whether a compound is plausibly neuroactive, commonly sedating, associated with oxidative-stress pathways, or likely to cross the blood–brain barrier may require synthesis across many sources rather than retrieval of one standardized measurement. The developers of these resources acknowledge the ceiling directly, identifying publication volume as a barrier to continued manual expansion (Harding et al., 2024; Wishart et al., 2022; Zdrazil et al., 2024). What is largely missing is a resource at once broad in the compounds it covers, broad in the kinds of properties it records, current, inexpensive, and internally consistent across all of it—a standardized attribute layer that could complement established databases even where it is not precise enough to replace source-specific records.

Large Language Models for Scientific Data Extraction

Large language models (LLMs) create a plausible mechanism for producing such a layer. Because they interpret natural language and can be prompted to follow user-defined schemas, they translate varied terminology into consistent labels without a task-specific labeled corpus, and when connected to web search they synthesize evidence dispersed across sources rather than relying solely on parametric memory. This is especially relevant to chemical information, where the same compound may appear under systematic names, salts, brands, abbreviations, or historical synonyms, and where a single question

may require connecting structural, mechanistic, and clinical descriptions. LLM outputs are, however, stochastic, sensitive to prompt and model version, and vulnerable to unsupported inference, so the methodological question is not whether an LLM can produce an answer but whether a complete pipeline can produce a structured data product whose errors, uncertainty, provenance, and appropriate uses are empirically characterized.

A rapidly growing body of work has established that LLMs can convert scientific prose into structured data with accuracy approaching, and workflows exceeding, what earlier rule-based systems achieved. In materials science and chemistry the dominant pattern is a staged, prompt-engineered extraction pipeline: systems mining alloys, metal–organic frameworks, photophysical properties, solid-state syntheses, and a century of reaction procedures report F1 or field-level accuracy at or above 0.93 by decomposing the task into relevance checking, schema-constrained retrieval, and validation, and by coupling flexible language reasoning to deterministic cheminformatics such as stoichiometry parsing and name-to-structure resolution (Grougan et al., 2026; Kamatchi Sundaram et al., 2026; Kang et al., 2025; Lee et al., 2025; Li et al., 2026; Roy et al., 2026). The same capability has been demonstrated across the life sciences—in pest-control synthesis, enzyme–substrate mining, herbarium and taxonomic transcription, microbiome metadata standardization, and single-cell type annotation—where accuracy at or near expert baselines is achievable at negligible cost when prompts specify ordered fields, null behavior, units, and output format (Austin et al., 2026; Gaio et al., 2026; Hou & Ji, 2024; Scheepens et al., 2024; Schmidt-Lebuhn & Knerr, 2025; Smith et al., 2025).

These studies also mapped the failure modes any large-scale application must confront. Residual errors concentrate in multi-step symbolic reasoning rather than recognition; fabricated entries scale with output length and model self-review catches them poorly; repeated runs disagree about whether a field should be populated at all; and longer chain-of-thought prompting and plurality voting do not reliably outperform a single concise pass (Crowley et al., 2025; Kamatchi Sundaram et al., 2026; Scheepens et al., 2024; Schmidt-Lebuhn & Knerr, 2025). The most effective responses have been architectural rather than prompt-level: abstention on uncertain cases, explicit grounding to supporting document regions, second-pass self-validation, and multi-agent or multi-model deliberation that converts disagreement into a reviewable quality score—one consensus procedure driving biologically implausible annotations from 56% to 4% over successive rounds (Cheng & Zhang, 2025; Li et al., 2025; Wang et al., 2025; Xie et al., 2026; Yang et al., 2026). Supplementary Text S2 reviews this literature in full.

From Extraction to Structured Annotation

For all their diversity, these systems share a defining premise: the LLM is an extraction engine whose task is to read a document and report faithfully what it states. Their errors are accordingly errors of fidelity—a misparsed formula, a fabricated row, a value attributed to the wrong compound—and their validation is validation against source text. That framing is exactly right when the goal is to harvest facts already measured and written down, but it leaves untouched a complementary use of the same models, one better suited to the coverage problem that defeats extraction when the evidence for a compound is thin or scattered. The alternative is to use the model as a structured annotator: to pose a fixed battery of questions and ask for a graded judgment on each, synthesizing across the training corpus and, with web search, the live literature, rather than copying an answer stated in any single source. On this framing an individual answer is not a retrieved fact but a computational estimate, and the resulting matrix is evaluated as any statistical data product is—not cell by cell against a source sentence, but in aggregate, through external agreement, calibration, and downstream coherence.

The line between extraction and annotation is not absolute; pest-control work already distinguished reading a taxonomy stated in the text from predicting one the abstract omitted, and cell-typing and image-annotation systems assign labels that synthesize evidence rather than transcribe it (Pahuja et al., 2026; Scheepens et al., 2024; Xie et al., 2026). What the annotator framing does is make that synthesis the

explicit object of the exercise, which reorganizes the methodological priorities: faithfulness to a single source is no longer the target, so the burden shifts onto calibration, external validation, and structural coherence. The framing also inherits the failure modes of generative models, above all a tendency toward affirmative hallucination, which must be managed through confidence stratification, independent adjudication, and treating high-confidence negatives very differently from low-confidence affirmatives (Grougan et al., 2026; Scheepens et al., 2024; Yang et al., 2026). LLMs should not, however, be treated as generic numerical or expert oracles: an evaluation of models as sources of Bayesian priors and zero-shot tabular imputations found many priors poorly calibrated and conventional methods generally superior, arguing for categorical tasks with explicit metadata, calibrated response options, external ground truth where available, and downstream analyses that test aggregate usefulness separately from cell-level accuracy (Selby et al., 2025).

Initial work suggests such estimates can carry scientifically useful multivariate structure. An early proof of concept generated trait profiles for land animals and questionnaire responses for individuals: the biological estimates correlated strongly with external measurements of body mass and metabolic rate, the psychometric matrix produced interpretable factors, and the recovered factor structure differed from the model's direct verbal prediction of that structure, suggesting that repeated structured judgments can expose regularities unavailable in a single declarative answer (Moore, 2025). The approach was then demonstrated at scale in the LifeDive project, in which a web-search-enabled Grok reasoning model assigned 196 ordinal phenotype attributes to 35,856 animal, plant, and fungal species, producing more than seven million estimates. That resource was validated not by checking each cell against a source but by correlating its aggregate structure against expert-curated trait resources, auditing a random sample with an architecturally distinct model, and confirming the biological coherence of the recovered dimensions; the audit placed overall directional accuracy at roughly 95 to 97%, found the model's confidence gradations informative, and framed the output as a versioned statistical data product (Moore, 2026). LifeDive is the closest methodological precedent for the present work.

The pharmaceutical and chemical domain is an unusually favorable setting in which to stress-test this approach. The phenotype of interest is exactly the kind of multi-scale target that resists any single method, ranging from deterministic structural descriptors, which can be computed and therefore checked exactly, through mechanism-of-action and molecular-target judgments, to broad clinical characterizations that depend on dose, species, route, formulation, and strength of evidence. The domain is anchored by an exceptionally well-characterized reference population—the few thousand marketed and heavily studied drugs curated by DrugBank, ChEMBL, and the IUPHAR/BPS Guide—against which a model's estimates can be checked, while the long tail of research compounds supplies exactly the sparsely documented cases in which an annotator is most needed (Harding et al., 2024; Knox et al., 2024; Zdrazil et al., 2024). Structural attributes further provide a built-in objective anchor: because a molecule's atoms, rings, and bonds follow deterministically from its structure, part of the battery can be graded against ground truth with no human judgment, calibrating confidence in the harder pharmacological items on the same compounds. Compound names introduce a countervailing difficulty—salts, combination brand names, and nonunique abbreviations—that makes chemistry a stringent test of confidence calibration and error asymmetry. Supplementary Text S3 develops this rationale.

The Present Study

I report PharmaDive, a compound-level property resource built by applying the structured-annotator methodology to the pharmaceutical and chemical domain. Candidate substances were assembled from four public pharmaceutical sources—the FDA Orange Book, RxNorm, structured product labeling accessed through openFDA, and the World Health Organization Model List of Essential Medicines—and supplemented by targeted literature searching, yielding a sample of more than 16,000 substances spanning approved drugs, investigational agents, endogenous and natural substances, discontinued

products, research chemicals, venoms, plant extracts, and cosmetic and fragrance ingredients. Most entries are discrete molecular species, but the collection deliberately extends to multicomponent preparations that have no single defining structure. Each was characterized across a battery of attribute questions covering pharmacological class, mechanism of action, molecular target, biological pathways, metabolic and physicochemical properties, therapeutic and clinical context, and structural features, using a web-search-enabled Grok reasoning model that rated every compound on each attribute along a five-point confidence scale, from definitely no to definitely yes, with an explicit neutral option for cases that could not be determined. In keeping with the annotator framing, the model was instructed to search the primary and academic literature before responding, requests were issued in small batches with strict tabular output, and automated quality-control logic aligned response columns to attribute identifiers, converted unparseable content to unknown, checkpointed successful output, and retried failures. After harmonization, the corpus comprised 130 unique attributes.

I evaluate the resource along the three axes the annotator framing demands. To assess accuracy, a random sample of compounds was independently reviewed by a second, architecturally distinct large language model along two tracks: an objective track, in which structural attributes were scored against ground truth computed from molecular structure, and an adjudication track, in which pharmacological and clinical attributes were graded against expert knowledge with confidence intervals obtained by compound-clustered resampling. Because a model-adjudicated audit is not fully independent of the object it audits, I additionally benchmarked the matrix against curated external references—deterministic cheminformatics ground truth and expert-annotated mechanisms, targets, and therapeutic classes—for the subset of compounds matching a structure-resolved library record. To assess structure, the harmonized matrix was subjected to item factor analysis for ordered-categorical data, and the recovered dimensions were examined for interpretability and for the face validity of the compound-level scores they induce. Together these analyses separate three questions often conflated in evaluations of LLM-generated data: whether individual ratings are correct, whether the model's stated confidence is informative, and whether the matrix supports scientifically interpretable aggregate patterns.

The contribution is both a dataset and a demonstration, and it is not a claim that LLM estimates should supersede curated chemistry resources, which remain authoritative within the domains defined by their data models and validation procedures. As a dataset, PharmaDive offers a broad, internally consistent, explicitly uncertainty-graded property layer that is inexpensive to produce and straightforward to extend, supporting hypothesis generation, comparison, dimensional summaries, candidate prioritization, and identification of records warranting expert review. As a demonstration, it carries the structured-annotator methodology—previously established across the tree of life (Moore, 2026)—into a domain that couples a partly deterministic, exhaustively documented core with a sparsely documented periphery, testing whether the approach's central promises of breadth at low cost, informative calibration, and coherent latent structure, together with its central hazard of confident affirmative error, transfer. Throughout, I treat the individual estimates as what they are—model-generated judgments of varying reliability, preserved with their confidence categories, provenance, and validation results—and locate the scientific value of the resource in the aggregate patterns they compose.

Method

Compound Sampling

Candidate compounds were assembled from four publicly available pharmaceutical data sources. Approved drug products and their active ingredients were obtained from the U.S. Food and Drug Administration (FDA) Orange Book product file (U.S. Food and Drug Administration, 2024). Generic and brand nomenclature was retrieved programmatically from RxNorm via the RxNav REST application programming interface (API; Nelson et al., 2011), querying 10 term types spanning ingredients, precise

ingredients, brand names, multiple-ingredient concepts, and semantic clinical and branded drug and dose-form concepts. Additional brand, generic, and substance names were drawn from structured product labeling through the openFDA drug-label endpoint (Kass-Hout et al., 2016), and a reference set of essential medicines was included from the World Health Organization Model List of Essential Medicines (World Health Organization, 2023).

Records were merged across sources, and compound names were normalized by removing dosage strengths, dosage forms, and routes of administration, with duplicate entries collapsed to a single canonical form. This yielded a candidate pool of approximately 20,000 compounds, of which approximately 15,500 constituted the analytic sample after characterization and harmonization (described below). A subsequent manual search of Google Scholar obtained names of research compounds, venoms, plant extracts, cosmetic and fragrance compounds, and other obscure substances possibly missing from the initial search; these were added, resulting in a final list of 16,116. The corpus is therefore not restricted to discrete molecular entities: venoms, botanical extracts, and other multicomponent preparations are represented alongside single compounds, and the full attribute battery was posed of every entry regardless of whether it has one defining structure.

Phenotypic Characterization via Large Language Model

Each compound was characterized across a battery of attribute questions, each expressed as a yes/no proposition (e.g., pharmacological class membership, mechanism of action, or physicochemical property) and rated on an ordinal confidence scale. Attribute ratings were generated using the xAI Grok reasoning model (grok-4-1-fast-reasoning; xAI, 2026) with web-search access, accessed through its asynchronous batch API. For every attribute, the model rated each compound on a five-point scale: 0 (definitely no), 1 (probably no), 2 (unknown or not applicable), 3 (probably yes), and 4 (definitely yes).

Compounds were grouped into small batches (three compounds per request) and submitted as structured prompts that (a) listed the target compounds, (b) enumerated the attribute questions, (c) specified the response scale, and (d) instructed the model to search the internet thoroughly—including the primary and academic literature for obscure or novel research compounds—before responding, to return output strictly as comma-separated values with one row per compound and one column per question, and to record 2 (unknown) rather than leaving any cell blank when a determination could not be made. Web search was enabled as a tool for every request. Requests were issued through a batch pipeline that generated and validated the request file, uploaded it, created the batch, polled for completion, and paginated through the returned results.

Model output was parsed with automated quality-control logic. Response columns were aligned to their intended attributes by parsing the header labels; batches whose columns could not be unambiguously aligned were discarded and recorded as missing rather than risk silent misalignment of ratings. Row counts were reconciled to the number of compounds in each batch, and residual unparseable cells were set to 2 (unknown). Batches that failed parsing, returned no usable text, or could not be aligned were logged and reissued in a retry batch, with previously successful results preserved through incremental checkpointing. The battery was administered in several successive question sets across multiple collection runs.

Accuracy Validation

To estimate the accuracy of the model-generated characterizations, a random sample of 200 compounds (drawn with a fixed random seed) was independently reviewed by a second, architecturally distinct large language model (Claude Opus 4.8, configured at maximum reasoning effort; Anthropic, 2026), which was provided the sampled compounds together with their generated ratings. Compounds that could not be unambiguously identified (e.g., homeopathic products, organism names, generic English words, and obsolete brand names) were excluded from adjudication. The reviewer evaluated the remaining ratings

along two tracks. In an objective track, structural attributes with determinate answers were scored against ground truth computed with RDKit (RDKit, n.d.) from SMILES representations, each validated against an independently recalled molecular formula to guard against transcription error. In an adjudication track, pharmacological and clinical attributes were graded against expert knowledge using a sample stratified by response valence—equal numbers of endorsed and non-endorsed ratings, given the strong negative skew of the response distribution—with per-response-option accuracy reweighted to the population response proportions and 95% confidence intervals obtained by compound-clustered bootstrap resampling. Because the reviewing model had access to the generated ratings while grading, the adjudication track was unblinded, whereas the objective track carried no such dependency.

Validation Against Curated Databases

The procedure above evaluates the characterizations against a second language model. To complement it with a direct comparison against independently curated data, and following the external-database validation used for organismal traits in the companion resource (Moore, 2026), I benchmarked the attribute matrix against established chemical and pharmacological references. Compound names were normalized—lowercased, stripped of parenthetical qualifiers, stereodescriptors, and salt, ester, and hydrate designations, and reduced to alphanumeric tokens—and matched by exact normalized name to the Drug Repurposing Hub (Corsello et al., 2017), a curated library of approved, investigational, and preclinical compounds with confirmed structures (SMILES) and expert-annotated mechanisms of action, molecular targets, disease areas, and clinical phases. Of the 16,116 compounds, 2,079 (12.9%) matched a Hub record with a resolvable structure. Coverage was highest for approved and clinically studied small molecules and lowest for brand-named combination products, biologics, cosmetic and excipient ingredients, venoms and botanical extracts, and homeopathic or organism-derived entries, which are largely absent from structure-based drug libraries and in many cases have no single structure to resolve; the matched subset therefore constitutes a well-characterized reference sample rather than a random one.

For structural attributes that are deterministic functions of molecular structure, ground truth was computed directly from each matched compound’s SMILES with RDKit (RDKit, n.d.), yielding an exact reference that requires no human or model judgment. Eighteen attributes were operationalized as substructure, atom-composition, ring-topology, or descriptor queries spanning specified atoms and functional groups, ring systems, stereochemistry, chain length, and rotatable-bond count; the graded lipophilicity attribute was compared against the Crippen calculated logP. Supplementary Text S4 gives the full operationalization of each attribute. Because these references are resolved deterministically from curated structures, this track is wholly independent of the model under evaluation.

For pharmacological attributes, external criteria were built from the Hub’s mechanism, target, disease-area, and clinical-phase annotations and from the World Health Organization Anatomical Therapeutic Chemical (ATC) classification (World Health Organization Collaborating Centre for Drug Statistics Methodology, 2024), matched by the same normalized name. Specific criteria mapped an attribute to a directly corresponding annotation, such as monoamine-oxidase or xanthine-oxidase activity to those enzymes in a compound’s target or mechanism annotation; broad criteria mapped diffuse attributes, such as effects on hormones or blood–brain-barrier penetration, to first-level ATC groups or disease areas, which provide only an approximate referent. Every criterion is listed in Supplementary Text S4. An attribute was treated as endorsed when its rating was definitely or probably yes and as not endorsed when definitely or probably no; agreement was summarized by directional accuracy, sensitivity, specificity, and the tetrachoric correlation between endorsement and the external indicator, with lipophilicity assessed by Spearman correlation. Because curated target lists and primary indications are not exhaustive, these criteria understate true prevalence, so the statistics are conservative and index recovered signal rather than adjudicating individual cells.

Data Compilation and Harmonization

Response matrices from all question sets and collection runs were merged by compound. Because the neutral category conveyed no directional information, all values of 2 (unknown or not applicable) were recoded as missing. When a compound had been rated more than once for a given attribute, its ratings were averaged and rounded to the nearest scale point; averaged values resolving to the neutral midpoint were likewise treated as missing. Attributes administered in more than one question set were identified by canonicalizing item labels (case-folding and removing non-alphanumeric characters) and collapsed into single columns. The compiled matrix comprised 160 attribute columns before this step and 130 unique attributes after it.

Dimensionality Analysis and Factor Scoring

The harmonized attribute matrix was analyzed with item factor analysis for ordered-categorical data as implemented in the *psych* package (Revelle, 2024) in R (R Core Team, 2025). Prior to analysis, the retained categories were recoded to a contiguous four-level ordinal scale (0 = definitely no, 1 = probably no, 2 = probably yes, 3 = definitely yes), the neutral category having already been set to missing. A factor analysis based on the matrix of polychoric item correlations was estimated using maximum-likelihood extraction with oblimin (oblique) rotation, specifying an 11-factor solution and interpreting the loadings within a two-parameter item response theory (IRT) framework (*irt.fa*). Latent-trait (theta) scores were then computed for all compounds using a two-parameter IRT scoring routine (*scoreIrt*), retaining item loadings at or above .30 and bounding scores to the interval [−6, 6].

The 11 resulting dimensions were labeled, on the basis of their highest-loading attributes, as Redox/Stress-Modulating, Anti-Neurodegenerative, Immunomodulatory, Neuroactive, Phytochemical, Nucleic-Acid–Targeting, Epigenetic, Cardiometabolic, Lipophilic, Synthetic Small-Molecule, and Basic Heterocyclic. To limit the influence of extreme values, factor scores were winsorized (0.2% trimmed per tail), and each dimension was then linearly rescaled to a common range of −4 to 4 and rounded to two decimal places for reporting and downstream use.

Software and Figures

All data retrieval, processing, and analysis were conducted in R (R Core Team, 2025), using httr (Wickham, 2023) and jsonlite (Ooms, 2014) for web-service requests and JSON handling, data.table (Barrett et al., 2024) for tabular ingestion, the tidyverse packages (Wickham et al., 2019) for data manipulation, psych (Revelle, 2024) for psychometric analysis, and scales (Wickham et al., 2023) for value rescaling. Figures were produced with ggplot2 (Wickham, 2016), ggrepel (Slowikowski, 2024), and patchwork (Pedersen, 2024); label placement used a fixed random seed for reproducibility. One portion of analysis (validation against existing databases) was done in Python. Complete details, including all code and data, are publicly available at https://osf.io/4f2b5/files/osfstorage.

Results

The harmonized corpus comprised more than 16,000 substances, each characterized on 130 attributes. Results are organized in three parts: the accuracy of the model-generated ratings, the dimensional structure recovered from the attribute matrix, and the face validity of the resulting compound-level factor scores.

Accuracy of the Compound Characterizations

The validation sample comprised 26,000 rating cells (200 compounds × 130 attributes). Overall, 16.5% of cells were missing, and missingness was concentrated in particular attributes; one attribute was missing for 107 of the 200 compounds. The response distribution was strongly skewed toward negative

endorsements: approximately 70% of non-missing ratings were negative (definitely no or probably no), and 51% were definitely no. Independent adjudication indicated that the true prevalence of affirmative attributes was approximately 23%. Forty-seven compounds could not be unambiguously identified and were excluded, leaving 153 for evaluation.

Agreement on objectively verifiable structural attributes was high. Across fully objective structural items, model ratings matched RDKit-computed ground truth for 95.7% of cells; excluding a single poorly performing item—whether a compound has more than five rotatable bonds, the only structural item requiring enumeration rather than recall—agreement rose to 99.0% (10 discrepancies among 1,038 cells). The rotatable-bond item itself was correct for only 61% of compounds, below the 79% expected from constant responding, and every one of its 32 errors was a false-positive endorsement.

On pharmacological and clinical attributes, weighted accuracy against expert adjudication was approximately 92%. This figure is best interpreted against the majority-class baseline: because the response distribution was so heavily negative, always responding definitely no would itself achieve 77% accuracy, so the ratings' net lift over that baseline was roughly 15 percentage points.

Accuracy differed sharply by response option, revealing a systematic positive-response bias (Table 2). Negative endorsements were highly accurate (definitely no, 99.2%; probably no, 100%), whereas affirmative endorsements were less reliable (definitely yes, 89.5%; probably yes, 56.6%, near chance). Sensitivity was 97.8% and specificity 89.4%; 29 of 30 adjudicated pharmacological errors, and all structural errors, were false-positive endorsements. Affirmative ratings were least reliable for niche mechanistic attributes (e.g., Nrf2, MAPK/ERK, inflammasome, ferroptosis, and histone-deacetylase involvement), for which an affirmative rating was correct approximately 48% of the time, and occasional internally inconsistent pairs were observed (e.g., one compound rated as both primarily hepatically metabolized and primarily renally cleared). Because the excluded compounds were disproportionately the most obscure entries, the approximately 92% figure should be regarded as an upper bound for the full corpus.

Agreement With Independent Databases

Direct comparison against curated external data reinforced the adjudication findings on a far larger sample of 2,079 compounds and localized the model's strengths and weaknesses more precisely (Table 3; Figure 2). Agreement with deterministic structural ground truth was high and, for most attributes, near ceiling: across the 18 structural attributes, median directional accuracy was 95.2% and the median tetrachoric correlation with RDKit-derived truth was .98. Attributes defined by atom composition or ring topology were recovered almost perfectly—phosphorus (99.7% accuracy, r = 1.00), quaternary ammonium (99.4%, 1.00), halogen (97.4%, 1.00), macrocycle (99.1%, .99), multiple stereocenters (95.6%, .99), and a steroid nucleus (97.2%, .98)—confirming that the model reliably associated compound names with their gross molecular structure. Graded lipophilicity correlated with the calculated logP at Spearman r = .59.

Two exceptions were informative, and both reproduced patterns from the smaller adjudication sample. The one attribute requiring enumeration rather than recognition—whether a compound has more than five rotatable bonds—again failed badly, with directional accuracy of only 59.0% (below the 61.5% obtainable by constant negative responding) and specificity of .34; the model endorsed it for 78.6% of compounds against a true prevalence of 38.5%, and essentially all errors were false-positive endorsements. The two structural attributes phrased so as to invite broad interpretation—containing an amine and containing an alcohol or hydroxyl group—showed the same directionality more mildly, recovered with high sensitivity but reduced specificity (57.0% and 72.5%). The recurrence of this asymmetry on exactly checkable structural attributes across roughly 2,000 compounds corroborates the affirmative-response bias seen in the model-adjudicated audit using an entirely independent and exact reference.

Stratifying these deterministic comparisons by the model's response option reproduced the calibration profile of the adjudication audit at far greater scale (Figure 2B). Across 36,026 structural cells with computed ground truth, ratings of definitely no were 98.6% accurate and probably no 95.9% accurate, whereas probably yes was correct only 21.8% of the time and definitely yes 79.0%; pooled, negative ratings were 98.5% accurate, affirmative ratings 73.1%, and 88% of all errors were false-positive endorsements. That an affirmative hedge is wrong more often than right even for exactly checkable structural properties provides strong, independent support for treating low-confidence affirmations as review flags rather than as evidence—precisely what the response-option analysis of the adjudication sample indicated.

Agreement with curated pharmacological annotations was also strong for attributes with specific referents, despite curated targets and indications being incomplete and thus biasing these comparisons downward. Across the 14 mechanistically specific attributes, median directional accuracy was 93.7% and the median tetrachoric correlation was .91. Attributes mapping to a single, well-annotated target or class agreed closely with the external indicator—renin–angiotensin action (r = 1.00), xanthine oxidase (.99), GLP-1 agonism (1.00), histone deacetylase (.98), antineoplastic use (.98), nucleoside-analog status (.92), antimicrobial action (.93), and monoamine-oxidase inhibition (.90)—indicating that the matrix recovers genuine pharmacology and not merely structural regularities. Broad clinical attributes mapped only to coarse ATC groups or disease areas agreed less well (median accuracy 76.5%, median tetrachoric .71), with the model again endorsing diffuse effects for more compounds than a primary-indication criterion records; these attributes are the least suitable for cell-level adjudication. Full item-level results appear in Table 3.

Dimensional Structure

Item factor analysis yielded an interpretable 11-factor solution with well-separated dimensions; together, the factors accounted for approximately 53% of the total item variance (Table 1). Peak loadings ranged from .65 to .94, and each factor was defined by a coherent set of attributes.

The first and largest factor, Redox/Stress-Modulating (17 salient loadings ≥ .40; peak λ = .79), was defined by attributes concerning reactive oxygen species, glutathione, the Nrf2/Keap1 antioxidant pathway, oxidative-stress markers, autophagy, mitochondrial function, and endoplasmic-reticulum stress. Anti-Neurodegenerative (λ = .88) loaded on tau aggregation and hyperphosphorylation, α-synuclein and amyloid-β aggregation, and neurotrophic signaling. Immunomodulatory (λ = .81) captured immune-cell trafficking, interleukin and cytokine modulation, eosinophil counts, and T-lymphocyte function. Neuroactive (λ = .94) was defined by effects on neurotransmitters, dependence liability, sedation, and GABAergic and glutamatergic signaling. Phytochemical (λ = .93) loaded on polyphenol and phenolic-antioxidant classification, saponin status, and xanthine-oxidase and urate-transport activity.

Nucleic-Acid–Targeting (λ = .77) captured effects on DNA, RNA, and protein synthesis, cellular internalization, nucleic-acid binding or intercalation, and nucleoside-analog status. Epigenetic (λ = .65) loaded on histone methylation and methyltransferases, Notch signaling, and DNA methylation. Cardiometabolic (λ = .84) was defined by the renin–angiotensin system, cardiac remodeling, hormone levels, and blood pressure and heart rate. Lipophilic (λ = .76) captured fat solubility, cytochrome-P450 and hepatic metabolism, ester linkages, and volume of distribution. Synthetic Small-Molecule (λ = .67) loaded on room-temperature stability, halogen content, salt form, oral administration, and crystallinity, and Basic Heterocyclic (λ = .70) was defined by amine groups, heterocyclic rings, rotatable-bond count, sulfur content, and inhibitory mechanism. The full pattern matrix is presented in Table 1.

Compound-Level Factor Scores

Rescaled factor scores for 60 widely recognized compounds are displayed in Figure 1 and showed strong face validity. On the Neuroactive dimension, recognized centrally acting agents—including morphine,

fentanyl, caffeine, nicotine, ethanol, diazepam (Valium), fluoxetine (Prozac), and lithium—occupied the high end of the distribution, whereas pharmacologically inert substances such as sodium chloride, sucrose, and titanium dioxide anchored the low end. On the Lipophilic dimension, steroids and fat-soluble compounds (e.g., hydrocortisone, testosterone, retinol, β-carotene, and ivermectin) scored highest, and small ionic or water-soluble species (e.g., sodium chloride, fluoride, iodine, and lithium) scored lowest.

The joint distribution of Nucleic-Acid–Targeting and Basic Heterocyclic scores (Figure 1, lower panel) similarly separated compounds by mechanism: antimicrobial agents with nucleic-acid–directed activity, such as ciprofloxacin and tetracycline, together with folic acid, scored high on both dimensions, whereas testosterone was a clear outlier, projecting high on Nucleic-Acid–Targeting but very low on Basic Heterocyclic, consistent with its steroidal, non-heterocyclic structure. Inert and structurally simple substances again clustered at the low end of both dimensions.

Discussion

PharmaDive applies a structured-annotator methodology to the pharmaceutical and chemical domain, testing whether a web-search-enabled large language model can generate a standardized, uncertainty-graded pharmacochemical phenotype across a broad compound collection—more than 16,000 substances characterized on 130 attributes. Three findings stand out. First, objectively verifiable structural ratings agreed with cheminformatics ground truth for 95.7% of cells, and for 99.0% once a single item requiring multi-step enumeration was set aside, whereas that one item fell well below even a constant-response baseline. Second, pharmacological and clinical ratings reached roughly 92% weighted accuracy against independent adjudication, a net lift of about 15 percentage points over the majority-class baseline, yet almost every error was a false-positive endorsement, and “probably yes” was far less reliable than either “definitely yes” or the negative categories. Third, item factor analysis recovered 11 interpretable dimensions that reproduce recognizable pharmacological and structural constellations and induce compound-level scores with strong face validity. Together these results support PharmaDive as a broad, inexpensive, internally coherent exploratory resource while defining clear limits on any individual cell: it is most trustworthy precisely where it is least surprising.

Accuracy, Calibration, and Error Asymmetry

The objective structural validation identifies an instructive boundary between tasks that general-purpose language models perform reliably and tasks better delegated to deterministic cheminformatics. Excluding the rotatable-bond item, agreement with RDKit-derived ground truth was 99.0%, indicating that the model reliably associated compound names with the correct broad molecular structure. By contrast, the item asking whether a compound has more than five rotatable bonds was correct for only 61% of compounds—below the 79% expected from constant negative responding—and every one of its 32 errors was a false-positive endorsement. Recognizing a familiar structural pattern is evidently far easier for the model than exact graph-based counting, a failure mode reported elsewhere in scientific extraction (Kamatchi Sundaram et al., 2026; Schmidt-Lebuhn & Knerr, 2025). The lesson generalizes: any attribute that is a deterministic function of the molecular graph is better computed than asked, and a mature version of PharmaDive should reserve the language model for evidence synthesis and qualitative classification while treating RDKit or an equivalent toolkit as authoritative for atom counts, ring topology, formal charge, and related descriptors.

The pharmacological and clinical results demand more qualified interpretation than the headline weighted accuracy of approximately 92% might suggest, and the most consequential result is not that figure but the structure of the errors beneath it. Across both validation tracks, essentially every error was a false-positive endorsement—29 of 30 adjudicated pharmacological errors and all structural errors—so the instrument is highly sensitive (97.8%) but less specific (89.4%): it rarely misses a genuine attribute but systematically

over-attributes absent ones. Base rates make this consequential. Independent adjudication placed the true prevalence of affirmative attributes at roughly 23%, so even specificity near 90% allows the large mass of true negatives to generate many false positives relative to the smaller pool of true positives; the positive predictive value of an affirmation is therefore bounded well below its face value and degrades further as an attribute becomes rarer. In practical terms, a negative rating—especially a confident one—is close to dispositive, whereas an affirmative rating is a hypothesis whose strength depends jointly on the model's stated confidence and on the base rate of the attribute.

The per-option accuracies bear this out. Both negative categories were almost perfectly accurate (definitely no, 99.2%; probably no, 100%), "definitely yes" was correct 89.5% of the time, "probably yes" only 56.6%, and affirmations of the rarest mechanistic attributes were correct only about 48% of the time, essentially chance. These are not scattered failures but a coherent consequence of combining a sensitive but imperfectly specific generator with low-prevalence targets, and they imply that the confidence scale should be read less as a smooth probability than as a threshold: high-confidence responses in either direction carry real information, low-confidence affirmatives carry little, and the ordinal options should not be collapsed into a simple positive-versus-negative indicator without calibration. In this dataset, "probably yes" is better understood as a review flag than as moderate affirmative evidence. The neutral category, recoded here as missing, gave the model a way to decline rather than guess, which appears to have kept some of the weakest affirmations out of the matrix entirely.

This pattern likely reflects both the information environment and the wording of the annotation task. Positive associations are easier to retrieve than evidence of absence: a compound may be mentioned once in an in vitro pathway study or a speculative review, whereas few sources explicitly establish that it does not affect a given pathway. Broad propositions such as "affects inflammation" also admit many kinds of evidence, from a primary molecular mechanism to a downstream biomarker change observed only at high concentration, so a web-enabled model can convert a weak or context-limited association into a general endorsement. Future uses should accordingly preserve response valence and confidence separately, estimate item-specific positive predictive values, and require direct source verification for low-base-rate mechanisms even when the model answers "definitely yes."

This directionality is the signature the generative-model literature would predict, and the broader calibration pattern resembles the earlier LifeDive study, in which high-confidence phenotype responses were substantially more accurate than lower-confidence ones—though the two audits are not directly comparable (Moore, 2026). It extends findings from other annotation tasks in which high accuracy is reported only after strict schema design, validation, and abstention, with coverage or reproducibility remaining central limitations (Grougan et al., 2026; Hou & Ji, 2024; Schmidt-Lebuhn & Knerr, 2025). PharmaDive adds a complementary result: when the model synthesizes a fixed compound profile rather than copying values from a source, confidence can be informative, but calibration depends strongly on the direction and epistemic type of the claim, not merely on confidence magnitude.

Latent Structure of the Compound Matrix

The factor analysis provides evidence that the generated ratings were not an unstructured collection of independent guesses. The 11-factor solution was coherent and well separated, accounting for approximately 53% of item variance with strong peak loadings. Several dimensions recovered familiar pharmacological domains—Neuroactive, Immunomodulatory, Nucleic-Acid–Targeting, Cardiometabolic, Anti-Neurodegenerative—while others integrated chemical form with biological disposition: Lipophilic combined fat solubility, hepatic and cytochrome-P450 metabolism, ester linkages, and volume of distribution, and Synthetic Small-Molecule combined room-temperature stability, halogenation, salt form, oral use, and crystallinity. The coexistence of mechanistic, physicochemical, structural, and clinical attributes within the same factors is not a defect but a reflection of the fact that compounds are developed

and used as integrated objects whose structure, formulation, distribution, targets, and therapeutic context covary.

The compound-level scores behave as domain knowledge would require. Recognized centrally acting agents anchored the high end of the Neuroactive dimension while pharmacologically inert substances anchored the low end; steroids and fat-soluble vitamins topped the Lipophilic dimension while small ionic species fell to the bottom; and the joint distribution of Nucleic-Acid–Targeting and Basic Heterocyclic scores separated nucleic-acid–directed antimicrobials from a steroidal outlier in exactly the way their structures predict. The factors can therefore serve as “soft coordinates” for comparing heterogeneous compounds even when no single conventional class label applies, supporting exploratory clustering, retrieval of analogs defined by broad phenotype rather than substructure alone, and identification of compounds whose profiles are unusual relative to their nominal class—echoing the organismal precedent to this work (Moore, 2026).

Coherence at this level should not, however, be mistaken for cell-level validity. The largest dimension, Redox/Stress-Modulating, is defined by 17 salient loadings drawn substantially from the very mechanistic attributes on which affirmative ratings were least reliable. That a stable, interpretable factor can be assembled from items whose individual affirmations are correct only about half the time is not a contradiction: factor structure is driven by covariation across many compounds, and unstructured error largely averages out. But structured error does not. These processes are genuinely biologically related, yet they are also frequently co-mentioned across broad experimental literatures, so the factor may capture a mixture of shared biology, publication practice, and model prior, and a systematic tendency to co-endorse a cluster of fashionable mechanistic attributes could inflate or even induce a dimension that looks principled. The factor labels are therefore best read as shorthand for loading patterns rather than as discovered mechanisms: the Anti-Neurodegenerative dimension may distinguish compounds studied in Alzheimer’s or Parkinson’s disease models without implying demonstrated disease modification. The Basic Heterocyclic factor also included the poorly validated rotatable-bond item, a concrete reason to repeat the analysis after replacing that rating with a computed descriptor.

More broadly, internal coherence is not independent proof of truth. A model can reproduce stable associations from the literature, but it can also reproduce stable misconceptions, semantic shortcuts, or correlated hallucinations; and because the factors and the face-validity examples were derived from the same ratings, plausible structure is evidence that the matrix contains organized signal, not that every contributing item is correct. The face validity observed here is additionally drawn from widely recognized compounds—precisely the cases the model characterizes best—whereas the sparsely documented periphery, where the annotator approach is most needed and least checkable, is exactly where such artifacts would be hardest to detect. The dimensional summaries are accordingly best used as a guide to where to look rather than as measurements.

External Validation Against Curated Resources

The database-grounded validation strengthens each of these conclusions with evidence that does not depend on a language model serving as adjudicator. Because structural ground truth was computed deterministically from curated structures, the near-ceiling agreement on atom-composition and ring-topology attributes (median tetrachoric r = .98 across 2,079 compounds) shows that the model’s competence at gross structural recognition is real rather than an artifact of a single reviewer, while the identical failure of the rotatable-bond item and the identical concentration of error in affirmative responses confirm that both patterns are properties of the annotation process itself rather than of any one evaluation method. Agreement with curated targets and ATC classes for attributes such as renin–angiotensin action, xanthine-oxidase and histone-deacetylase activity, and GLP-1 agonism extends the point from chemistry to pharmacology (Moore, 2026). The reassurance is nonetheless bounded: only 12.9% of compounds matched a structure-resolved reference record, and the matched subset was enriched for approved, well-

studied drugs—precisely those the model characterizes best—so these statistics describe the well-documented core rather than the periphery that most motivates the resource. The mechanistic criteria are themselves incomplete, so a model affirmation the external record does not confirm cannot be scored as an error; these comparisons therefore validate the recovery of major contrasts more than they adjudicate individual cells, and the deterministic structural track remains the most decisive component of the validation. Supplementary Text S5 situates PharmaDive relative to the curated resources it is designed to complement rather than replace.

Limitations

Several limitations qualify these conclusions, the most important concerning validation itself. Accuracy was estimated against a single, architecturally distinct but still general-purpose language model acting as an unblinded adjudicator rather than against human experts. Architectural separation reduces the risk of identical model-specific errors but does not eliminate shared training data, shared web sources, or common cultural priors, so an audit by another model can understate exactly the errors both models are prone to; and because the reviewer saw the ratings it was assessing, anchoring effects are possible. The objective track avoids this dependency but covered only a limited set of structural attributes. The external comparison reported above (Table 3) supplies part of the stronger validation required, while blinded review by multiple pharmacologists or medicinal chemists would strengthen it further.

The exclusion of 47 of the 200 sampled compounds from adjudication is itself informative. These records were disproportionately obscure, ambiguous, obsolete, or not readily identifiable—that is, the periphery the approach most needs to characterize—which makes the approximately 92% pharmacological estimate an upper bound for the full corpus. Compound identity is unusually difficult in a name-centered pipeline: salts and free bases may share pharmacology but differ in molecular mass and solid-state behavior; brand names may denote combinations or formulations; abbreviations may be nonunique; and historical names may map imperfectly to modern structures. A subset of structural and physicochemical items is therefore ill-posed without a fixed protonation and salt state, and is ill-posed in a stronger sense for the venoms, extracts, and other multicomponent preparations the corpus deliberately includes, for which no single structure exists to interrogate; because the structural track required a resolvable SMILES, such entries are absent from it and their structural ratings remain unaudited.

Missingness raises a related concern. Unknown and not-applicable responses were combined into a single category and then recoded as missing, although epistemic uncertainty and logical inapplicability are different states; missingness was concentrated in particular attributes—one was undetermined for more than half of the validation compounds—and is unlikely to be random across compounds. The factor model may consequently reflect the better-documented portion of the corpus more strongly than the long tail that motivated the project. Averaging repeated ratings and rounding to the nearest category yields a single analyzable matrix but can conceal disagreement between runs. Unlike the organismal precedent, I did not impute, preferring to preserve the distinction between an observed and an inferred category and heeding evidence that model-based imputations are often poorly calibrated (Selby et al., 2025).

The attribute battery further mixes propositions of very different scope. A deterministic question such as whether a molecule contains chlorine is fundamentally unlike whether it affects inflammation, and some broad items do not distinguish a primary mechanism from a downstream consequence, therapeutic-dose evidence from high-concentration experiments, human evidence from animal or cell models, or an approved use from exploratory research. These ambiguities can inflate affirmative responses and complicate factor labels.

Finally, reproducibility and provenance remain central limitations. The annotation model is proprietary, web-search results change, and model updates can alter responses even when prompts are held constant (Hou & Ji, 2024; Schmidt-Lebuhn & Knerr, 2025). General-purpose models may also recognize well-

known compounds from training, making performance on familiar entities an uncertain mixture of synthesis and memorization (Selby et al., 2025). The current design does not attach a source trail to each rating, so exact reruns may be impossible and an individual cell cannot be audited without repeating the search. The selected 11-factor solution and its labels add further analytic discretion, and face validity was assessed on the same matrix from which the factors were extracted; the recovered dimensions are thus a model of the data rather than a ground truth about chemical space. Supplementary Text S6 elaborates these limitations and the analytic choices involved.

Implications and Future Directions

The error profile documented here maps directly onto its own remedies, and the most defensible path forward is a hybrid architecture organized by epistemic type. Exact structural attributes should be computed from standardized structures; measured properties and curated target relations should be retrieved from authoritative databases together with their conditions and provenance; and LLM annotation should be reserved for the broad, qualitative, or cross-source judgments that are otherwise expensive to synthesize (Li et al., 2026; Roy et al., 2026). The guiding rule is to compute what can be computed, retrieve what can be retrieved, and estimate only what genuinely requires synthesis. Each estimated cell should then retain the compound identifier and form, the prompt, the model and version, the date, the response category, the missingness state, and any deterministic checks, so that the matrix functions as a versioned statistical object rather than a static fact table.

Calibration should likewise become item-specific and decision-oriented. Because the dominant failure is confident and correlated over-endorsement, “probably yes” ratings should enter a mandatory review queue rather than be treated as positive labels, and even “definitely yes” ratings on rare mechanisms should be verified before use in mechanistic, clinical, or regulatory claims; multi-model consensus, adversarial review, or targeted human adjudication may be especially valuable for affirmative and low-base-rate items (Xie et al., 2026; Yang et al., 2026). Several empirical extensions would clarify scientific utility: broadening the external validation begun here to attribute-specific benchmarks against additional curated resources; deliberately oversampling obscure compounds, ambiguous names, salts, and entities published after the model’s training period; using replicate annotations to quantify temporal and cross-model stability; and testing the latent dimensions as inexpensive features in downstream tasks such as retrieval, adverse-effect prediction, and repurposing. Supplementary Text S7 sets out these directions in detail.

Conclusion

The structured-annotator paradigm, established across organismal biology, transfers to pharmacology and chemistry—a domain chosen precisely because it couples a partly deterministic, exhaustively documented core with a sparsely documented periphery. In that setting PharmaDive delivers what the paradigm promises—breadth at low cost, informative if asymmetric calibration, and coherent latent structure—while exhibiting the hazard the paradigm predicts: a systematic bias toward confident affirmative error, concentrated in the rarest and least verifiable claims. Its principal weakness is thus not random noise but a specific and consequential error asymmetry. The resource should accordingly be understood neither as a replacement for expert-curated databases nor as a collection of measured facts, but as a versioned layer of computational estimates whose value is aggregate and comparative rather than cell-level and authoritative.

The empirical lessons that shaped this judgment are portable to LLM-generated scientific data products more generally: trust high-confidence negatives, treat affirmatives as leads whose weight falls with the model’s confidence and the attribute’s rarity, compute what can be computed rather than asking for it, and evaluate coherence and accuracy as distinct criteria rather than letting either vouch for the other. Held to those disciplines—and equipped in future versions with deterministic chemistry, source-linked retrieval,

item-specific validation, and targeted expert review—a broad, inexpensive, uncertainty-graded attribute layer of this kind can become a useful complement to the existing chemical information ecosystem, without displacing the curated resources that remain authoritative within their domains.

References

Anthropic. (2026). *Claude* (Opus 4.8) [Large language model]. https://claude.ai

Austin, M. W., Linan, A. G., Blomberg, M., Schmidt, H. H., & Teisher, J. K. (2026). Using large language models to automate herbarium specimen transcription: A case study at the Missouri Botanical Garden. *Plants, People, Planet, 8*(4), 1068–1075. https://doi.org/10.1002/ppp3.70128

Barrett, T., Dowle, M., & Srinivasan, A. (2024). *data.table: Extension of "data.frame"* [Computer software]. https://CRAN.R-project.org/package=data.table

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242.

Cheng, Y., & Zhang, H. (2025). L2D: A versatile literature-to-data set pipeline for complex, user-specific data extraction in environmental research. *ACS ES&T Water, 5*(8), 4897–4907. https://doi.org/10.1021/acsestwater.5c00551

Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., Johnston, S. E., Vrcic, A., Wong, B., Khan, M., Asiedu, J., Narayan, R., Mader, C. C., Subramanian, A., & Golub, T. R. (2017). The Drug Repurposing Hub: A next-generation drug library and information resource. *Nature Medicine, 23*(4), 405–408. https://doi.org/10.1038/nm.4306

Crowley, G., Tabula Sapiens Consortium, & Quake, S. R. (2025). Benchmarking cell type and gene set annotation by large language models with AnnDictionary. *Nature Communications, 16*, 9511. https://doi.org/10.1038/s41467-025-64511-x

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., & Ashburner, M. (2008). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research, 36*(Database issue), D344–D350. https://doi.org/10.1093/nar/gkm791

Gaio, D., Tackmann, J., Perez-Molphe-Montoya, E., Näpflin, N., Patsch, D., Malfertheiner, L., Peluso, M. E., & von Mering, C. (2026). Enhanced semantic classification of microbiome sample origins using large language models (LLMs). *GigaScience, 15*, giag015. https://doi.org/10.1093/gigascience/giag015

Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials, 72*, 171–179. https://doi.org/10.1107/S2052520616003954

Grougan, M., Subasinghe, J., Hix, M. A., & Walker, A. R. (2026). Librarian of Alexandria: A modular chemical data extraction pipeline to compare LLM performance. *Journal of Chemical Information and Modeling, 66*, 6921–6932. https://doi.org/10.1021/acs.jcim.6c00374

Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P. H., Davenport, A. P., Spedding, M., & Davies, J. A. (2024). The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Research, 52*(D1), D1438–D1449. https://doi.org/10.1093/nar/gkad944

Hou, W., & Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods, 21*(8), 1462–1465. https://doi.org/10.1038/s41592-024-02235-4

Kamatchi Sundaram, A., Chakraborty, M., Devathi, S. M. K., Prusty, B. P., & Batra, R. (2026). Automated extraction of multicomponent alloy data using large language models for sustainable design. *Advanced Science*, Article e75916. https://doi.org/10.1002/advs.75916

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 28*(1), 27–30.

Kang, Y., Lee, W., Bae, T., Han, S., Jang, H., & Kim, J. (2025). Harnessing large language models to collect and analyze metal–organic framework property data set. *Journal of the American Chemical Society, 147*(6), 3943–3958. https://doi.org/10.1021/jacs.4c11085

Kass-Hout, T. A., Xu, Z., Mohebbi, M., Nelsen, H., Baker, A., Levine, J., Johanson, E., & Bright, R. A. (2016). OpenFDA: An innovative platform providing access to a wealth of FDA's publicly available data. *Journal of the American Medical Informatics Association, 23*(3), 596–600. https://doi.org/10.1093/jamia/ocv153

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2025). PubChem 2025 update. *Nucleic Acids Research, 53*(D1), D1516–D1525. https://doi.org/10.1093/nar/gkae1059

Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N. E., Strawbridge, S. A., Garcia-Patino, M., Kruger, R., Sivakumaran, A., Sanford, S., Doshi, R., Khetarpal, N., Fatokun, O., Doucet, D., Zubkowski, A., …

Wishart, D. S. (2024). DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Research, 52*(D1), D1265–D1275. https://doi.org/10.1093/nar/gkad976

Lee, S., Cruse, K., Baibakova, V., Ceder, G., & Jain, A. (2025). Text-mined dataset of solid-state syntheses with impurity phases using large language model. *Scientific Data, 12*, Article 1969. https://doi.org/10.1038/s41597-025-06222-y

Li, J., Li, M., Yang, Q., & Luo, S. (2026). ReactionSeek: LLM-powered literature data mining and knowledge discovery in organic synthesis. *Nature Communications, 17*, 3356. https://doi.org/10.1038/s41467-026-70180-1

Li, X., Huang, Z., Quan, S., Peng, C., & Ma, X. (2025). SLM-MATRIX: A multi-agent trajectory reasoning and verification framework for enhancing language models in materials data extraction. *npj Computational Materials, 11*, Article 241. https://doi.org/10.1038/s41524-025-01719-x

Linstrom, P. J., & Mallard, W. G. (2001). The NIST Chemistry WebBook: A chemical data resource on the Internet. *Journal of Chemical & Engineering Data, 46*(5), 1059–1063. https://doi.org/10.1021/je000236i

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research, 35*(Database issue), D198–D201. https://doi.org/10.1093/nar/gkl999

Moore, T. M. (2025). *Using large language models to obtain quantitative data* [Preprint]. PsyArXiv.

Moore, T. M. (2026). *A language model generates seven million phenotype estimates across the tree of life* [Preprint]. LifeDive.org.

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., & Moore, R. (2011). Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association, 18*(4), 441–448. https://doi.org/10.1136/amiajnl-2011-000116

Ooms, J. (2014). *The jsonlite package: A practical and consistent mapping between JSON data and R objects*. arXiv. https://doi.org/10.48550/arXiv.1403.2805

Pahuja, V., Stevens, S., East, A., Record, S., & Su, Y. (2026). Automatic image-level morphological trait annotation for organismal images. In *Proceedings of the International Conference on Learning Representations (ICLR 2026)*. https://arxiv.org/abs/2604.01619

Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S. A., Pettersson, J., Goncharoff, N., Hersey, A., & Overington, J. P. (2016). SureChEMBL: A large-scale, chemically annotated patent document database. *Nucleic Acids Research, 44*(D1), D1220–D1228. https://doi.org/10.1093/nar/gkv1253

Pedersen, T. L. (2024). *patchwork: The composer of plots* [Computer software]. https://CRAN.R-project.org/package=patchwork

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

RDKit. (n.d.). *RDKit: Open-source cheminformatics* [Computer software]. https://www.rdkit.org/

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* [Computer software]. Northwestern University. https://CRAN.R-project.org/package=psych

Roy, A., Grisan, E., Buckeridge, J., & Gattinoni, C. (2026). ComProScanner: A multi-agent based framework for composition-property structured data extraction from scientific literature. *Digital Discovery, 5*, 1794–1808. https://doi.org/10.1039/d5dd00521c

Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM, 65*(11), 1501–1509. https://doi.org/10.1007/s11837-013-0755-4

Scheepens, D., Millard, J., Farrell, M., & Newbold, T. (2024). Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods in Ecology and Evolution, 15*, 1261–1273. https://doi.org/10.1111/2041-210X.14341

Schmidt-Lebuhn, A. N., & Knerr, N. (2025). Large language models can extract morphological data from taxonomic descriptions, but their stochastic nature makes automation challenging: A test on Australian Asteraceae. *PhytoKeys, 261*, 189–210. https://doi.org/10.3897/phytokeys.261.158396

Selby, D., Iwashita, Y., Spriestersbach, K., Saad, M., Bappert, D., Warrier, A., Mukherjee, S., Kise, K., & Vollmer, S. (2025). Had enough of experts? Quantitative knowledge retrieval from large language models. *Stat, 14*(2), e70054. https://doi.org/10.1002/sta4.70054

Slowikowski, K. (2024). *ggrepel: Automatically position non-overlapping text labels with "ggplot2"* [Computer software]. https://CRAN.R-project.org/package=ggrepel

Smith, N., Yuan, X., Melissinos, C., & Moghe, G. (2025). FuncFetch: An LLM-assisted workflow enables mining thousands of enzyme-substrate interactions from published manuscripts. *Bioinformatics, 41*(1), btae756. https://doi.org/10.1093/bioinformatics/btae756

Ursu, O., Holmes, J., Knockel, J., Bologa, C. G., Yang, J. J., Mathias, S. L., Nelson, S. J., & Oprea, T. I. (2017). DrugCentral: Online drug compendium. *Nucleic Acids Research, 45*(D1), D932–D939. https://doi.org/10.1093/nar/gkw993

U.S. Food and Drug Administration. (2024). *Orange Book: Approved drug products with therapeutic equivalence evaluations* [Data set]. https://www.fda.gov/drugs/drug-approvals-and-databases/approved-drug-products-therapeutic-equivalence-evaluations-orange-book

Wang, X., Huey, S. L., Sheng, R., Mehta, S., & Wang, F. (2025). SciDaSynth: Interactive structured data extraction from scientific literature with large language model. *Campbell Systematic Reviews, 21*, e70073. https://doi.org/10.1002/cl2.70073

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H. (2023). *httr: Tools for working with URLs and HTTP* [Computer software]. https://CRAN.R-project.org/package=httr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *scales: Scale functions for visualization* [Computer software]. https://CRAN.R-project.org/package=scales

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., & Richard, A. M. (2017). The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *Journal of Cheminformatics, 9*, 61. https://doi.org/10.1186/s13321-017-0247-6

Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B. L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V. W., Varshavi, D., Varshavi, D., … Gautam, V. (2022). HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research, 50*(D1), D622–D631. https://doi.org/10.1093/nar/gkab1062

World Health Organization. (2023). *WHO model list of essential medicines—23rd list, 2023*. https://www.who.int/publications/i/item/WHO-MHP-HPS-EML-2023.02

World Health Organization Collaborating Centre for Drug Statistics Methodology. (2024). *Guidelines for ATC classification and DDD assignment 2024*. WHO Collaborating Centre for Drug Statistics Methodology. https://www.whocc.no/atc_ddd_index/

xAI. (2026). *Grok* (grok-4-1-fast-reasoning) [Large language model]. https://x.ai/

Xie, E., Cheng, L., Shireman, J., Cai, Y., Liu, J., Mohanty, C., Dey, M., & Kendziorski, C. (2026). CASSIA: A multi-agent large language model for automated and interpretable cell annotation. *Nature Communications, 17*, 389. https://doi.org/10.1038/s41467-025-67084-x

Yang, C., Zhang, X., & Chen, J. (2026). Large language model consensus substantially improves the cell type annotation accuracy for scRNA-seq data. *Communications Biology*. Advance online publication. https://doi.org/10.1038/s42003-026-10420-8

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Mendez Lopez, D., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research, 52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004

Table 1. Factor Pattern Matrix for the 130-Attribute Compound Characterization (11-Factor Oblimin Solution) N = 130 attributes. Entries are pattern loadings from an item factor analysis (maximum-likelihood extraction, oblimin rotation) of polychoric attribute correlations for the full compound corpus. Column headings are abbreviated factor labels (in order: Redox/Stress-Modulating, Anti-Neurodegenerative, Immunomodulatory, Neuroactive, Phytochemical, Nucleic-Acid–Targeting, Epigenetic, Cardiometabolic, Lipophilic, Synthetic Small-Molecule, Basic Heterocyclic). The leading item stem ("Does this compound/substance …") has been omitted for brevity. Loadings with absolute value below .30 are omitted; salient loadings (≥ .40) are shown in boldface. Proportions of variance are approximate under oblique rotation because the factors are intercorrelated.

Item	Redox/ Stress	Anti-Neurodeg.	Immuno-modulatory	Neuro-active	Phyto-chemical	Nucleic-Acid	Epi-genetic	Cardio-metabolic	Lipo-philic	Synthetic SM	Basic Heteroc.
Affect reactive oxygen species (ROS) accumulation	0.79										
Affect glutathione (GSH) levels	0.78										
Affect the Nrf2 (nuclear factor erythroid 2-related factor 2) pathway	0.76										
Modulate oxidative stress markers (e.g., SOD, MDA)	0.76										
Modulate autophagy	0.76										
Modulate mitochondrial function	0.73										
Enhance autophagy or protein clearance pathways	0.69										
Modulate the NRF2/Keap1 antioxidant pathway	0.68										
Affect endoplasmic reticulum stress pathways	0.67										
Modulate the MAPK/ERK signaling pathway	0.63										
Modulate the PI3K/AKT signaling pathway	0.62										
Affect proteostasis or protein degradation pathways	0.61										
Affect NADPH oxidase (NOX) enzymes	0.49		0.30								
Inhibit ferroptosis	0.45	0.38									
Affect the TGF-β/SMAD signaling pathway	0.42							0.33			
Affect tau protein aggregation		0.88									
Affect tau hyperphosphorylation		0.87									
Affect alpha-synuclein aggregation		0.86									
Affect amyloid-beta aggregation or clearance		0.83									
Affect neuroplasticity or neurotrophic factors like BDNF		0.59		0.41							
Inhibit microglial activation		0.54	0.44								
Have demonstrated insulin-sensitizing effects		0.46			0.30			0.44			
Affect acetylcholinesterase		0.45			0.43						
Act as a protein folding corrector or pharmacological chaperone		0.38			0.33		0.35				
Modulate the AGE-RAGE pathway		0.33			0.31						
Specifically target immune cell trafficking or migration			0.81								
Target interleukins			0.80								
Modulate eosinophil counts			0.77								
Neutralize cytokines			0.73								
Modulate T lymphocyte function or numbers			0.68								
Modulate TNF (tumor necrosis factor)	0.42		0.67								

Item	Redox/ Stress	Anti-Neurodeg.	Immuno-modulatory	Neuro-active	Phyto-chemical	Nucleic-Acid	Epi-genetic	Cardio-metabolic	Lipo-philic	Synthetic SM	Basic Heteroc.
A complement cascade inhibitor	−0.34		0.66								
Affect inflammation			0.62								
Modulate inflammasome activity	0.46		0.61								
Affect the immune system			0.47			0.35				−0.39	
Used in pediatric populations			0.32								
Affect neurotransmitters				0.94							
Carry addiction/dependence risk				0.80							
Commonly cause drowsiness/sedation				0.73							
Modulate GABAergic pathways		0.32		0.69							
Modulate glutamate signaling or metabolism		0.46		0.69							
Cross the blood-brain barrier				0.68							
Affect ion channels				0.64							
Modulate α2-adrenergic receptors				0.60							
Affect adenosine receptors				0.59	0.33						
Interact with 5-HT1A receptors				0.59							
Affect pain signaling			0.41	0.58							
Affect nicotinic acetylcholine receptors	−0.30	0.35		0.55							
A polyphenolic compound					0.93						
A phenolic antioxidant					0.89						
Classified as a saponin					0.68						
Affect xanthine oxidase					0.58						
Modulate urate transporters					0.57		0.31				
Have monoamine oxidase (MAO) inhibitory activity				0.32	0.53						
Contain terpenoid structures					0.45				0.34		−0.30
Available over-the-counter					0.39						
Affect DNA/RNA/protein synthesis						0.77					
Require internalization into cells to exert its therapeutic effect						0.70					
Intercalate with or directly bind to nucleic acids (DNA/RNA)						0.65					
Affect enzymes						0.58					
A nucleoside or nucleoside analog						0.56					
Contain fused ring systems						0.50			0.36		
Contain a ketone or aldehyde group						0.46			0.34		
Exist as multiple tautomers						0.43					
Have a long aliphatic chain						−0.42			0.41		
Used in cancer treatment						0.41	0.39				
A prodrug (activated after administration)						0.39					
Planar (flat molecule)					0.32	0.36				0.30	
Contraindicated in pregnancy						0.35					

Item	Redox/ Stress	Anti-Neurodeg.	Immuno-modulatory	Neuro-active	Phyto-chemical	Nucleic-Acid	Epi-genetic	Cardio-metabolic	Lipo-philic	Synthetic SM	Basic Heteroc.
Commonly cause GI upset						0.34					
Sensitive to light						0.26					
Affect histone methylation							0.65				
Affect histone methyltransferase enzymes							0.63				
Affect the Notch signaling pathway							0.51				
Modulate epigenetic mechanisms							0.48				
Affect DNA methylation							0.47				−0.33
Affect acetylation/deacetylation of proteins		0.30	0.35				0.44				
Affect the ubiquitin-proteasome pathway	0.32	0.35					0.44				
Affect histone deacetylase (HDAC)			0.39				0.43				
Modulate the Wnt/β-catenin signaling pathway	0.33						0.36				
Affect the renin-angiotensin system								0.84			
Affect cardiac remodeling								0.65			
Affect hormone levels								0.60			
Affect blood pressure/heart rate				0.55				0.58			
Used for chronic conditions						0.30		0.53			
Reduce triglyceride levels		0.36			0.32			0.49			
Compound/drug's mechanism involve disrupting cell membrane integrity								−0.44			
Have significant drug-drug interactions				0.30		0.34		0.44			0.30
Kill or stop growth of organisms (antibiotic/antiviral/antifungal)				−0.31		0.40		−0.44			
Used prophylactically (preventively)								0.37			
A macrocycle (large ring structure)								−0.26			
Compound/drug's therapeutic index primarily limited by on-target toxicity rather than off-target effects								0.21			
Lipophilic (fat-soluble)									0.76		
Contain a metal atom									−0.64		
Charged at physiological pH									−0.61		0.40
Primarily metabolized by cytochrome P450 enzymes									0.59		
Metabolized primarily by the liver				0.31					0.52		
Primarily cleared through renal (kidney) excretion									−0.51		
Contain an ester linkage									0.43		
Have a volume of distribution greater than 5 L/kg									0.42		
Bind strongly to plasma proteins									0.38		
Typically used in emergency situations									−0.33		
Contain phosphorus atoms									−0.28		
A biologic (protein/antibody)										−0.79	
Stable at room temperature										0.67	
Have multiple chiral centers										−0.60	

Item	Redox/ Stress	Anti-Neurodeg.	Immuno-modulatory	Neuro-active	Phyto-chemical	Nucleic-Acid	Epi-genetic	Cardio-metabolic	Lipo-philic	Synthetic SM	Basic Heteroc.
Derived from natural products					0.36					−0.52	
Have halogen atoms (F, Cl, Br, I)										0.49	
Contain an alcohol/hydroxyl group										−0.48	
A salt form									−0.44	0.47	
A GLP-1 receptor agonist								0.38		−0.44	
Exist as a zwitterion at physiological pH									−0.31	−0.42	0.39
Usually taken orally								0.35		0.38	
Contain a carboxylic acid group										−0.38	
Affect protein-protein interactions										−0.37	
Form crystals easily										0.31	
Have a long half-life (>24 hours)										−0.29	
A weak acid										0.27	
Have quaternary ammonium groups										0.25	
Exhibit non-linear pharmacokinetics at therapeutic doses										−0.22	
Require specific formulation technology for adequate bioavailability										−0.21	
Contain an amine group				0.30							0.70
Contain a heterocyclic ring (non-carbon atoms in ring)						0.31					0.64
A steroid structure			0.33					0.30	0.35		−0.56
Have rotatable bonds (>5)									0.41		0.46
Contain sulfur atoms											0.44
Work by blocking/inhibiting something						0.36					0.41
Contain a Schiff base (imine/C=N) functional group											0.35
Can this compound/substance be applied topically if in the right medium/formulation											−0.33
Contain a linker molecule connecting two distinct functional components											0.31
Sum of squared loadings	**9.12**	**6.81**	**7.14**	**8.06**	**6.00**	**6.30**	**4.62**	**5.61**	**5.45**	**5.42**	**4.49**
Proportion of variance	**0.07**	**0.05**	**0.06**	**0.06**	**0.05**	**0.05**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**

Table 2. Accuracy of Model-Generated Compound Ratings Against Independent Adjudication Estimates derive from an independent review of a random sample of 200 compounds (26,000 rating cells) by a separate large language model; 47 compounds that could not be unambiguously identified were excluded. Structural items were scored against ground truth computed with RDKit; pharmacological items were hand-graded against expert knowledge, reweighted to the population response distribution, with 95% confidence intervals from compound-clustered bootstrap resampling (single, unblinded grader). Across both tracks, nearly all errors were false-positive endorsements, indicating a systematic positive-response bias. Because the excluded compounds were disproportionately obscure, the weighted pharmacology estimate is an upper bound for the full corpus.

Evaluation track and metric	Estimate
Objective structural items (RDKit ground truth)	
Excluding rotatable-bond item	99.0%
All structural items	95.7%
Rotatable-bond item (requires enumeration)	61%
Pharmacological and clinical items (expert adjudication)	
Weighted accuracy, reweighted to population	~92%
Majority-class baseline (always “definitely no”)	77%
Sensitivity (true-positive rate)	97.8%
Specificity (true-negative rate)	89.4%
Accuracy by response option (adjudication track)	
0 = Definitely no	99.2%
1 = Probably no	100%
3 = Probably yes	56.6%
4 = Definitely yes	89.5%

Table 3. External Validation of Compound Attributes Against Independent Reference Data Model attribute ratings (endorsed = a rating of definitely or probably yes; not endorsed = definitely or probably no) were compared, across 2,079 compounds matched by normalized name, against deterministic structural ground truth computed with RDKit and against curated pharmacological criteria from the Drug Repurposing Hub (Corsello et al., 2017) and the World Health Organization Anatomical Therapeutic Chemical (ATC) classification. n = compounds with a non-missing rating and a defined criterion; Endorsed % = percentage the model rated definitely or probably yes; Ref. % = criterion prevalence in the reference data; Acc. = directional accuracy; Sens. and Spec. = sensitivity and specificity of model endorsement against the criterion; *r* = tetrachoric correlation between endorsement and the criterion. Structural criteria are exact; mechanistic and clinical criteria derived from curated target lists and primary indications are incomplete and therefore understate true prevalence, so their agreement statistics are conservative. Attributes are ordered by accuracy within each section. [a] Lipophilicity, a graded attribute, was compared against the Crippen calculated logP by Spearman correlation and is reported without binary statistics.

Attribute	n	Endorsed %	Ref. %	Acc.	Sens.	Spec.	r
Structural attributes (deterministic ground truth: RDKit)							
Contains phosphorus atom(s)	2056	1.9	1.6	99.7	100.0	99.7	1.00
Has quaternary ammonium group	2067	3.1	2.5	99.4	100.0	99.4	1.00
Is a macrocycle (large ring)	2069	3.8	4.1	99.1	85.9	99.7	0.99
Has halogen atom(s)	1849	25.6	24.8	97.4	96.3	97.8	1.00
Is a steroid structure	2069	5.8	7.7	97.2	69.8	99.5	0.98
Contains a metal atom	2068	4.6	1.4	96.5	89.7	96.6	0.91
Contains sulfur atom(s)	2067	21.4	19.5	96.1	95.0	96.4	0.99
Has multiple chiral centers	1769	38.7	38.3	95.6	94.8	96.2	0.99
Contains fused ring systems	1711	51.4	51.2	95.3	95.7	95.0	0.99
Contains an ester linkage	1817	17.8	14.0	95.2	96.1	95.0	0.98
Is a polyphenolic compound	2030	6.1	5.8	94.5	55.1	97.0	0.82
Contains a heterocyclic ring	2065	71.4	68.2	93.4	97.4	84.6	0.98
Contains a carboxylic acid group	2065	24.6	18.7	92.4	95.3	91.8	0.97
Has a long aliphatic chain	2066	11.6	2.8	90.3	84.2	90.4	0.80
Contains ketone or aldehyde	2067	27.4	13.4	83.9	92.1	82.7	0.88
Contains an alcohol/hydroxyl group	2064	56.3	42.3	82.3	95.6	72.5	0.91
Contains an amine group	2069	69.9	48.4	77.1	98.6	57.0	0.91
Has >5 rotatable bonds	2058	78.6	38.5	59.0	98.7	34.1	0.79
Is lipophilic[a]	2058	—	—	—	—	—	.59
Mechanistic / therapeutic attributes – specific criteria (Drug Repurposing Hub, WHO ATC)							
GLP-1 receptor agonist	2048	0.1	0.1	100.0	66.7	100.0	1.00
Nucleoside/nucleoside analog	2041	2.6	1.9	98.3	73.7	98.8	0.92
MAO inhibitory activity	1977	3.8	1.0	97.0	89.5	97.0	0.90
Affects renin-angiotensin system	1887	4.6	1.4	96.8	100.0	96.8	1.00
Affects acetylcholinesterase	1957	4.0	1.4	96.1	51.9	96.7	0.70
Affects xanthine oxidase	1926	4.3	0.3	96.0	100.0	96.0	0.99
Modulates alpha-2 adrenergic receptors	1973	7.7	2.8	93.9	78.2	94.3	0.83

Attribute	n	Endorsed %	Ref. %	Acc.	Sens.	Spec.	r
Interacts with 5-HT1A receptors	1949	8.1	3.0	93.5	77.6	94.0	0.83
Affects nicotinic ACh receptors	1932	7.2	1.6	93.5	73.3	93.8	0.75
Affects adenosine receptors	1890	7.5	1.0	92.8	63.2	93.1	0.64
Used in cancer treatment	2075	16.2	8.9	92.5	98.9	91.9	0.98
Modulates GABAergic pathways	1943	11.2	1.7	90.1	85.3	90.2	0.78
Affects histone deacetylase (HDAC)	1810	12.5	0.7	88.1	100.0	88.0	0.98
Antibiotic/antiviral/antifungal	2074	25.8	14.1	86.9	95.2	85.6	0.93
Clinical / therapeutic attributes – broad criteria (Drug Repurposing Hub, WHO ATC)							
Affects neurotransmitters	1765	33.5	28.8	87.3	86.2	87.7	0.91
Affects pain signaling	2056	18.3	4.7	85.6	90.7	85.3	0.83
Affects ion channels	2048	27.4	12.1	79.3	77.8	79.6	0.71
Affects hormone levels	2045	30.9	11.5	76.5	82.2	75.8	0.71
Affects the immune system	2059	40.5	4.7	62.6	82.5	61.6	0.51
Crosses blood-brain barrier	2003	51.1	17.4	60.7	83.9	55.8	0.56
Affects blood pressure/heart rate	2042	54.3	14.1	57.5	92.0	51.9	0.65

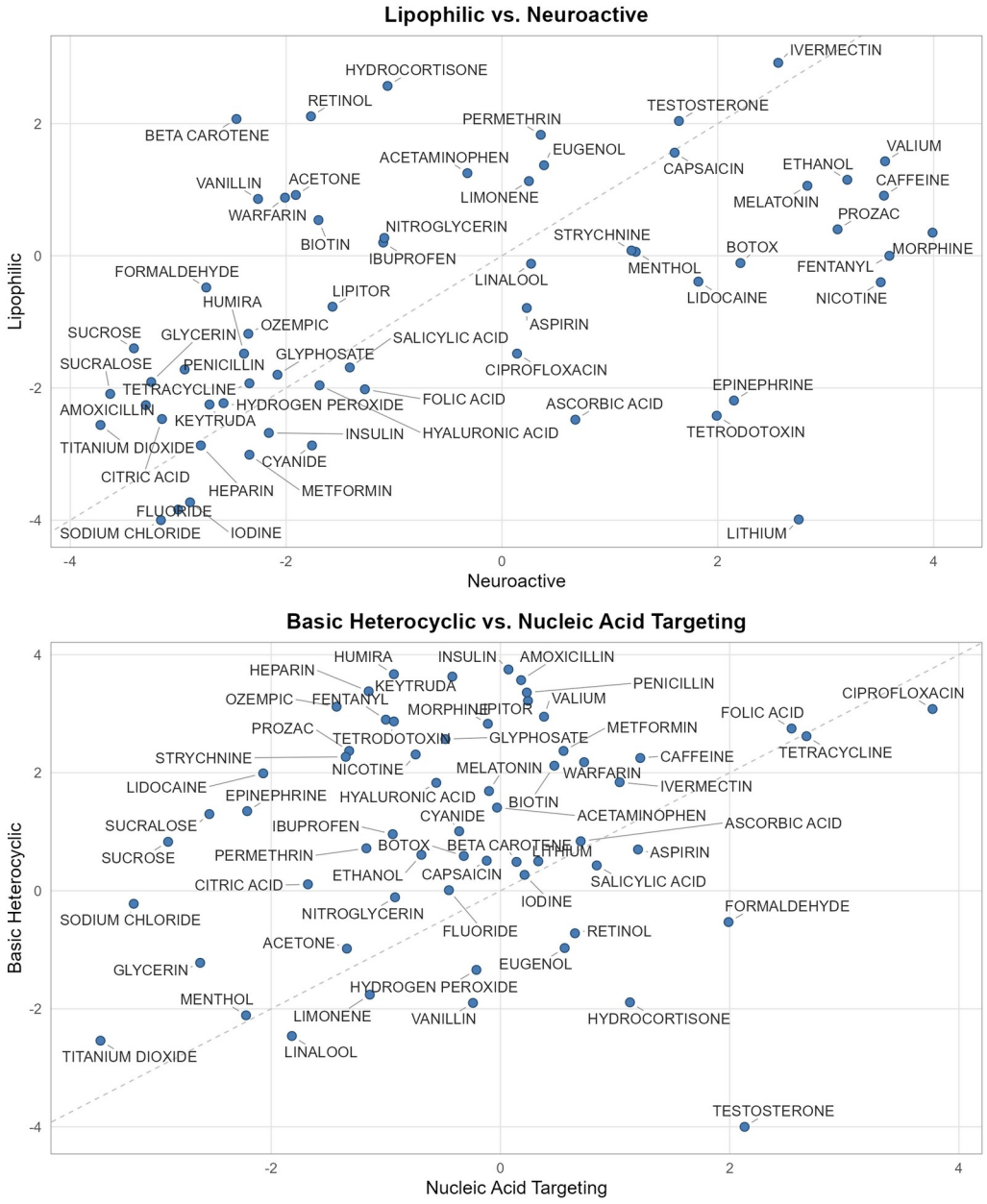

Figure 1. Rescaled Factor Scores for 60 Common Compounds on Four Selected Dimensions Each point is a compound; positions are winsorized factor scores rescaled to a common −4 to 4 range. The dashed line marks equality (y = x). Upper panel: Lipophilic (y-axis) versus Neuroactive (x-axis). Lower panel: Basic Heterocyclic (y-axis) versus Nucleic-Acid–Targeting (x-axis). Higher scores indicate a greater estimated likelihood that a compound expresses the attributes defining each dimension.

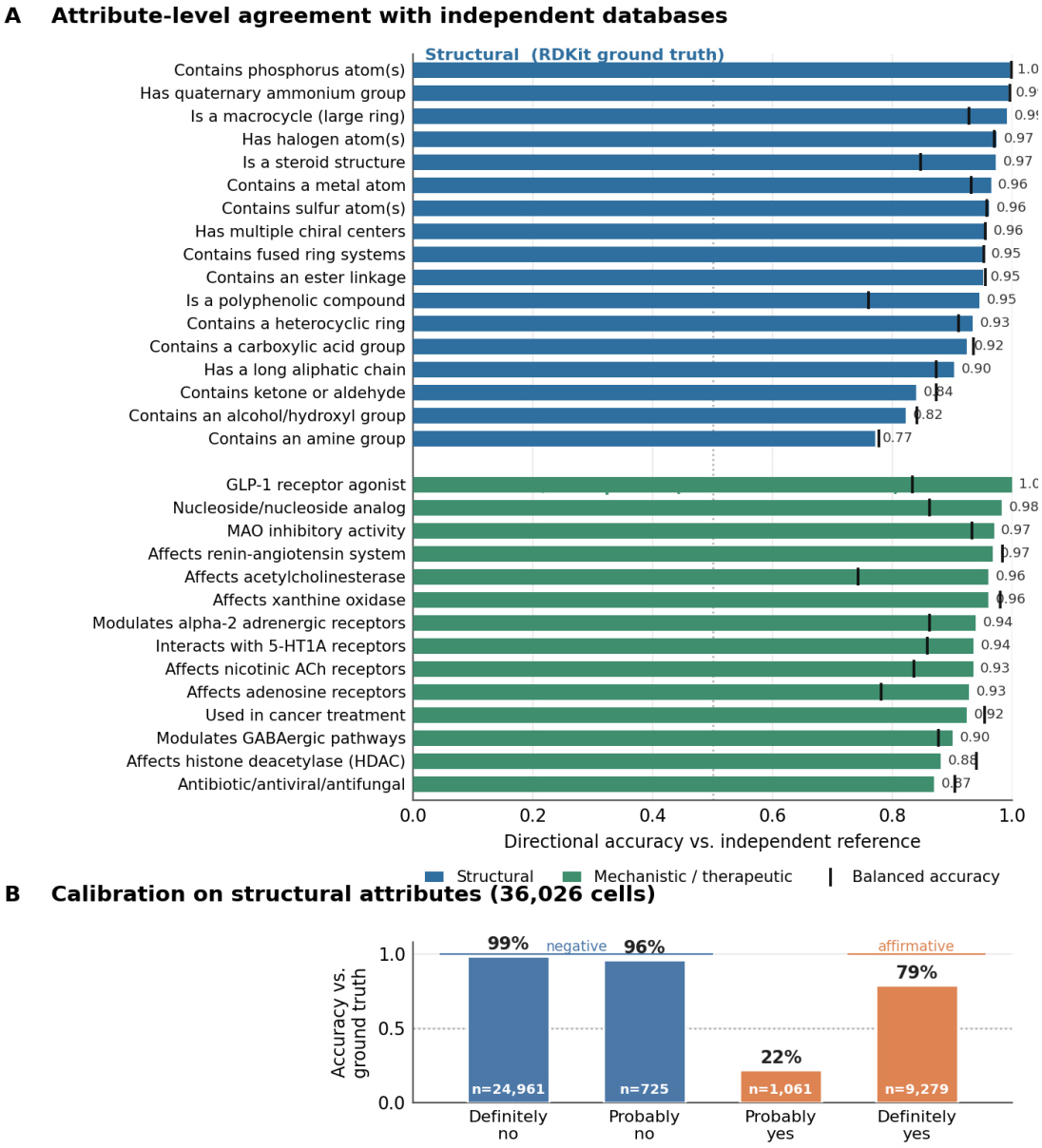

Figure 2. External Validation of Compound Attributes Against Independent Structural and Pharmacological References Panel A shows directional accuracy (bars) and balanced accuracy (vertical ticks) for each attribute against its independent reference, for the 17 structural attributes with deterministic RDKit ground truth and the 14 attributes with specific mechanistic or therapeutic referents (Drug Repurposing Hub and WHO ATC classification), across 2,079 matched compounds. Panel B shows accuracy against deterministic structural ground truth as a function of the model's response option, across 36,026 rated structural cells: negative responses are 98–99% accurate, whereas the affirmative option probably yes is correct only 22% of the time, and 88% of all errors are false-positive endorsements. Values in Panel B are cell counts.

Supplementary Material

PharmaDive: A Language Model Generates an Uncertainty-Graded Property Matrix for More Than 16,000 Compounds

Tyler M. Moore

Corresponding author. Email: tymoore@pennmedicine.upenn.edu

Contents

This supplement contains material referenced in the main manuscript: extended reviews of the curated chemical-knowledge and large-language-model extraction literatures (S1–S2), the full rationale for selecting the pharmaceutical and chemical domain (S3), the complete operationalization of the external validation criteria (S4), an extended comparison with curated chemical resources (S5), extended limitations (S6), and extended implications and future directions (S7). All tables and figures appear in the main manuscript.

Supplementary Text S1

Structured Chemical Knowledge and Its Curation Bottleneck: Extended Review

This section expands the condensed treatment of existing chemical and pharmacological data resources given in the Introduction.

The public infrastructure that addresses this problem is substantial and, within its domains, mature, and its history illustrates both the value of structured chemical information and the demanding standards required to make it reusable. At the largest scale, PubChem aggregates chemical information from more than a thousand data sources; its 2025 update reported over 300 million substance records, more than 118 million unique compounds, and hundreds of millions of bioactivity measurements, cross-linked to proteins, genes, pathways, taxa, patents, and the literature (Kim et al., 2025). Its strengths are coverage and interoperability, but its model is fundamentally that of an aggregator and archive—it organizes what others deposit—and its authors note that the automated literature pipeline, keyed to indexed titles and abstracts, misses much of the chemistry published in non-indexed venues (Kim et al., 2025). Dense, curated, compound-level pharmacological characterization is not what such a resource is designed to supply.

Foundational archives established the standards on which the field still depends. The Protein Data Bank created a global archive for experimentally determined macromolecular structures, coupling deposition with dictionary-based annotation, validation, stable identifiers, and broad dissemination (Berman et al., 2000), and the Crystallography Open Database, Cambridge Structural Database, and Inorganic Crystal Structure Database applied related principles to small-molecule and materials structures, emphasizing open access, persistent identifiers, expert editorial review, and the careful separation of experimental from theoretical records (Gražulis et al., 2009; Groom et al., 2016; Zagorac et al., 2019). The NIST Chemistry WebBook, organized by chemical species, likewise preserves multiple experimental determinations together with their source, phase, method, conditions, units, and reported precision rather than reducing each property to an undocumented canonical

number (Linstrom & Mallard, 2001). Across these resources, utility depends not only on scale but on standardization, provenance, explicit evidence classes, and the preservation of uncertainty.

For dense pharmacological characterization, the field relies on manually curated knowledgebases, each authoritative within a defined remit. DrugBank pairs each drug with detailed annotations of targets, enzymes, transporters, interactions, and mechanisms; its sixth release grew FDA-approved coverage from 2,646 to 4,563 drugs and expanded documented drug–drug interactions roughly threefold, curated by a team working from hundreds of standard operating procedures with mandatory secondary review (Knox et al., 2024). DrugCentral similarly integrates structure, bioactivity, mechanism, and indication data for approved active ingredients, distinguished by the manual curation of every ingredient's structure to a single canonical parent (Ursu et al., 2017). ChEBI provides a manually annotated ontology of small molecules of biological interest, positioned explicitly as a curation-quality gold standard (Degtyarenko et al., 2008); the Therapeutic Target Database organizes drugs and ligands around targets, pathways, and indications (Chen et al., 2002); and BindingDB concentrates on a single property class—experimentally measured protein–ligand binding affinities extracted from the primary literature, with the experimental conditions that affect those values preserved (Liu et al., 2007). Function-centered resources extend the same rigor from structure to biology: KEGG connects genes, compounds, enzymes, reactions, and curated pathway diagrams so that organism-specific systems can be reconstructed from standardized nodes (Kanehisa & Goto, 2000); the Human Metabolome Database integrates chemical identity with physiological location, concentrations, pathways, and analytical observables while explicitly distinguishing detected from biochemically predicted compounds (Wishart et al., 2022); the IUPHAR/BPS Guide to PHARMACOLOGY emphasizes selective expert curation of well-characterized ligands and quantitative target interactions (Harding et al., 2024); and ChEMBL integrates millions of bioactivity measurements across publications, patents, and assays, its curators stressing that source type, assay category, target confidence, organism, endpoint, and units must be retained because the records cannot be treated as a flat activity matrix (Zdrazil et al., 2024). ChemSpider and ZINC round out the landscape, respectively aggregating structures, identifiers, properties, and supplier links from hundreds of sources (Pence & Williams, 2010) and preparing commercially available compounds in ready-to-dock forms with calculated properties and procurement information (Irwin & Shoichet, 2005).

A third strategy replaces manual curation with automation or first-principles computation, buying scale at the cost of confinement to a particular kind of property. SureChEMBL mines chemical structures from patents through a fully automated daily pipeline, holding more than 17 million compounds from over 14 million patents (Papadatos et al., 2016); the EPA's CompTox Chemistry Dashboard federates nine databases to deliver structure-curated environmental and toxicological data for roughly 760,000 substances (Williams et al., 2017); and where the science permits, high-throughput density functional theory sidesteps the literature entirely, as in the Open Quantum Materials Database, the Materials Project, the NOMAD Laboratory, and the AFLOW framework, which generate internally consistent computed properties for hundreds of thousands of inorganic materials under standardized protocols with automated error recovery (Curtarolo et al., 2012; Draxl & Scheffler, 2019; Jain et al., 2013; Saal et al., 2013). These resources demonstrate that scale is attainable without hand curation, but each buys scale by restricting kind: chemical identity and its location in a document, a family of environmental endpoints, or a physically tractable quantity such as the total energy of a crystalline solid—which does not extend to the multi-scale, weakly formalized phenotype of a drug acting in a human body.

Supplementary Text S2

Large Language Models for Scientific Data Extraction: Extended Review

This section expands the condensed review of LLM-based extraction pipelines in materials science, chemistry, and the life sciences given in the Introduction.

A rapidly growing body of work has established that LLMs can convert scientific prose into structured data with accuracy approaching, and workflows exceeding, what earlier rule-based systems achieved. In materials science and chemistry, the dominant pattern is a staged, prompt-engineered extraction pipeline. A fully automated system parsing roughly eleven thousand alloy articles used retrieval-augmented few-shot prompting and chain-of-thought staging to extract compositions, processing conditions, and properties, reaching table-extraction F1 as

high as 0.96 with residual errors traced almost entirely to multi-step symbolic reasoning over composition formulas (Kamatchi Sundaram et al., 2026). A metal–organic-framework mining system decomposed extraction into categorization, inclusion, and retrieval stages—injecting only the schemas relevant to the properties actually detected—and reported F1 at or above 0.93 while cutting cost roughly twentyfold, and its authors found that models trained on the text-mined experimental data substantially outperformed models trained on simulated data, underscoring that literature extraction can capture signal computation alone misses (Kang et al., 2025). Subsequent systems elaborated the paradigm: the Librarian of Alexandria decoupled relevance checking from extraction and built prompts programmatically from user-defined schemas, extracting roughly 150,000 photophysical fields at 95.37% accuracy while candidly documenting low coverage as its chief limitation (Grougan et al., 2026); MOF-ChemUnity anchored extraction to a retrieval-augmented agent that resolved names at 94% accuracy and assembled a knowledge graph of more than seventy thousand properties (Pruyn et al., 2025); a multi-agent framework of heterogeneous small language models used tree search and a mask-and-reconstruct consistency test to reach accuracy competitive with much larger models while explicitly abstaining on uncertain cases (Li et al., 2025); and a few-shot model applied across some 8.4 million publications built a corpus of more than 80,000 solid-state reactions, deliberately capturing the impurity phases most datasets omit (Lee et al., 2025). Chemistry-specific systems reinforce that natural-language reasoning must be coupled to deterministic chemical representations: ComProScanner paired specialized agents and retrieval-augmented generation with a stoichiometry parser to normalize composition–property records (Roy et al., 2026), and ReactionSeek combined multimodal LLMs with optical chemical-structure recognition, name-to-structure resolution, and schema enforcement to mine a century of reaction procedures, its strongest results coming from the hybridization of flexible language models with deterministic cheminformatics and explicit negative constraints (Li et al., 2026). Recent materials databases display both the promise and the failure modes at scale: sequential prompts over full high-entropy-alloy articles produced 12,427 records but extracted composition more accurately than phase identity and introduced further downstream parsing errors, prompting retention of article-level provenance and raw outputs for audit (Chizhevskiy et al., 2026), while a nanocrystal-synthesis pipeline trained with rephrasing, negative examples, and confidence labels yielded roughly 160,000 routes whose proposed procedures still required chemical-plausibility checks and laboratory modification (Gu et al., 2026).

The same capability has been demonstrated across the life sciences, where several studies also began to probe its limits. A general-purpose model in a multi-step zero-shot protocol extracted species, ecological roles, and locations from pest-control abstracts with per-abstract recall near 99%, exceeding specialized named-entity tools, while documenting that fabricated entries were tied to output length and that model self-review was largely ineffective at catching them (Scheepens et al., 2024). FuncFetch combined literature search, screening, full-text processing, structured extraction, and human verification to mine over 32,000 enzyme–substrate entries, performing best when high-recall output was treated as a candidate set for downstream verification rather than autonomous curation (Smith et al., 2025). Herbarium-transcription and taxonomic-description systems showed that scale is achievable when prompts specify ordered fields, null behavior, units, and output format, but that domain ambiguity and model stochasticity remain consequential even when the evidence is present in the source text, with records held inactive until human review and repeated runs sometimes disagreeing about whether a field should be populated at all (Austin et al., 2026; Schmidt-Lebuhn & Knerr, 2025). Interactive literature-to-data systems responded by making grounding and validation explicit interface components, linking each extracted value back to supporting document regions or decomposing extraction into relevance screening, schema-constrained extraction, and a second self-validation pass (Cheng & Zhang, 2025; Wang et al., 2025). And a series of single-cell and metadata-classification studies mapped where the approach succeeds and fails: GPT-4 matched manual cell-type labels for most cell types at negligible cost but with imperfect reproducibility (Hou & Ji, 2024); general-purpose models standardized more than two million microbiome sample origins at near-human accuracy when categories were operationally defined (Gaio et al., 2026); benchmarking across more than a million cells found that stronger models labeled abundant populations well but that longer chain-of-thought prompts and plurality voting did not surpass a single concise pass (Crowley et al., 2025); multi-agent architectures such as CASSIA and model-consensus deliberation improved difficult annotations and converted uncertainty into a reviewable quality score, with one consensus procedure driving biologically implausible annotations from 56% to 4% over successive rounds (Xie et al., 2026; Yang et al., 2026); and localized vision-language prompting reduced hallucination in organismal images by directing attention to species-discriminative anatomy rather than whole-image gestalt

(Pahuja et al., 2026). These life-science studies matter here for two reasons: they show the approach generalizing well beyond chemistry, and they foreground the central methodological problems—hallucination, calibration, and verification—that any large-scale application must confront.

Supplementary Text S3

Why the Pharmaceutical and Chemical Domain: Extended Rationale

This section gives in full the rationale, summarized in the Introduction, for selecting pharmacology and chemistry as the setting in which to stress-test the structured-annotator approach.

The pharmaceutical and chemical domain is an unusually favorable setting in which to develop and stress-test the annotator approach, for three reasons. First, the phenotype of interest is exactly the kind of multi-scale, heterogeneous target that resists any single method, ranging from deterministic structural descriptors, which can be computed and therefore checked exactly, through mechanism-of-action and molecular-target judgments, to broad pharmacological and clinical characterizations that depend on dose, species, route, formulation, and strength of evidence. Second, the domain is anchored by an exceptionally well-characterized reference population—the few thousand marketed and heavily studied drugs curated by resources such as DrugBank, DrugCentral, ChEBI, BindingDB, ChEMBL, and the IUPHAR/BPS Guide (Degtyarenko et al., 2008; Harding et al., 2024; Knox et al., 2024; Liu et al., 2007; Ursu et al., 2017; Zdrazil et al., 2024)—against which a model's estimates can be checked, while the long tail of research compounds supplies exactly the sparsely documented cases in which an annotator, rather than an extractor, is most needed. Third, structural attributes provide a built-in objective anchor: because a molecule's atoms, rings, and bonds follow deterministically from its structure, a subset of the annotation battery can be graded against ground truth with no human judgment, calibrating confidence in the harder pharmacological items on the same compounds. Compound names introduce a countervailing difficulty—a base and its salt may share pharmacology but differ in mass and solid-state properties, brand names may denote combinations, and abbreviations can be nonunique—which, together with the tendency of web sources to collapse preclinical effects into clinical claims or repeat unsupported assertions, makes chemistry a stringent setting for testing confidence calibration and error asymmetry. A false affirmative mechanistic claim can be more misleading than a conservative negative, yet a broad matrix can remain useful if the direction and concentration of its errors are known. Beyond accelerating lookup, such a matrix makes higher-order questions tractable: structural motifs, physicochemical behavior, metabolism, mechanisms, and therapeutic contexts covary—lipophilic compounds tend to share distribution and metabolic tendencies, nucleoside analogs cluster around nucleic-acid synthesis and oncology or anti-infective use, centrally active drugs combine blood–brain-barrier penetration with neurotransmitter effects and dependence risk—so that factor analysis can summarize these patterns into a smaller number of interpretable dimensions whose value lies not in matching single textbook classes but in revealing recurring constellations of attributes, with the caveat that latent coherence and cell-level accuracy remain complementary criteria.

Supplementary Text S4

Operationalization of the External Validation Criteria

This section specifies, in full, the deterministic structural queries and the curated pharmacological criteria summarized in the Method.

Deterministic structural criteria. Eighteen attributes were operationalized as substructure, atom-composition, ring-topology, or descriptor queries evaluated with RDKit against each matched compound's curated SMILES: the presence of a ketone or aldehyde; metal, sulfur, phosphorus, or halogen atoms; a quaternary ammonium group; an amine; an alcohol or phenolic hydroxyl (excluding carboxyl hydroxyls); a carboxylic acid; an ester linkage; a heterocyclic ring; a fused ring system; a macrocycle (a ring of at least twelve atoms); multiple stereocenters; a long aliphatic chain; a steroid nucleus (a fused 6–6–6–5 ring system); a polyphenolic substructure (at least two phenolic groups); and more than five rotatable bonds. The graded lipophilicity attribute was not binarized; it was compared against the Crippen calculated logP by Spearman correlation.

Specific pharmacological criteria. Specific criteria mapped an attribute to a directly corresponding annotation in the Drug Repurposing Hub or the WHO ATC classification: monoamine-oxidase, acetylcholinesterase, xanthine-oxidase, and histone-deacetylase activity to those enzymes in a compound's target or mechanism annotation; GLP-1, α2-adrenergic, 5-HT1A, nicotinic, adenosine-receptor, and GABAergic activity to the corresponding targets; and nucleoside-analog status and antimicrobial, antineoplastic, and renin–angiotensin action to matching mechanism terms or ATC groups.

Broad pharmacological and clinical criteria. Broad criteria mapped diffuse attributes—effects on neurotransmitters, blood pressure or heart rate, hormones, the immune system, ion channels, pain signaling, and blood–brain-barrier penetration—to first-level ATC groups or disease areas, which provide only an approximate and incomplete referent. Because a curated target list or primary indication does not enumerate every activity a compound possesses, these criteria understate true prevalence; the resulting agreement statistics are therefore conservative and index recovered signal rather than adjudicating individual cells. Item-level results for all criteria appear in Table 3 of the main manuscript.

Supplementary Text S5

Relationship to Curated Chemical Knowledge Resources

This section develops in full the comparison, summarized in the Discussion, between PharmaDive and the expert-curated resources it is intended to complement.

PharmaDive occupies a different scientific layer from established chemical and pharmacological databases, and it is a complement to that curated infrastructure rather than a substitute for it. Resources such as ChEMBL preserve assay type, target confidence, organism, endpoint, units, source, and structure standardization because a bioactivity value is uninterpretable without its experimental context (Zdrazil et al., 2024); the NIST Chemistry WebBook preserves multiple determinations, original units, methods, phases, and source provenance rather than reducing a property to an undocumented canonical value (Linstrom & Mallard, 2001); HMDB explicitly distinguishes experimentally detected metabolites from predicted compounds and links chemical identity to physiology and analytical evidence (Wishart et al., 2022); and the Cambridge Structural Database combines automated processing with expert chemical review to resolve identity, disorder, and unusual bonding (Groom et al., 2016). DrugBank, the IUPHAR/BPS Guide to PHARMACOLOGY, and ChEBI likewise supply provenance, measurement detail, and sustained expert stewardship that a model-generated matrix cannot reproduce (Degtyarenko et al., 2008; Harding et al., 2024; Knox et al., 2024). PharmaDive cannot reproduce those evidentiary standards, and where a source-specific, experimentally grounded value exists—a measured potency, a pharmacokinetic parameter, a regulatory status, a documented mechanism of action—it should always be preferred to a model estimate.

Its contribution is instead breadth and uniformity. The same 130 questions are posed across approved drugs, natural and endogenous substances, investigational agents, discontinued products, research chemicals, venoms, plant extracts, and cosmetic and fragrance ingredients—including many substances for which no single curated resource provides a dense profile—so that broad qualitative questions that are rarely encoded as searchable fields, such as whether a compound is plausibly neuroactive, structurally basic and heterocyclic, or associated with oxidative-stress pathways, are asked of every compound. That uniform layer is the gap the resource fills, and it is most valuable exactly along the long tail of sparsely documented substances that motivated the project.

Used as an overlay rather than an authority, the matrix is well suited to exploratory comparison across large compound sets, to similarity search and the mapping of chemical space, to hypothesis generation about mechanism or class, to candidate prioritization, and to quality-control triage in which model–source disagreements flag records for expert attention; it can help users decide which databases or primary sources to consult, surface apparently inconsistent records, and summarize a large collection before more expensive analysis. It is not suited to adjudicating an individual mechanistic claim, populating a safety-critical field, or standing in for a measured property. This role is analogous to a search index or probabilistic annotation layer: it reduces the cost of navigating heterogeneous evidence without replacing the evidence itself. Used in that spirit, the confidence categories

become a usable instrument in their own right—negatives can be trusted, confident affirmatives taken as strong leads, and low-confidence affirmatives treated as candidates awaiting confirmation rather than as data.

Supplementary Text S6

Extended Limitations

This section gives the fuller statement of the limitations summarized in the Discussion, including the specific analytic choices that qualify the reported factor solution.

Several limitations qualify these conclusions, the most important concerning validation itself. Accuracy was estimated against a single, architecturally distinct but still general-purpose language model acting as an unblinded adjudicator rather than against human experts. Architectural separation reduces the risk of identical model-specific errors, but it does not eliminate shared training data, shared web sources, or common cultural priors, so an audit by another model can understate exactly the errors both models are prone to; and because the reviewer saw the ratings it was assessing, anchoring or confirmation effects are possible. The compound-clustered bootstrap addresses sampling uncertainty but not systematic adjudicator bias. The objective track avoids this dependency—it is anchored to deterministic cheminformatics and carries no such circularity—but it covered only a limited set of structural attributes and depended on successful structure resolution. The external comparison reported above (Table 3) provides part of this stronger validation—a large-scale benchmark against curated databases and deterministic cheminformatics selected independently of the prompts—while blinded review by multiple pharmacologists or medicinal chemists and explicit adjudication rules would strengthen it further.

Missingness raises a related concern. Unknown and not-applicable responses were combined into a single category and then recoded as missing, although epistemic uncertainty and logical inapplicability are different states; missingness was concentrated in particular attributes—one was undetermined for more than half of the validation compounds—and is unlikely to be random across compounds, since obscure substances and poorly operationalized questions probably contribute disproportionately. The factor model may consequently reflect the better-documented portion of the corpus more strongly than the long tail that motivated the project, and because the polychoric correlations were estimated from available cases, dimensions defined largely by sparse items are correspondingly less stable. Averaging repeated ratings and rounding to the nearest category yields a single analyzable matrix but can conceal disagreement between runs. Unlike the organismal precedent, I did not impute, preferring to preserve the distinction between an observed and an inferred category and heeding evidence that model-based imputations are often poorly calibrated and bettered by conventional methods (Selby et al., 2025); future versions should instead separate “unknown,” “not applicable,” and “conflicting evidence,” preserve every raw replicate, and report item- and compound-level missingness as part of the data product.

The attribute battery further mixes propositions of very different scope. A deterministic question such as whether a molecule contains chlorine is fundamentally unlike whether it affects inflammation, is anti-neurodegenerative, or requires a particular formulation, and some broad items do not distinguish a primary mechanism from a downstream consequence, therapeutic-dose evidence from high-concentration experiments, human evidence from animal or cell models, or an approved use from exploratory research. These ambiguities can inflate affirmative responses and complicate factor labels; more operational definitions, evidence-tiered variants of broad questions, and explicit context fields for species, dose, route, and evidence class would improve interpretability.

Finally, reproducibility and provenance remain central limitations. The annotation model is proprietary, web-search results change, and model updates can alter responses even when prompts are held constant, consistent with documented run-to-run and model-to-model variation in scientific extraction and annotation (Hou & Ji, 2024; Schmidt-Lebuhn & Knerr, 2025). General-purpose models may also recognize well-known compounds or heavily used datasets from training, making performance on familiar entities an uncertain mixture of synthesis and memorization (Selby et al., 2025). Although the model, prompt design, and collection procedure are documented, the current design does not attach a source trail to each rating, so exact reruns may be impossible and an individual cell cannot be audited without repeating the search. The selected 11-factor solution and its labels add further analytic discretion—formal retention criteria would generally indicate more dimensions, and the collapse of

the five-point scale to four ordered levels, the .30 loading threshold used in scoring, and the winsorizing and rescaling of scores are all defensible but consequential choices—and face validity was assessed on the same matrix from which the factors were extracted. The recovered dimensions are thus a model of the data rather than a ground truth about chemical space, and replication across model versions, alternate models, held-out compounds, and alternative factor solutions will be important.

Supplementary Text S7

Implications and Future Directions: Extended Discussion

This section sets out in full the hybrid architecture, calibration practices, and empirical extensions summarized in the Discussion.

The error profile documented here maps directly onto its own remedies, and the most defensible path forward is a hybrid architecture organized by epistemic type. Exact structural attributes should be computed from standardized structures; measured properties and curated target relations should be retrieved from authoritative databases together with their conditions and provenance; and LLM annotation should be reserved for the broad, qualitative, or cross-source judgments that are otherwise expensive to synthesize—pairing flexible language reasoning with exact chemical computation in the manner that has proved most effective in materials and reaction mining (Li et al., 2026; Roy et al., 2026). The guiding rule is to compute what can be computed, retrieve what can be retrieved, and estimate only what genuinely requires synthesis. Each estimated cell should then retain the compound identifier and form, the prompt, the model and version, the date, the response category, the missingness state, any supporting sources, and any deterministic or database-derived checks, so that the matrix functions as a versioned statistical object rather than a static fact table.

Calibration should likewise become item-specific and decision-oriented. Because the dominant failure is confident and correlated over-endorsement, “probably yes” ratings should enter a mandatory review queue rather than be treated as positive labels, and even “definitely yes” ratings on rare mechanisms should be verified before use in mechanistic, clinical, or regulatory claims; negative ratings were highly reliable in the present audit, but high-stakes absence claims should still be checked, because absence of evidence can be mistaken for evidence of absence. Multi-model consensus, adversarial review, or targeted human adjudication may be especially valuable for affirmative and low-base-rate items: querying several independent models and combining their judgments through structured deliberation has, in single-cell annotation, driven biologically implausible calls down markedly over successive rounds and converted disagreement into a reviewable quality score rather than hiding it by averaging (Xie et al., 2026; Yang et al., 2026). Directing such a procedure specifically at the low-confidence affirmatives on rare mechanistic attributes—the cells with the weakest positive predictive value—would address the resource’s specific weakness rather than its aggregate average.

Several empirical extensions would clarify scientific utility. First, the external validation begun here (Table 3) should be broadened to attribute-specific benchmarks against additional resources such as ChEMBL, DrugBank, HMDB, KEGG, measured physicochemical datasets, and prospectively curated primary literature—an approach the domain’s unusual wealth of authoritative reference data makes feasible at a scale far beyond a 200-compound audit. Second, evaluations should deliberately oversample obscure compounds, ambiguous names, salts, tautomers, and entities published after the model’s training period. Third, replicate annotations should quantify temporal and cross-model stability, and because the resource is versioned, it invites longitudinal re-annotation as models improve, with calibration tracked across releases. Fourth, the latent dimensions and the raw matrix should be tested as inexpensive features in downstream tasks—retrieval, adverse-effect prediction, target-family classification, repurposing, and active-learning prioritization—to assess whether they add incremental value beyond standard fingerprints and curated descriptors, operationalizing the distinction between aggregate usefulness and cell-level accuracy that runs throughout this work.

Supplementary References

Austin, M. W., Linan, A. G., Blomberg, M., Schmidt, H. H., & Teisher, J. K. (2026). Using large language models to automate herbarium specimen transcription: A case study at the Missouri Botanical Garden. *Plants, People, Planet, 8*(4), 1068–1075. https://doi.org/10.1002/ppp3.70128

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242.

Chen, X., Ji, Z. L., & Chen, Y. Z. (2002). TTD: Therapeutic target database. *Nucleic Acids Research, 30*(1), 412–415.

Cheng, Y., & Zhang, H. (2025). L2D: A versatile literature-to-data set pipeline for complex, user-specific data extraction in environmental research. *ACS ES&T Water, 5*(8), 4897–4907. https://doi.org/10.1021/acsestwater.5c00551

Chizhevskiy, V., Marković, G., Benrazzouq, S.-E., Cvelbar, U., Nominé, A., & Zavašnik, J. (2026). High entropy alloys database generated with large language model. *Scientific Data, 13*, 612. https://doi.org/10.1038/s41597-026-06930-z

Crowley, G., Tabula Sapiens Consortium, & Quake, S. R. (2025). Benchmarking cell type and gene set annotation by large language models with AnnDictionary. *Nature Communications, 16*, 9511. https://doi.org/10.1038/s41467-025-64511-x

Curtarolo, S., Setyawan, W., Hart, G. L. W., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M. J., Stokes, H. T., Demchenko, D. O., & Morgan, D. (2012). AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science, 58*, 218–226. https://doi.org/10.1016/j.commatsci.2012.02.005

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., & Ashburner, M. (2008). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research, 36*(Database issue), D344–D350. https://doi.org/10.1093/nar/gkm791

Draxl, C., & Scheffler, M. (2019). The NOMAD laboratory: From data sharing to artificial intelligence. *Journal of Physics: Materials, 2*(3), 036001. https://doi.org/10.1088/2515-7639/ab13bb

Gaio, D., Tackmann, J., Perez-Molphe-Montoya, E., Näpflin, N., Patsch, D., Malfertheiner, L., Peluso, M. E., & von Mering, C. (2026). Enhanced semantic classification of microbiome sample origins using large language models (LLMs). *GigaScience, 15*, giag015. https://doi.org/10.1093/gigascience/giag015

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., & Le Bail, A. (2009). Crystallography Open Database—An open-access collection of crystal structures. *Journal of Applied Crystallography, 42*(4), 726–729. https://doi.org/10.1107/S0021889809016690

Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials, 72*, 171–179. https://doi.org/10.1107/S2052520616003954

Grougan, M., Subasinghe, J., Hix, M. A., & Walker, A. R. (2026). Librarian of Alexandria: A modular chemical data extraction pipeline to compare LLM performance. *Journal of Chemical Information and Modeling, 66*, 6921–6932. https://doi.org/10.1021/acs.jcim.6c00374

Gu, K., Liang, Y., Peng, S., Guo, A., Fu, Y., & Zhong, H. (2026). A large-scale nanocrystal database with aligned synthesis and properties, enabling generative inverse design. *ACS Nano, 20*, 17413–17422. https://doi.org/10.1021/acsnano.6c03070

Harding, S. D., Armstrong, J. F., Faccenda, E., Southan, C., Alexander, S. P. H., Davenport, A. P., Spedding, M., & Davies, J. A. (2024). The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Research, 52*(D1), D1438–D1449. https://doi.org/10.1093/nar/gkad944

Hou, W., & Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods, 21*(8), 1462–1465. https://doi.org/10.1038/s41592-024-02235-4

Irwin, J. J., & Shoichet, B. K. (2005). ZINC—A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling, 45*(1), 177–182. https://doi.org/10.1021/ci049714+

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials, 1*(1), 011002. https://doi.org/10.1063/1.4812323

Kamatchi Sundaram, A., Chakraborty, M., Devathi, S. M. K., Prusty, B. P., & Batra, R. (2026). Automated extraction of multicomponent alloy data using large language models for sustainable design. *Advanced Science*, Article e75916. https://doi.org/10.1002/advs.75916

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 28*(1), 27–30.

Kang, Y., Lee, W., Bae, T., Han, S., Jang, H., & Kim, J. (2025). Harnessing large language models to collect and analyze metal–organic framework property data set. *Journal of the American Chemical Society, 147*(6), 3943–3958. https://doi.org/10.1021/jacs.4c11085

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2025). PubChem 2025 update. *Nucleic Acids Research, 53*(D1), D1516–D1525. https://doi.org/10.1093/nar/gkae1059

Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N. E., Strawbridge, S. A., Garcia-Patino, M., Kruger, R., Sivakumaran, A., Sanford, S., Doshi, R., Khetarpal, N., Fatokun, O., Doucet, D., Zubkowski, A., … Wishart, D. S. (2024). DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Research, 52*(D1), D1265–D1275. https://doi.org/10.1093/nar/gkad976

Lee, S., Cruse, K., Baibakova, V., Ceder, G., & Jain, A. (2025). Text-mined dataset of solid-state syntheses with impurity phases using large language model. *Scientific Data, 12*, Article 1969. https://doi.org/10.1038/s41597-025-06222-y

Li, J., Li, M., Yang, Q., & Luo, S. (2026). ReactionSeek: LLM-powered literature data mining and knowledge discovery in organic synthesis. *Nature Communications, 17*, 3356. https://doi.org/10.1038/s41467-026-70180-1

Li, X., Huang, Z., Quan, S., Peng, C., & Ma, X. (2025). SLM-MATRIX: A multi-agent trajectory reasoning and verification framework for enhancing language models in materials data extraction. *npj Computational Materials, 11*, Article 241. https://doi.org/10.1038/s41524-025-01719-x

Linstrom, P. J., & Mallard, W. G. (2001). The NIST Chemistry WebBook: A chemical data resource on the Internet. *Journal of Chemical & Engineering Data, 46*(5), 1059–1063. https://doi.org/10.1021/je000236i

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research, 35*(Database issue), D198–D201. https://doi.org/10.1093/nar/gkl999

Pahuja, V., Stevens, S., East, A., Record, S., & Su, Y. (2026). Automatic image-level morphological trait annotation for organismal images. In *Proceedings of the International Conference on Learning Representations (ICLR 2026)*. https://arxiv.org/abs/2604.01619

Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Irvine, S. A., Pettersson, J., Goncharoff, N., Hersey, A., & Overington, J. P. (2016). SureChEMBL: A large-scale, chemically annotated patent document database. *Nucleic Acids Research, 44*(D1), D1220–D1228. https://doi.org/10.1093/nar/gkv1253

Pence, H. E., & Williams, A. (2010). ChemSpider: An online chemical information resource. *Journal of Chemical Education, 87*(11), 1123–1124. https://doi.org/10.1021/ed100697w

Pruyn, T. M., Aswad, A., Khan, S. T., Huang, J., Black, R., & Moosavi, S. M. (2025). MOF-ChemUnity: Literature-informed large language models for metal–organic framework research. *Journal of the American Chemical Society, 147*(45), 43474–43486. https://doi.org/10.1021/jacs.5c11789

Roy, A., Grisan, E., Buckeridge, J., & Gattinoni, C. (2026). ComProScanner: A multi-agent based framework for composition-property structured data extraction from scientific literature. *Digital Discovery, 5*, 1794–1808. https://doi.org/10.1039/d5dd00521c

Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM, 65*(11), 1501–1509. https://doi.org/10.1007/s11837-013-0755-4

Scheepens, D., Millard, J., Farrell, M., & Newbold, T. (2024). Large language models help facilitate the automated synthesis of information on potential pest controllers. *Methods in Ecology and Evolution, 15*, 1261–1273. https://doi.org/10.1111/2041-210X.14341

Schmidt-Lebuhn, A. N., & Knerr, N. (2025). Large language models can extract morphological data from taxonomic descriptions, but their stochastic nature makes automation challenging: A test on Australian Asteraceae. *PhytoKeys, 261*, 189–210. https://doi.org/10.3897/phytokeys.261.158396

Selby, D., Iwashita, Y., Spriestersbach, K., Saad, M., Bappert, D., Warrier, A., Mukherjee, S., Kise, K., & Vollmer, S. (2025). Had enough of experts? Quantitative knowledge retrieval from large language models. *Stat, 14*(2), e70054. https://doi.org/10.1002/sta4.70054

Smith, N., Yuan, X., Melissinos, C., & Moghe, G. (2025). FuncFetch: An LLM-assisted workflow enables mining thousands of enzyme-substrate interactions from published manuscripts. *Bioinformatics, 41*(1), btae756. https://doi.org/10.1093/bioinformatics/btae756

Ursu, O., Holmes, J., Knockel, J., Bologa, C. G., Yang, J. J., Mathias, S. L., Nelson, S. J., & Oprea, T. I. (2017). DrugCentral: Online drug compendium. *Nucleic Acids Research, 45*(D1), D932–D939. https://doi.org/10.1093/nar/gkw993

Wang, X., Huey, S. L., Sheng, R., Mehta, S., & Wang, F. (2025). SciDaSynth: Interactive structured data extraction from scientific literature with large language model. *Campbell Systematic Reviews, 21*, e70073. https://doi.org/10.1002/cl2.70073

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., & Richard, A. M. (2017). The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *Journal of Cheminformatics, 9*, 61. https://doi.org/10.1186/s13321-017-0247-6

Wishart, D. S., Guo, A., Oler, E., Wang, F., Anjum, A., Peters, H., Dizon, R., Sayeeda, Z., Tian, S., Lee, B. L., Berjanskii, M., Mah, R., Yamamoto, M., Jovel, J., Torres-Calzada, C., Hiebert-Giesbrecht, M., Lui, V. W., Varshavi, D., Varshavi, D., … Gautam, V.

(2022). HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research, 50*(D1), D622–D631. https://doi.org/10.1093/nar/gkab1062

Xie, E., Cheng, L., Shireman, J., Cai, Y., Liu, J., Mohanty, C., Dey, M., & Kendziorski, C. (2026). CASSIA: A multi-agent large language model for automated and interpretable cell annotation. *Nature Communications, 17*, 389. https://doi.org/10.1038/s41467-025-67084-x

Yang, C., Zhang, X., & Chen, J. (2026). Large language model consensus substantially improves the cell type annotation accuracy for scRNA-seq data. *Communications Biology*. Advance online publication. https://doi.org/10.1038/s42003-026-10420-8

Zagorac, D., Müller, H., Ruehl, S., Zagorac, J., & Rehme, S. (2019). Recent developments in the Inorganic Crystal Structure Database: Theoretical crystal structure data and related features. *Journal of Applied Crystallography, 52*, 918–925. https://doi.org/10.1107/S160057671900997X

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., de Veij, M., Ioannidis, H., Mendez Lopez, D., Mosquera, J. F., Magarinos, M. P., Bosc, N., Arcila, R., Kizilören, T., Gaulton, A., Bento, A. P., Adasme, M. F., Monecke, P., Landrum, G. A., & Leach, A. R. (2024). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research, 52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004