Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PACSYLLM: natural-language access to an expanded PACSY database of protein structures and chemical shifts

Abigail Chiu
University of Colorado Denver
Woonghee Lee
woonghee.lee@ucdenver.edu
University of Colorado Denver

Research Article

Keywords: PACSY, chemical shift, protein structure, BMRB, text-to-SQL, natural-language interface

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11816282/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PACSYLLM: natural-language access to an expanded PACSY database of protein structures and chemical shifts

Abigail Chiu[1,2,3], Woonghee Lee[2,3,*]

[1]Department of Integrative Biology, University of Colorado Denver, CO, USA

[2]Department of Chemistry, University of Colorado Denver, CO, USA

[3]Center for Advanced Computational Molecular Science, University of Colorado Denver, CO, USA

*Corresponding author: woonghee.lee@ucdenver.edu

Abstract

Chemical shifts are the first data obtained in protein NMR studies, and their relationship to three-dimensional structure underlies resonance assignment, secondary-structure analysis, and structure determination. In 2012, PACSY made this relationship searchable by integrating NMR structures from the Protein Data Bank (PDB), chemical shifts from the Biological Magnetic Resonance Data Bank (BMRB), and fold classifications from the Structural Classification of Proteins (SCOP). Many analyses, however, still require SQL queries or custom scripts. Here, we describe a rebuilt, quality-controlled PACSY database and PACSYLLM, a natural-language interface to its data. The database holds 11,537 PDB/BMRB chain pairs, 10,881 of them protein chains, with 7.9 million chemical shifts and 239 million atomic coordinates, 1.6 to 3.9 times the 2012 content. It also stores derived annotations for shift-based secondary-structure and torsion-angle prediction, chemical-shift re-referencing, fold classification, ensemble quality, and covalent geometry. Chemical shifts are matched to the correct molecular entity and checked against their NMR-STAR source files. PACSYLLM keeps scientific computation outside the language model: deterministic programs handle well-defined outputs, while a self-hosted 27-billion-parameter model generates one validated, read-only SQL query and summarizes the returned data. PACSYLLM answered all 26 held-out questions correctly in three repeats and answered reworded versions correctly. The database queries of six published PACSY analyses, each posed as a single natural-language request, returned the published or independently verified results.

Keywords: PACSY; chemical shift; protein structure; BMRB; text-to-SQL; natural-language interface

1. Introduction

Protein NMR spectroscopy produces several types of information that must be interpreted together to determine a structure. An early and essential step is resonance assignment, in which each signal in the spectra is assigned to a specific nucleus in the protein. Distances between protons are then derived from NOE measurements, while backbone torsion angles can be predicted from the chemical shifts (Cornilescu et al. 1999; Shen and Bax 2013). Structure-calculation programs, such as CYANA, PONDEROSA-C/S and Xplor-NIH (Schwieters et al. 2003; Lee et al. 2014; Güntert and Buchner 2015), use these and other restraints to compute an ensemble of models consistent with experimental data.

As such, chemical shifts are particularly central to the NMR structure-determination process. They are the first data an NMR spectroscopist obtains, and their values depend on each atom's sensitivity to its local structural environment. For example, the Cα chemical shift of a residue in an α-helix differs from that of the same residue type in a β-strand. This relationship forms the basis of the chemical shift index (Wishart et al. 1992; Wishart and Sykes 1994), torsion-angle prediction programs, resonance assignment methods, and chemical shift validation. In practice, researchers often encounter unusual resonances during their analysis. They may wonder whether comparable shifts have been reported previously, which residue types or structural environments produce them, or whether an observed value is consistent with the proposed local structure. These questions arise naturally while assigning spectra, evaluating unexpected peaks, and interpreting a newly determined protein structure.

Answering these questions and generating the statistics for new analysis programs at scale requires combining information from several public resources. The Protein Data Bank (PDB) stores atomic coordinates (Berman et al. 2007), whereas the Biological Magnetic Resonance Data Bank (BMRB) stores chemical shifts and other NMR data (Ulrich et al. 2008; Hoch et al. 2022). Structural classification resources, such as the Structural Classification of Proteins (SCOP; Murzin et al. 1995), its successor SCOPe (Chandonia et al. 2022), and CATH (Sillitoe et al. 2021), group protein domains into classes, folds, and families by their structural and evolutionary relationships. Although most NMR structures are deposited in both PDB and BMRB, they are not often classified in SCOP. These classifications can let a user restrict a question to all-α proteins, or compare proteins of the same fold. Each of these three resources is maintained separately, with different file formats and identifiers, and the PDB and BMRB often number residues differently.

Consequently, even a conceptually simple question can require substantial data integration. For example, a simple question, such as *"what is the typical Cα shift of an alanine in a helix of an all-α protein?"*, requires matching the appropriate PDB and BMRB entries, reconciling chains and residue numbering, aligning sequences, deriving structural annotations from the coordinates, and integrating fold classifications before the desired statistic can be calculated. Investigating an unusual experimental observation can be even less straightforward because the researcher may not know in advance which database fields, structural annotations, or filters are needed. The difficulty

is therefore not simply the availability of NMR data, but the ability to interrogate those data in the form of the scientific questions that arise during analysis.

1.1. PACSY

PACSY (Protein structure And Chemical Shift NMR spectroscopY) was developed in 2012 to solve this underlying integration problem (Lee et al. 2012). This relational database linked structures from the PDB, chemical shifts from the BMRB, and fold classifications from SCOP so that these data could be retrieved together using SQL (Codd 1970). The initial release was built from 3,745 PDB/BMRB pairs and held 7,395 chains, 2.0 million chemical shifts, and 75.9 million atomic coordinates. The data were organized into six table types, four of which were further divided into twenty amino acid tables. Users accessed PACSY either by typing SQL on the PACSY website or through PACSY Analyzer, a graphical program that generated SQL from user-selected input and output fields. Still, both approaches required users to understand how a biological question mapped onto the database organization. This limitation is illustrated by an example in the original PACSY paper that asked for the range of alanine $^{13}C\alpha$ shifts expected in an all-α protein of about 90 residues at low pH. Expressing this question required seven input filters and twelve output fields. More extensive analysis, such as backbone-shift statistics across secondary-structure classes and the dependence of alanine $^{1}H\alpha$ shifts on solvent exposure, required additional Python scripts.

Despite these barriers to direct querying, PACSY became widely used, and the literature that cites it shows two patterns of use (Figure 1; SI Table S11). Within the NMRFAM-SPARKY and POKY software, PACSY operates behind the interface. Its statistics drive predict-and-confirm side-chain assignment (Lee et al. 2015, 2016a), which POKY extended to solid-state spectra (Dwarasala et al. 2022), the NOE checks of AUDANA (Lee et al. 2016b), and the expected shift distributions of I-PINE webserver (Lee et al. 2019), and further tools and BMRB annotations build on it (Lee et al. 2016a; Yokochi et al. 2016; Lee and Markley 2018; Manthey et al. 2022; Dwarasala

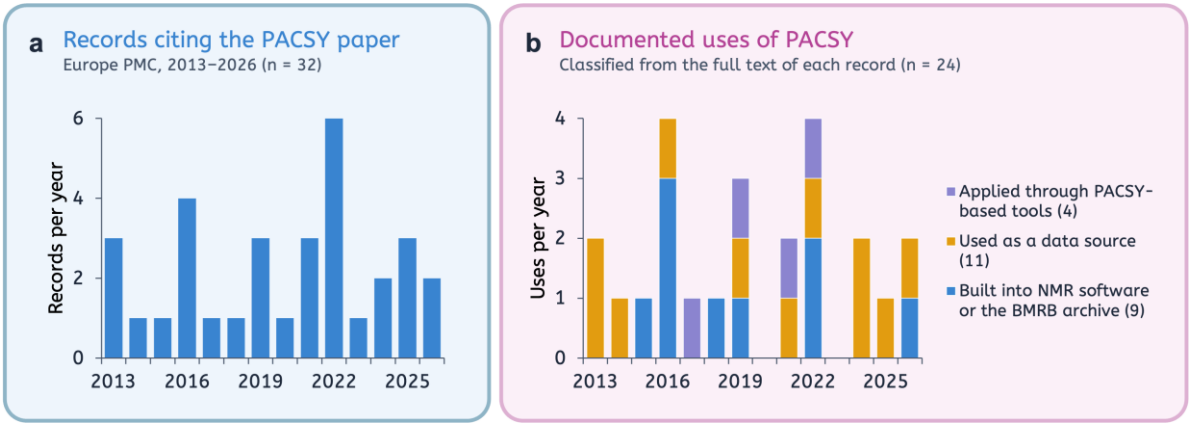


Figure 1. Use of PACSY in the literature. (a) Europe PMC records citing Lee et al. (2012) per year (32 records, 2013–2026; a preprint of a journal article is counted once). (b) Documented uses per year, classified from the full text of each record (24 uses; SI Table S11): built into NMR software or the BMRB archive, used as a data source (directly or through PLUQ and PLUQin), or applied through these PACSY-based tools.

et al. 2022; Chiu and Lee 2026). Outside this ecosystem, researchers exported PACSY and reprocessed it before analysis. Fritzsching et al. built the PLUQ residue-typing tool on a reduced SQLite copy (Fritzsching et al. 2013) and identified misreferenced ^{13}C shifts in a local MySQL import (Fritzsching et al. 2016), work that later studies built on (Yang et al. 2013; Pérez-Conesa et al. 2021; Emmanouilidis et al. 2024; Somberg et al. 2025; Weiss et al. 2026), and other groups cleaned exports (Burakova et al. 2022), wrote their own extraction scripts (Pritchard and Hansen 2019), or added separately computed predictions (Li et al. 2024).

The remaining barrier, therefore, is not whether PACSY can support these analyses but how directly a researcher can ask for them. Most users think in terms of experimental questions rather than database schemas, such as whether an unusual peak has precedents, which structural environments are associated with a chemical shift, how a protein compares with related structures, or whether an observation is typical of a particular residue or secondary-structure class.

1.2. PACSYLLM

Two developments now make natural-language access to PACSY practical. First, commonly needed results, including torsion angles and secondary-structure predictions from chemical shifts, corrections for chemical shift referencing errors, CATH and SCOPe fold classifications, and statistics on the well-ordered core of each structural ensemble, can be computed in advance and stored with existing PACSY data (Eghbalnia et al. 2005; Wang et al. 2005; Kirchner and Güntert 2011; Shen and Bax 2013; Sillitoe et al. 2021; Nielsen and Mulder 2021; Chandonia et al. 2022). Second, large language models (LLMs) can translate a plain-text question into SQL. This task, known as text-to-SQL, has established benchmarks (Yu et al. 2018; Li et al. 2023) and prompted strategies (Pourreza and Rafiei 2023). An openly available model can also be run on the database provider's own server, which allows questions and database contents to remain local. Still, because an LLM can generate a valid query that does not match the intended question, or distort numerical results when summarizing them, an interface is needed to safeguard scientific computation and validation outside the language model.

Here, a rebuilt and expanded PACSY is described together with PACSYLLM, a natural-language interface to its data. The new PACSY uses a modern relational database management system while retaining the 2012 table names for compatibility with existing software. The chemical shift data underwent systematic quality control. Each shift is paired with the correct molecule and residue and checked against its source file, and values that represent missing data are stored as missing. New structural and chemical shift-derived data layers were added, so that analyses previously performed separately can be queried directly. Finally, PACSYLLM was designed around a simple design principle, which is to keep any operation that can be performed deterministically outside the language model.

PACSYLLM provides natural-language access to a quality-controlled database, delegates well-defined scientific computations to conventional programs, and limits the language model to generating validated, read-only SQL and summarizing returned results. The system was evaluated

using held-out questions, examples from its user documentation, and six analyses from the original PACSY literature, each reformulated as a single natural-language question. Complete results and generated SQL queries are provided in the Supplementary Information (SI).

2. Materials and methods

2.1. Data sources and entry matching

Similar to the 2012 release, the coordinate file from PDB, chemical shifts from BMRB, and fold classification from SCOPe and CATH were obtained for each protein chain. External programs and in-house routines computed the derived layers, and the complete chain was written to the database. The data are organized in normalized tables centered on two tables, ENTRY (one row per PDB/BMRB chain pair) and RESIDUE. Users and programs can reach the data through SQL, views that reproduce the 2012 table names, or through PACSYLLM.

Coordinate files (mmCIF) from the RCSB PDB and NMR-STAR files from the BMRB are paired for each chain. As in the original PACSY database, only validated PDB/BMRB matches deposited by the same authors were included (Lee et al. 2012). Each PACSY entry corresponds to one PDB/BMRB/chain combination. mmCIF and NMR-STAR files were parsed with GEMMI (Wojdyr 2022) and PyNMRSTAR (Hoch et al. 2022). Structural classifications were obtained from SCOPe 2.08 (Chandonia et al. 2022), the updated successor of the SCOP database used in the original PACSY, and from CATH (Sillitoe et al. 2021). Classifications for each entry were stored in the CLASSIFICATION table (Table 1) with the full SCOPe reference index (Table 3).

Because a BMRB entry may contain multiple molecular entities, such as a protein and a bound peptide, each PDB chain was aligned with every BMRB entity with the Smith–Waterman algorithm (Smith and Waterman 1981), and the best match was accepted only when the sequences were sufficiently concordant (SI Section S3.1). Only residue pairs of the same amino-acid type were mapped, and chemical shifts were imported only from the matched BMRB entity, using the shift list and assembly member with the most assignments. Chains without a matching BMRB entity were retained without chemical shifts or chemical shift-derived annotations.

Entries were constructed one at a time. Each chain was written in a single database transaction, so an unsuccessful build could not leave a partially populated entry (Figure 2). Separate update procedures add new derived layers to existing entries and incorporate newly deposited PDB/BMRB pairs as they appear (SI Section S2.2).

2.2. Derived structural and chemical shift annotations

2.2.1. Residue and ensemble level annotations

Table 1 lists the derived data layers and the typical use of each. STRIDE (Frishman and Argos 1995) was applied to every deposited model to assign secondary structure (B, isolated β-bridge; C, coil; E, β-strand; G, 3_{10} helix; H, α-helix; I, π-helix; T, turn), backbone torsion angles φ and ψ, and solvent-accessible surface area. DSSP (Kabsch and Sander 1983) was also applied to every model

as a second, independent secondary-structure assignment. The hydrophobicity scale HDO_PBT of the original PACSY was calculated as the accessible area divided by the Gly-X-Gly reference value of Lee et al. (2012). LACS (Wang et al. 2005), which detects systematic referencing errors in ^{13}C shifts, was run with pyLACS, its Python reimplementation (Manthey et al. 2022). The resulting atom-class-specific offsets were applied to the corresponding chemical shifts.

Several additional layers describe the ensemble of models that make up an NMR structure. Cyrange (Kirchner and Güntert 2011) identified the well-ordered core, defined as the residues that are sufficiently well-determined for comparison among models. The ensemble RMSD measures how much the models differ. This was quantified by superimposing the core Cα atoms of each model onto the mean structure using three iterations of Kabsch superposition (Kabsch 1976). Backbone RMSDs were then calculated per atom, per residue, and per model. A separate conformer layer tests whether an ensemble shows evidence for more than one distinct state by clustering Cα–Cα distance maps of the core, testing the split against a matched single-conformer null model, and requiring a minimum structural separation between clusters (SI Section S3.6). This layer is used as an exploratory indicator of possible conformational heterogeneity rather than as a definitive assignment of conformational states. Finally, PDB2PQR (Dolinsky et al. 2004) was used to assign AMBER charges and radii to every atom of the representative model for subsequent electrostatics calculations.

Table 1. Derived and integrated data layers in the rebuilt PACSY database. "Entries" gives the number of the 11,537 entries with at least one row in the layer. Of these entries, 10,881 are protein chains with residue data, and 656 are nucleic-acid chains (Section 3.1).

Layer	Method	Typical use	Table(s)	Entries
Assigned chemical shifts	BMRB deposition	Experimental shifts for every assigned atom	CHEMSHIFT	10,518
Secondary structure, φ/ψ, accessibility	STRIDE, per model	Structure and exposure of each residue	STRUCTURE	10,036
Torsion angles and secondary structure from shifts	TALOS-N	Shift-based φ/ψ restraints	TALOSN_PREDICTION	10,465
Secondary-structure populations from shifts	PECAN	Shift-based secondary structure	PECAN_PREDICTION	10,517
Secondary-structure class from shifts	CheSPI	Shift-based secondary structure	CHESPI_PREDICTION	10,189
Referencing offsets	LACS	Detecting referencing errors	LACS	9,442
LACS-corrected shifts	LACS	Statistics on re-referenced shifts	LACS_CHEMSHIFT	9,438
Well-ordered core	Cyrange	Residues to compare between models	CYRANGE_REGION	9,439

Layer	Method	Typical use	Table(s)	Entries
Ensemble RMSD, per atom / residue / model	Kabsch superposition on Cyrange core	Precision of the ensemble	RMSD_*	9,318
Conformer clustering	PCA of core Cα–Cα maps, Ward, null-model test	Signs of more than one state (exploratory)	CONFORMER_*	9,232
Structure quality	MolProbity (Chen et al. 2010)	Geometric quality of the models	MOLPROBITY	9,455
Per-atom charge and radius (AMBER)	PDB2PQR	Electrostatics input	PQR_DB	10,881
Peptide-bond ω and cis/trans state	from coordinates, per model; wwPDB cis annotation	Cis peptide bonds	STRUCTURE.omega, PEPTIDE_BOND	10,854
Disulfide bonds	Sγ–Sγ distance per model; wwPDB _struct_conn	Disulfide connectivity	DISULFIDE	1,948[a]
Secondary structure and accessibility, second assignment	DSSP, per model	Second secondary-structure assignment	DSSP	10,871
Fold classification	SCOPe 2.08 / CATH	Selecting proteins by fold	CLASSIFICATION	4,165 / 5,613
Simulated 2D peak lists (N-HSQC / C-HSQC / NCA / NCO)	from assigned shifts	Expected spectra	*_CONTENT	7,750 / 7,286 / 7,862 / 5,802[b]
LACS-corrected peak lists (same four)	from LACS-corrected shifts	Expected spectra after re-referencing	LACS_*_CONTENT	7,452 / 6,862 / 7,496 / 5,640[b]
Deposition authors and citation	BMRB entry saveframes	Finding entries by author or paper	AUTHOR, CITATION	11,513 / 11,508

[a] Of the 5,716 entries that contain cysteine. [b] Entries whose simulated list contains at least one peak. The database stores a peak-list record for every entry with an NMR-STAR file; the record is empty when the entry has no assigned atom pair of the required type, such as an amide H–N pair for $^{1}H,^{15}N$-HSQC.

2.2.2. Covalent geometry

Two additional layers describe cis/trans peptide-bond geometry and disulfide connectivity, both computed from the deposited mmCIF coordinates of every model. The ω torsion of each peptide bond is stored with the following residue, so that an X–Pro bond is stored with the proline, and each model is classified as cis, trans, or twisted; PEPTIDE_BOND summarizes each bond over the ensemble and records whether the deposition lists it as cis. DISULFIDE holds every Cys–Cys pair whose Sγ atoms are within bonding distance in at least one model or that the deposition lists as a disulfide, with the number of models in which the bond is formed. Shift-based indicators

of both states, the proline Cβ−Cγ difference and the cysteine Cβ shift, are obtained by joining to CHEMSHIFT rather than stored separately (SI Section S3.7).

2.3. Quality control procedures

Chemical shift assignments underwent two independent consistency checks. First, every stored atom assignment was compared with the atoms expected for its residue type. For example, glycine has no Cβ and alanine no Cγ. Second, chemical shifts were independently re-derived from the original NMR-STAR files and compared row by row with the corresponding database records. These checks were also used to test the handling of multiple BMRB entities, shift lists, and residue mappings. Cases involving shifts from an incorrect entity or shift list, or mappings between different residue types, were rejected. Unit tests cover these cases (SI Section S3).

Numeric fields were screened for sentinel values used by depositors or external programs to represent missing data, including ±999, ±9999, and ±99999. These values were converted to NULL when they represented missing rather than numerical measurements. This procedure was applied both to metadata migrated from the original PACSY release and to derived prediction layers, including undefined TALOS-N φ, ψ, and χ1 values (SI Section S3.2). Coverage of the derived layers was also checked systematically. Because TALOS-N cannot process non-standard residues, sequences containing them were divided into contiguous stretches of standard amino acids, and TALOS-N was run separately on each stretch (SI Section S3.4).

Finally, stored atomic coordinates were compared with the corresponding deposited mmCIF files. Oxygen and sulfur atoms, which the original coordinate tables did not contain, were taken from the deposited structure for migrated entries, either directly or after model-wise superimposition, and entries stored from an older version of a deposition were built from the current mmCIF file. Structure-derived layers, including per-atom charge and ensemble RMSD, were computed from the completed coordinates (SI Section S3.5).

2.4. Database organization

The per-amino-acid tables used in the original release of PACSY were replaced by 40 normalized tables linked through entry and residue identifiers (Figure 2; SI Table S5). In the normalized schema, each piece of information is stored once, and attributes such as residue type, chain, and model number appear as columns rather than in table names. To preserve compatibility with existing software, the database also provides 114 views that reproduce the table and column names of the original schema. These views contain no independent data. The records from the normalized tables are presented using the same PACSY organizations (SI Tables S3 and S4). They include SEQ_DB and PDBSEQ_DB, the 100 per-residue views: X_DB, X_STRC_DB, X_CS_DB, and X_COORD_DB as in the original PACSY database, and X_LACS_CS_DB for LACS-corrected shifts, where X represents each of the 20 standard amino acids. The residue-level views keep the original column names and include an additional FIRSTKEY_ID field that links each residue to its parent chain.

2.5. PACSYLLM

2.5.1. Design principle

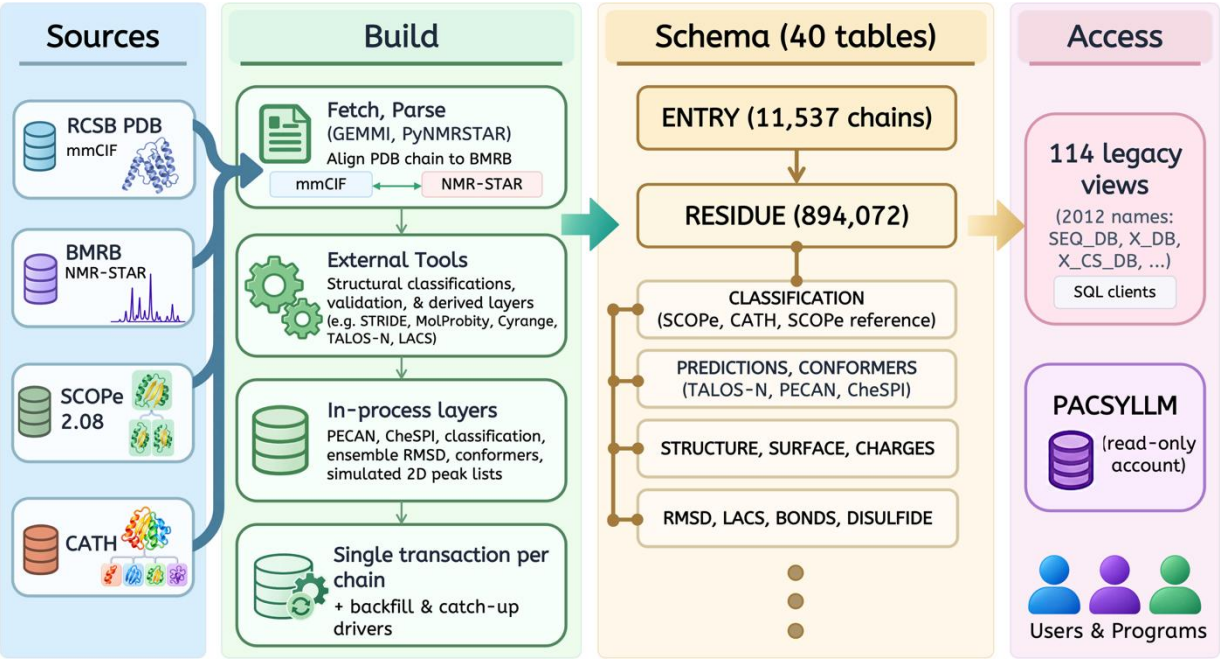


Figure 2. Building and organization of the rebuilt PACSY database. Each PDB/BMRB chain pair is fetched, parsed and aligned. The external tools and in-process layers are then run, and the chain is written to the database in one transaction. The normalized schema of 40 tables is centered on ENTRY and RESIDUE, with residue-level and entry-level tables. 114 views recreate the 2012 table names for SQL clients, and PACSYLLM reads the database through a read-only account.

PACSYLLM provides the database back end for the LLM assistant in the POKY software suite (Lee et al. 2021). It runs on the PACSY database server and follows a simple design principle: tasks with well-defined, deterministic solutions are handled by conventional programs rather than by the language model. When natural-language interpretation is needed, it mainly translates a question into a database query or summarizes information returned by the system. This separation reduces the risk that the model will alter numerical values or perform scientific calculations that deterministic software can carry out more reliably. Access to the database is also restricted. Model-generated SQL is validated before execution and is run through a read-only database account. Whenever possible, returned values are passed through without being rewritten by the model. Stored text, such as a peak list, is returned verbatim. For plotting requests, the returned numerical values are inserted directly into a matplotlib script.

PACSYLLM used a self-hosted 27-billion-parameter model derived from Qwen3.8, served with vLLM (Kwon et al. 2023), at temperature 0 (SI Section S2.3).

2.5.2. Request routing

Each user request passes through up to four stages (Figure 3). The stages are ordered so that deterministic programs are used whenever possible, and the language model is used only when the request requires interpretation.

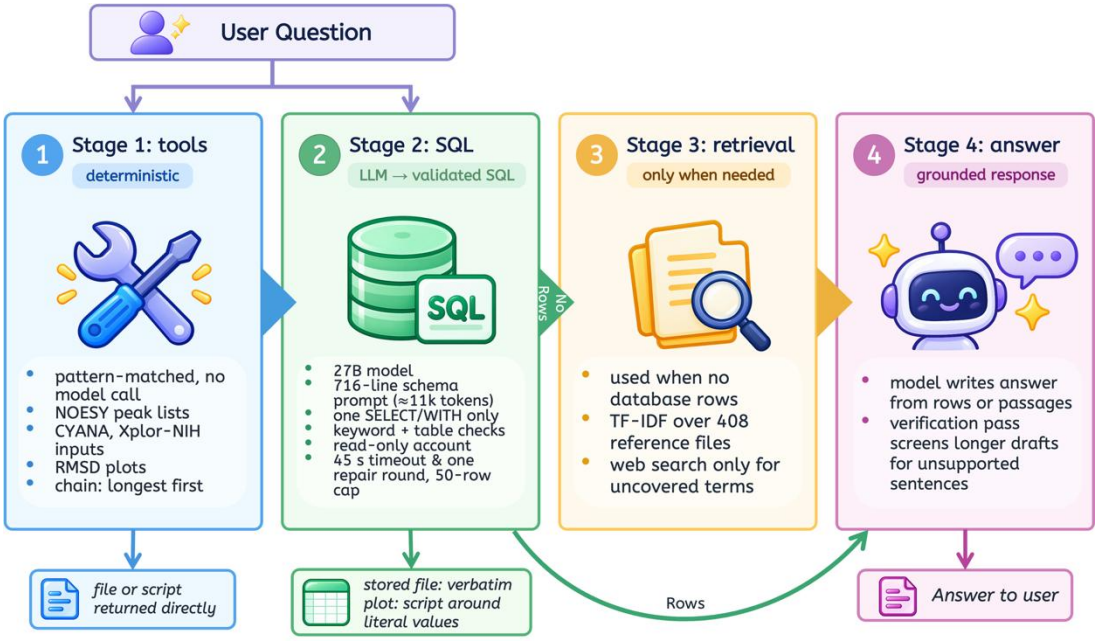


Figure 3. How PACSYLLM answers a question. Stage 1 answers requests for well-defined outputs with fixed programs. In Stage 2 the model writes one SQL statement, which is validated and executed read-only (Table 2). Stored file text and plot scripts leave Stage 2 directly, and other results go to Stage 4. Questions the database cannot answer go to Stage 3 retrieval; an empty query result is still passed to Stage 4, and a count of zero is returned as a database result. Stage 4 composes the answer and screens drafts of 400 characters or more for unsupported sentences.

Stage 1: deterministic outputs. Requests with a single well-defined output for a given PACSY entry are recognized by pattern matching and handled without calling the language model. These include 2D and 3D ^{15}N- or ^{13}C-edited NOESY peak lists generated by searching for proton pairs within 5.0 Å after averaging distances across models with a minimum sequence separation of two residues, FASTA sequences and input files for CYANA (Güntert and Buchner 2015) and Xplor-NIH (Schwieters et al. 2003), and per-residue Cα or backbone RMSD plots, returned as executable matplotlib scripts containing the stored RMSD values. The simulated NOESY lists apply a simple distance cutoff to the deposited coordinates, with peak heights scaled as r^{-6} on an arbitrary scale, and do not represent experimental NOESY data (SI Section S7). When a BMRB identifier maps to multiple PACSY chains, and no chain is specified, the longest chain is selected, with the entry identifier used to break ties.

Stage 2: validated SQL generation. Other database questions are translated into a single SQL statement by the language model. The model receives the user question together with a schema prompt of approximately 11,000 tokens. The prompt describes the available tables and columns, their relationships and permitted values, and includes 23 worked examples of natural-language questions and their corresponding SQL statements. It also contains rules for the

conventions of the database. For example, analyses that combine chemical shifts with per-model structural annotations use only model 1 when each residue should contribute once. Secondary-structure terms such as “helix” are translated into the appropriate STRIDE codes, and questions about a single BMRB entry are restricted to one chain. The generated SQL statement must pass the safeguards listed in Table 2 before and during execution. The main security control is execution through a database account with SELECT privileges only. Additional checks restrict the form of the statement, the tables that can be queried, the execution time, and the amount of data returned. These controls limit both modification of the database and expensive queries, but they do not guarantee that a valid query represents the scientific meaning intended by the user. Stored text returned by SQL, such as a deposited peak list, is returned verbatim. When a database query produces values for a plot, those numerical values are inserted directly into the plotting script rather than re-entered or recalculated by the model.

Table 2. Safeguards applied to model-generated SQL (SI Section S2.4).

Safeguard	Rule
Statement form	One statement; must begin with SELECT or WITH
Keyword blocklist	No INSERT, UPDATE, DELETE, DROP, ALTER, CREATE, TRUNCATE, GRANT, REVOKE, REPLACE, INTO OUTFILE/DUMPFILE, LOAD_FILE, LOCK TABLES, CALL, EXECUTE, SET or USE
Table allowlist	Every table after FROM or JOIN must be one of 54 documented tables and views, or one of the 100 per-residue legacy views
Default limit	LIMIT 200 appended when absent, except for a single-row aggregate
Least privilege	Executed as pacsyllm_ro, which holds SELECT privileges only
Timeout	45 s per statement
Repair round	A rejected or failed statement is returned to the model once with the error message; timeouts are not retried
Answer cap	At most 50 rows passed to the answer stage
Chain rule	Stated in the prompt; enforced in code only for Stage 1 outputs

Stage 3: document retrieval. If a question cannot be answered from the database, PACSYLLM searches 408 curated reference documents using TF–IDF similarity. These documents include POKY manual pages and PyMOL and Xplor-NIH command references (SI Section S7). A web search is used only when the curated documents do not cover the main terms in the question.

Stage 4: grounded answer generation. The language model generates the final response using only the database rows or retrieved passages provided to it. For drafts of 400 characters or more, a second model pass removes sentences containing specific factual claims that are not supported by the supplied material.

2.6. Evaluation

All questions were submitted to the production system. Three different sets were used. The first set, 26 held-out questions (SI Section S1.1), estimated accuracy on new questions. Each had a handwritten reference SQL statement. Each question was run three times. Generated SQL was scored automatically by comparing its returned values with those of the reference query. Numerical values were required to agree within half a unit of the final reported decimal place, and the number of returned rows had to match. The same 26 questions were then submitted three times each to the complete production pipeline. A final answer was considered numerically correct only if it contained every reference value. Each answer was also inspected for incorrect accompanying statements. The second set was the twelve user-manual questions (SI Section S1.3). Nine of them also appear as worked examples in the schema prompt, and the system was refined on this set. Their performance was used as a regression test rather than as an estimate of accuracy. The third set, six analyses from published PACSY studies (SI Section S4), tested whether research workflows could be reproduced from a single natural-language request. Each request was run through the deployed Stage 2 execution path, and the generated SQL, returned rows, and elapsed time were recorded. Plot requests were also run through the complete PACSYLLM pipeline (SI Sections S1.1 and S4). To test sensitivity to wording, the questions of all three sets were also posed in reworded forms with the same meaning, with each in a new conversation through the POKY client (SI Section S1.2).

All evaluation results were independently checked using handwritten SQL under the same read-only account and, where applicable, against expected line counts, residue counts, and stored RMSD values. The literature survey (Figure 1; SI Section S5) included all records in Europe PMC that cited Lee et al. (2012), with a keyword search for PLUQ and PLUQin. Because Europe PMC does not index every citing publication, the survey represents a lower bound on PACSY use. A publication was counted as a PACSY use only when its text describes actual use of PACSY or a PACSY-derived tool.

3. Results

3.1. The rebuilt PACSY database is larger and richer

Table 3 compares the rebuilt PACSY database with the 2012 release. The number of matched chains increased 1.6-fold, residues 2.4-fold, chemical shifts 3.9-fold, and atomic coordinates 3.1-fold. Of the 11,537 PDB/BMRB chain entries in the rebuilt database, 10,881 are protein chains with residue-level data, and 656 are nucleic-acid chains retained from the 2012 release. Most of the derived layers in Table 1 are new. These include TALOS-N, PECAN, and CheSPI predictions; LACS referencing corrections; well-ordered cores and ensemble RMSDs; conformer analysis; per-atom charges and radii; and peptide-bond and disulfide geometry. Fold classifications are available for 4,165 entries from SCOPe and 5,613 from CATH. Coverage varies among layers according to the information available for each deposition; coverage and exceptions are given in SI Sections S2.1 and S3.

Quality control confirms the stored data against their sources. Every protein entry has the oxygen and sulfur atoms of its deposited structure, except three entries whose depositions contain none, and the stored coordinates agree with the deposited mmCIF files (SI Section S3.5). Chemical shifts are paired with one BMRB entity per chain. This matters for complexes because 963 of the 7,735 checked BMRB entries (12%) contain shifts for more than one molecular entity or shift list. BMRB ID 17378, for example, contains a 105-residue protein and a six-residue bound peptide, and only the protein's shifts are stored for the protein chain (SI Section S3.1). The database contains 7,948,373 chemical shifts in 10,518 entries. No stored chemical shift is assigned to an atom absent from its residue type, and re-deriving the shifts from the available NMR-STAR files reproduces every stored value.

PACSY retains chemical shifts as deposited, but erroneous values present in the original source files can still appear in unfiltered statistics. The covalent-geometry layers agree with independent evidence. The coordinate-based cis/trans classification matches the wwPDB cis annotation, and the chemical shifts separate both states. The proline Cβ−Cγ difference is 9.34 ± 1.52 ppm for cis and 4.56 ± 1.51 ppm for trans prolines, and the mean cysteine Cβ shift is 41.00 ± 4.40 ppm for disulfide-bonded and 29.85 ± 4.37 ppm for other cysteines (SI Section S3.7).

Legacy views preserve the original table and column names, although queries that relied on the original numerical ordering of identifiers require a minor update (SI Section S4.4).

Table 3. Comparison of the original PACSY release and the rebuilt PACSY database. Counts are exact row counts of the current database.

Item	2012	2026	Ratio
PDB entries	3,745	8,891	2.4
BMRB entries	3,745	8,812	2.4
Matched chains (entries)	7,395	11,537	1.6
Residues	374,631	894,072	2.4
Chemical shifts	2,035,722	7,948,373	3.9
Per-model residue structure records	6,098,716	14,474,199	2.4
Atomic coordinates	75,899,756	238,583,538	3.1
SCOP(e) reference index	143,428	397,955	2.8
CATH/SCOPe entry classifications	—	9,778	new
Ensemble RMSD, per atom / residue / model	—	3,356,778 / 749,885 / 176,905	new
TALOS-N / PECAN / CheSPI residue predictions	—	853,348 / 854,022 / 845,109	new
LACS-corrected chemical shifts	—	7,844,339	new
Peptide-bond summaries (cis/trans)	—	882,678	new
Disulfide-bond records	—	5,026	new
DSSP residue summaries	—	893,621	new
Storage	5,639 MB	36,181 MB	≈6.4

3.2. Validation and accuracy of PACSYLLM

The held-out evaluation assessed PACSYLLM using 26 questions developed after the system and its schema prompt had been finalized. These questions do not appear in the prompt (SI Section S1.1). Each question has a handwritten reference query. With the full schema prompt, the current system generated correct SQL for all 26 questions in each of three repeats, with identical SQL in every repeat and a median time of 1.3 s (Table 4). The complete production pipeline returned the correct reference values in all 26 final answers, with identical answer text across repeats. Reworded versions of the 26 questions were also all answered correctly (Figure 4a; SI Section S1.2).

Table 4. Held-out evaluation of the current system (26 questions). The SQL stage alone and the final answers of the complete production pipeline were each evaluated in three repeats; the third row shows reworded versions of the same questions tested in two separate runs through the POKY client (SI Section S1.2).

Evaluated output	SQL correct	Reference value(s) correct	Identical in all three repeats	Median time
SQL stage	26/26 in every repeat	—	26/26 (SQL)	1.3 s
Complete pipeline (production)	—	26/26 in every repeat	26/26 (answer text)	3.3 s
Complete pipeline, reworded questions (POKY client)	—	26/26 in both runs	—	—

3.3. Representative user questions

The twelve questions from the PACSYLLM user manual illustrate the range of requests handled by the system (Table 5). They include four broad types of tasks, which are database statistics (questions 1–6), retrieval of stored content (question 7), deterministic outputs such as peak lists and input files (questions 8, 9, and 12), and analysis of structural ensembles (questions 10 and 11). Five questions use BMRB ID 17378, the protein–peptide complex described above, and therefore also test the handling of multiple chains. All twelve questions, and reworded versions of each, were answered correctly by the final system (Figure 4b; SI Section S1.2). Because these questions were used during development and nine also appear as worked examples in the schema prompt, their performance serves as a regression test of documented functionality rather than an independent measure of accuracy.

Stage 1 handles outputs whose content and chain selection can be defined deterministically, so peak lists, sequence files, and RMSD series are complete and belong to the intended chain. For BMRB ID 17378, the NOESY simulation produced a 1,453-line peak list in 0.56 s, the CYANA sequence contained the expected 105 residues, and the per-residue Cα RMSD output also contained all 105 residues (SI Figure S1). RMSD was lowest within the three Cyrange well-ordered segments and increased in loops and toward the disordered C-terminus, reaching 5.06 Å at Glu105.

Table 5. The 12 user-manual questions (verbatim). Stage, the stage that answers the question in the final system. All outputs were verified.

#	Question	Stage	Final verified output
1	What's the general chemical shift for CA for Alanine from IDPs?	2	matches direct SQL
2	What is the average CA chemical shift for isoleucine in a helix versus a coil?	2	matches direct SQL
3	Same, but using LACS reference correction.	2	matches direct SQL[a]
4	How many entries in the database are solid-state NMR?	2	matches direct SQL
5	What's the average helix propensity for tryptophan?	2	matches direct SQL
6	Break down entries by experiment method.	2	matches direct SQL
7	Can I get the NHSQC peak list for BMRB 6457?	2	stored peak list returned verbatim
8	Can you simulate the 2D NOESY peak list for BMRB 17378?	1	1,453-line peak list, chain A; 0.56 s
9	Give me the CYANA .seq file for BMRB 17378.	1	105 lines, Gly1 … Glu105
10	Which model of BMRB 17378 deviates most from the ensemble average?	2	model 8 of 25, 2.10 Å
11	Does BMRB 17378 show signs of having multiple conformers?	2	no signs (best cluster count 1)[b]
12	Can you plot the CA RMSD per residue for BMRB 17378?	1	105 residues, 0.29–5.06 Å, no model call (SI Figure S1)

[a] Asked as a follow-up in the same conversation as question 2. [b] Conformer criterion as in SI Section S3.6. The best split of chain A is significant ($p = 0.005$), but its clusters are only 0.18 Å further apart than models within a cluster.

3.4. Published PACSY analyses reproduced as natural-language questions

Six analyses from the original PACSY paper and later studies were reformulated as single natural-language requests to test whether PACSYLLM could reproduce published research workflows (Table 6; SI Section S4). With the current program, the database queries of all six returned the published or independently verified values. One step of the SSD-NMR workflow, the export of complete shift lists for spectrum simulation, lies outside the scope of an interactive query.

Table 6. Published PACSY analyses re-posed to PACSYLLM with the current program (SI Section S4). Times cover SQL generation and execution. The plot time is end-to-end.

Analysis (source)	Previous workflow	PACSYLLM result
Backbone ^{15}N statistics by secondary structure (Lee et al. 2012)	Python script over per-residue tables	Correct, 6.1 s. Ala helix/strand/coil 121.84/125.00/124.83 ppm (2012: 121.72/124.87/124.48)
Ala ^{1}Hα vs hydrophobicity scale (Lee et al. 2012)	Python script and plot	Correct, 17.6 s; runnable plot script with 23 bins
Misreferenced ^{13}C data (Fritzsching et al. 2016)	CSV export, local MySQL, custom analysis	Correct, 1.4 s. 8.01% of 9,442 entries have $\lvert\Delta C\alpha\rvert > 1$ ppm (LACS)
Ala ^{13}Cα assignment aid (Lee et al. 2012)	7 + 12 Analyzer fields, then a spreadsheet	Correct, 2.9 s. 54.38 ± 1.19 ppm (n = 99), edge = 'N'
Residue type from a Cα/Cβ pair (Fritzsching et al. 2013)	SQLite extract, Python, and MATLAB programs	Correct, 17.6 s. Leu ranked first (turn 271, coil 256, helix 231)

Analysis (source)	Previous workflow	PACSYLLM result
SSD-NMR benchmark (Li et al. 2024)	Own curation, spectra, and TALOS-N runs	Per-protein composition correct (3.7 s); curation 4,154 entries[a] (14.1 s); corpus-wide STRIDE/TALOS-N correlation r = 0.892 (29.6 s); spectrum simulation not attempted

[a] 4,155 when "20 residues" is read from the stored chain length rather than from the residue records; both readings are defensible.

PACSYLLM reproduced the classic PACSY analyses. Despite the larger database, the 2026 alanine ^{15}N means differed from the 2012 values by no more than 0.13 ppm for helix, strand, and turn and by no more than 0.35 ppm for coil and 3_{10} helix, small relative to the within-class standard deviations of 3.8–5.6 ppm (SI Figure S2a; SI Table S9). The relationship between alanine ^{1}Hα chemical shift and solvent exposure retained the opposing trends reported previously Mean ^{1}Hα increased from 3.90 to 4.18 ppm with exposure in helices and decreased from 4.98 to 4.73 ppm in β-strands (SI Figure S2b). Because the LACS offsets are stored, the referencing analysis of Fritzsching et al. (2016) became a single query; its 8.01% of entries with a Cα offset above 1 ppm is similar in scale to the approximately 7% reported previously, although the earlier study used a different estimator and database release.

The original PACSY practical example is the most demanding of these requests, because it combines a SCOP class, a sequence-length window, a pH window, model 1, and a helix-edge condition. When the request stated each condition explicitly, PACSYLLM applied every filter and returned 54.38 ± 1.19 ppm from 99 shifts in 13 PDB entries, including all six proteins of the 2012 analysis (SI Table S10). Requests that combine many conditions can benefit from clear and explicit wording. The example also shows why semantic validation matters. The helix-edge flag is stored as the letters Y and N, and a query that compares it with a number runs without error but removes the filter, returning a plausible mean of 54.33 ppm from 105 shifts (SI Section S4.4). A query can therefore satisfy syntactic and access-control checks while still misrepresenting the intended scientific condition.

The two remaining analyses involve corpus-scale queries. The PLUQ-style residue lookup, a self-join over approximately 7.9 million chemical shifts, completed within the 45 s execution limit. In the SSD-NMR workflow, the stored STRIDE and TALOS-N data replace the separate TALOS-N runs of Li et al. (2024), and the curation count and the database-wide correlation were answered directly (Table 6). The metal and paramagnetic criteria of the original curation could not be applied, because they are stored only as raw NMR-STAR text. Exporting every assigned ^{13}C shift list for spectrum simulation remains a bulk-export task that is better served by a dedicated program than by interactive model-generated SQL.

3.5. PACSYLLM in practice: investigating an unusual amide proton shift

Most backbone amide protons in proteins resonate between about 7 and 10 ppm, so a cross peak far upfield of this range is difficult both to assign and to interpret. The NMR structure of human myeloid-derived growth factor (hMYDGF; PDB 6O6W, BMRB 30584) provides a published example (Bortnov et al. 2019). The backbone amide proton of A157 resonates at 4.6 ppm, and an additional three-dimensional NOESY experiment was used to assign it. A researcher who observes such a peak may first ask whether similar shifts have been reported in other proteins. Using PACSYLLM, one can ask for every alanine backbone amide proton with a chemical shift between 4.0 and 5.0 ppm. PACSYLLM returned three residues, each with its PDB and BMRB identifiers, residue number, and chemical shift (Figure 5a; SI Section S6). They were Ala42 of the third fibronectin type II module of matrix metalloproteinase-2 (PDB 1J7M, BMRB 5012;

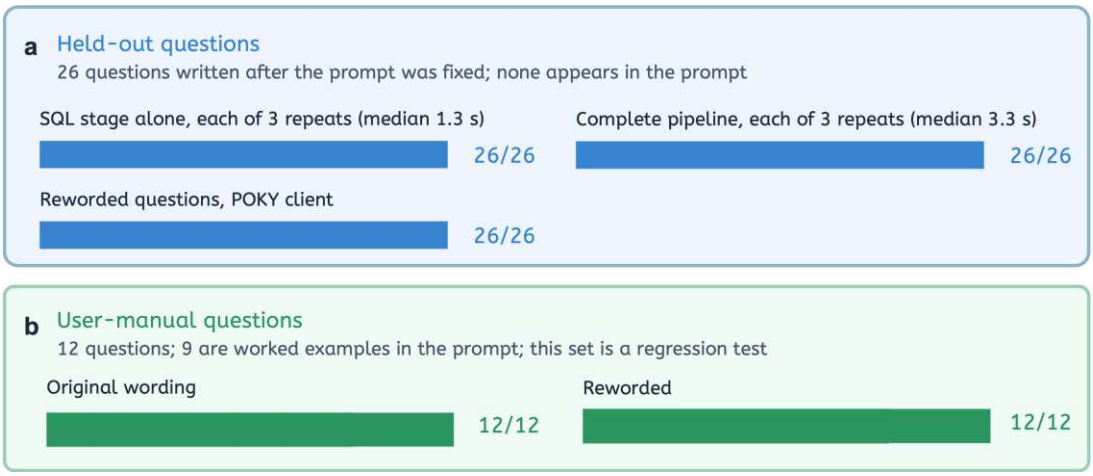


Figure 4. Evaluation of PACSYLLM on two question sets. (a) Held-out evaluation. All 26 questions were answered correctly by the SQL stage and by the complete production pipeline in each of three repeats. Reworded versions of all 26 questions were also answered correctly in two separate runs through the POKY client (Table 4; SI Sections S1.1 and S1.2). (b) User-manual regression test. All 12 documented example questions and their reworded versions were answered correctly. Because nine of these questions appear as worked examples in the schema prompt, this set tests preservation of documented functionality rather than independent accuracy (Table 5; SI Sections S1.2 and S1.3).

Briknarová et al. 2001) at 4.04 ppm, Ala37 of the G37A mutant of bovine pancreatic trypsin inhibitor (BPTI; PDB 1JV8, BMRB 5171; Battiste et al. 2002) at 4.36 ppm, and hMYDGF residue 131 at 4.60 ppm. A direct SQL query of the database returned the same three records, so the list was complete, and each value matches its deposited BMRB entry. Residue numbers follow the PDB author numbering, in which hMYDGF residue 131 is A157 of the original study.

Because each result carries its identifiers, the corresponding structures can be examined directly. In all three proteins, the alanine amide proton lies above the face of an aromatic ring, where ring currents shift nearby protons upfield (Figure 5b). In BPTI G37A, the amide proton lies 2.2–2.3 Å from the center of the Tyr35 ring, above the ring face in every deposited model, and the original authors described its large upfield shift as a ring-current shift arising from an interaction with the Tyr35 ring (Battiste et al. 2002). In 1J7M, the amide proton lies 2.2–2.8 Å from the center of the buried Phe19 ring (2.4 Å in model 1), again above the ring face in every deposited model.

a *Question: "List the PDB ID, BMRB ID, residue number and chemical shift of every alanine backbone amide proton (H) with a chemical shift between 4.0 and 5.0 ppm."*

PDB ID	BMRB ID	Residue number	Chemical shift (ppm)
1J7M	5012	42	4.04
1JV8	5171	37	4.36
6O6W	30584	131 (A157)	4.602

b

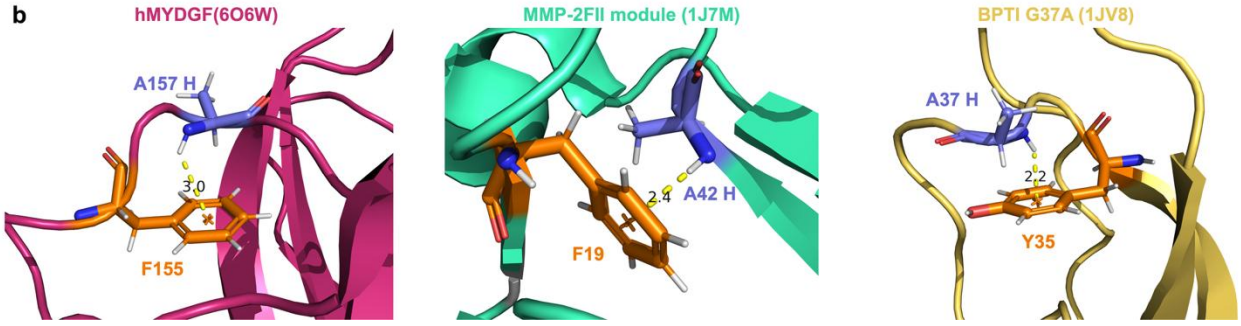


Figure 5. Using PACSYLLM to investigate an unusual amide proton shift. (a) The request and the three alanine backbone amide protons returned, with PDB and BMRB identifiers, residue numbers (PDB author numbering), and chemical shifts; a direct SQL query returned the same three records. hMYDGF residue 131 (PDB 6O6W, BMRB 30584) is A157 in Bortnov et al. (2019). (b) Model 1 of each structure, showing the alanine (blue), its amide proton (circle), the nearest aromatic side chain (orange) and the distance from the amide proton to the ring center (dashed line); nitrogen and oxygen atoms are dark blue and red. Left, hMYDGF A157 and F155; center, the MMP-2 fibronectin type II module, A42 and F19; right, BPTI G37A, A37 and Y35. The arrangement occurs in every deposited model of 1J7M and 1JV8 and in 9 of the 20 models of 6O6W. Except for BPTI G37A (Battiste et al. 2002), the ring-current interpretation is a structural observation of this work, not a conclusion of the original studies.

In hMYDGF, it lies 3.0 Å from the center of the F155 ring (PDB residue 129) in model 1, but this arrangement is present in only 9 of the 20 deposited models. For 1J7M and hMYDGF, the geometry is consistent with an aromatic ring-current contribution to the upfield shift; neither original study proposed a cause, and the present inspection does not establish one. In this workflow, PACSYLLM finds and organizes comparable experimental observations, whereas the physical interpretation rests on the structures and the literature.

4. Availability and use

PACSYLLM is accessed through the POKY AI Chat Assistant in the POKY software suite (two-letter code *AI*; Figure 6). Users submit questions in natural language through the chat interface, and PACSYLLM routes each request through the workflow described above. No separate installation of the PACSY database or language model is required by the user; these components run on the PACSY server. PACSYLLM is available in the latest version of POKY. Instructions and example questions are provided in the manual [https://poky.clas.ucdenver.edu/poky_releases/manual/poky_pacsyllm.html]. For researchers interested in developing their own applications similar to PACSYLLM, the enhanced PACSY

database is also publicly available as a standalone dataset through the Zenodo repository [https://doi.org/10.5281/zenodo.24407005].

5. Conclusions

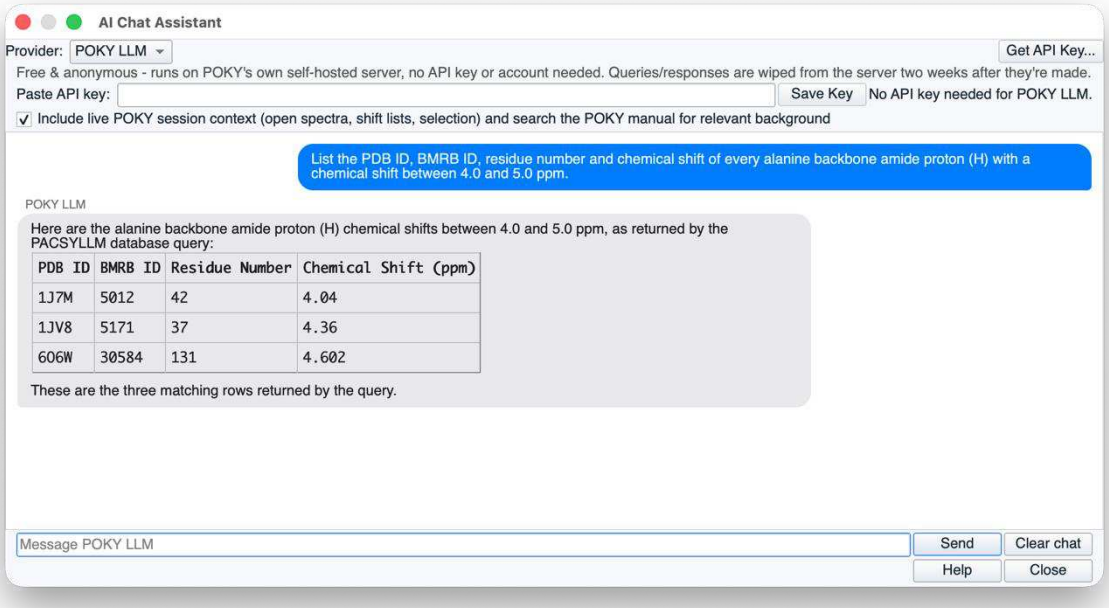


Figure 6. PACSYLLM is accessed through the POKY AI Chat Assistant in the POKY software suite (two-letter code *AI*).

The original PACSY integrated protein structures, chemical shifts, and fold classifications into a single searchable resource. The rebuilt PACSY expands this foundation with substantially more structural and chemical-shift data, new derived annotations, and updated SCOPe and CATH classifications. It also applies systematic quality control to link each chemical shift to the correct molecular ent ity and verify it against the source data. Together, these changes provide a larger and more consistent resource for studying relationships between protein structure and NMR chemical shifts.

PACSYLLM makes this information accessible through natural-language questions. A researcher can ask whether an unusual chemical shift has precedents, how a resonance compares with values observed in a particular structural environment, or how a property varies across proteins without first translating the question into database tables and SQL. In the current evaluation, PACSYLLM answered all 26 held-out questions correctly and also answered reworded versions of all 26 questions. More complex requests involving several simultaneous constraints were most reliable when those constraints were stated clearly and explicitly. The system is designed so that the language model does not replace established scientific computation. Deterministic programs handle well-defined tasks, while the model primarily translates natural-language questions into database queries and summarizes the returned information. This separation

allows natural-language access while keeping reproducible calculations and database controls outside the language model.

The present system nevertheless has important limitations. SQL validation can determine whether a generated query is safe and structurally valid, but it cannot guarantee that the query represents the intended scientific meaning. Clear question formulation therefore remains important, particularly for analyses involving multiple structural, experimental, or database constraints. Corpus-scale exports and specialized analyses are also better handled by dedicated deterministic workflows than by repeated natural-language queries. Finally, the present evaluation was based on questions developed by the authors and a single language-model configuration. Future work should include additional deterministic analysis tools, stronger semantic and type-aware validation, and evaluation with questions contributed by independent users. More broadly, PACSYLLM demonstrates how natural-language interfaces can make complex scientific databases more useful during data interpretation and analysis when language models are constrained to interpretation and communication, while scientific computation and data access remain reproducible and explicitly controlled.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

Author contributions

A.C. and W.L. writing–original draft, review & editing; conceptualization; database development and analysis; W.L. supervision & funding acquisition

Funding and additional information

This work was supported by the National Science Foundation [DBI-2413041 and DBI-2051595 to W.L.] and the University of Colorado Denver.

References

Battiste JL, Li R, Woodward C (2002) A Highly Destabilizing Mutation, G37A, of the Bovine Pancreatic Trypsin Inhibitor Retains the Average Native Conformation but Greatly Increases Local Flexibility. Biochemistry 41:2237–2245. https://doi.org/10.1021/bi011693e

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301-303. https://doi.org/10.1093/nar/gkl971

Bortnov V, Tonelli M, Lee W, et al (2019) Solution structure of human myeloid-derived growth factor suggests a conserved function in the endoplasmic reticulum. Nat Commun 10:5612. https://doi.org/10.1038/s41467-019-13577-5

Briknarová K, Gehrmann M, Bányai L, et al (2001) Gelatin-binding Region of Human Matrix Metalloproteinase-2. Journal of Biological Chemistry 276:27613–27621. https://doi.org/10.1074/jbc.M101105200

Burakova E, Vasa SK, Linser R (2022) Characterization of conformational heterogeneity via higher-dimensionality, proton-detected solid-state NMR. J Biomol NMR 76:197–212. https://doi.org/10.1007/s10858-022-00405-0

Chandonia J-M, Guan L, Lin S, et al (2022) SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. Nucleic Acids Research 50:D553–D559. https://doi.org/10.1093/nar/gkab1054

Chen VB, Arendall WB, Headd JJ, et al (2010) *MolProbity*: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12–21. https://doi.org/10.1107/S0907444909042073

Chiu A, Lee W (2026) Modernizing biomolecular NMR: The POKY suite. Journal of Biological Chemistry 302:111246. https://doi.org/10.1016/j.jbc.2026.111246

Codd EF (1970) A relational model of data for large shared data banks. Commun ACM 13:377–387. https://doi.org/10.1145/362384.362685

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302. https://doi.org/10.1023/A:1008392405740

Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Research 32:W665–W667. https://doi.org/10.1093/nar/gkh381

Dwarasala A, Rahimi M, Markley JL, Lee W (2022) ssPINE: Probabilistic Algorithm for Automated Chemical Shift Assignment of Solid-State NMR Data from Complex Protein Systems. Membranes 12:834. https://doi.org/10.3390/membranes12090834

Eghbalnia HR, Wang L, Bahrami A, et al (2005) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. Journal of Biomolecular NMR 32:71–81. https://doi.org/10.1007/s10858-005-5705-1

Emmanouilidis L, Bartalucci E, Kan Y, et al (2024) A solid beta-sheet structure is formed at the surface of FUS droplets during aging. Nat Chem Biol 20:1044–1052. https://doi.org/10.1038/s41589-024-01573-w

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579. https://doi.org/10.1002/prot.340230412

Fritzsching KJ, Hong M, Schmidt-Rohr K (2016) Conformationally selective multidimensional chemical shift ranges in proteins from a PACSY database purged using intrinsic quality criteria. J Biomol NMR 64:115–130. https://doi.org/10.1007/s10858-016-0013-5

Fritzsching KJ, Yang Y, Schmidt-Rohr K, Hong M (2013) Practical use of chemical shift databases for protein solid-state NMR: 2D chemical shift maps and amino-acid assignment with secondary-structure information. J Biomol NMR 56:155–167. https://doi.org/10.1007/s10858-013-9732-z

Güntert P, Buchner L (2015) Combined automated NOE assignment and structure calculation with CYANA. J Biomol NMR 62:453–471. https://doi.org/10.1007/s10858-015-9924-9

Hoch JC, Baskaran K, Burr H, et al (2022) Biological Magnetic Resonance Data Bank. Nucleic Acids Res 51:D368–D376. https://doi.org/10.1093/nar/gkac1050

Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Cryst A 32:922–923. https://doi.org/10.1107/S0567739476001873

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637. https://doi.org/10.1002/bip.360221211

Kirchner DK, Güntert P (2011) Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics 12:170. https://doi.org/10.1186/1471-2105-12-170

Kwon W, Li Z, Zhuang S, et al (2023) Efficient Memory Management for Large Language Model Serving with PagedAttention. In: Proceedings of the 29th Symposium on Operating Systems Principles. ACM, Koblenz Germany, pp 611–626

Lee W, Bahrami A, Dashti HT, et al (2019) I-PINE web server: an integrative probabilistic NMR assignment system for proteins. J Biomol NMR 73:213–222. https://doi.org/10.1007/s10858-019-00255-3

Lee W, Cornilescu G, Dashti H, et al (2016a) Integrative NMR for biomolecular research. Journal of Biomolecular NMR 64:307–332. https://doi.org/10.1007/s10858-016-0029-x

Lee W, Markley JL (2018) PINE-SPARKY.2 for automated NMR-based protein structure research. Bioinformatics 34:1586–1588. https://doi.org/10.1093/bioinformatics/btx785

Lee W, Petit CM, Cornilescu G, et al (2016b) The AUDANA algorithm for automated protein 3D structure determination from NMR NOE data. J Biomol NMR 65:51–57. https://doi.org/10.1007/s10858-016-0036-y

Lee W, Rahimi M, Lee Y, Chiu A (2021) POKY: a software suite for multidimensional NMR and 3D structure calculation of biomolecules. Bioinformatics 37:3041–3042. https://doi.org/10.1093/bioinformatics/btab180

Lee W, Stark JL, Markley JL (2014) PONDEROSA-C/S: client–server based software package for automated protein 3D structure determination. Journal of biomolecular NMR 60:73–75

Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics 31:1325–1327. https://doi.org/10.1093/bioinformatics/btu830

Lee W, Yu W, Kim S, et al (2012) PACSY, a relational database management system for protein structure and chemical shift analysis. Journal of Biomolecular NMR 54:169–179. https://doi.org/10.1007/s10858-012-9660-3

Li H, Tuttle MD, Zilm KW, Batista VS (2024) Rapid Quantification of Protein Secondary Structure Composition from a Single Unassigned 1D ^{13}C Nuclear Magnetic Resonance Spectrum. J Am Chem Soc 146:27542–27554. https://doi.org/10.1021/jacs.4c08300

Li J, Hui B, Qu G (2023) Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023)

Manthey I, Tonelli M, Clos li L, et al (2022) POKY software tools encapsulating assignment strategies for solution and solid-state protein NMR data. Journal of Structural Biology: X 6:100073. https://doi.org/10.1016/j.yjsbx.2022.100073

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247:536–540. https://doi.org/10.1016/S0022-2836(05)80134-2

Nielsen JT, Mulder FAA (2021) CheSPI: chemical shift secondary structure population inference. J Biomol NMR 75:273–291. https://doi.org/10.1007/s10858-021-00374-w

Pérez-Conesa S, Keeler EG, Zhang D, et al (2021) Informing NMR experiments with molecular dynamics simulations to characterize the dominant activated state of the KcsA ion channel. The Journal of Chemical Physics 154:165102. https://doi.org/10.1063/5.0040649

Pourreza M, Rafiei D (2023) DIN-SQL: decomposed in-context learning of text-to-SQL with self-correction. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023)

Pritchard RB, Hansen DF (2019) Characterising side chains in large proteins by protonless 13C-detected NMR spectroscopy. Nat Commun 10:1747. https://doi.org/10.1038/s41467-019-09743-4

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. Journal of Magnetic Resonance 160:65–73. https://doi.org/10.1016/S1090-7807(02)00014-9

Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. Journal of Biomolecular NMR 56:227–241. https://doi.org/10.1007/s10858-013-9741-y

Sillitoe I, Bordin N, Dawson N, et al (2021) CATH: increased structural coverage of functional space. Nucleic Acids Research 49:D266–D273. https://doi.org/10.1093/nar/gkaa1079

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. Journal of Molecular Biology 147:195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Somberg NH, Sučec I, Hong M (2025) Assignment-Free Determination of Ligand Binding Sites in Proteins by Solid-State NMR. J Am Chem Soc 147:26534–26545. https://doi.org/10.1021/jacs.5c06493

Ulrich EL, Akutsu H, Doreleijers JF, et al (2008) BioMagResBank. Nucleic Acids Research 36:D402–D408. https://doi.org/10.1093/nar/gkm957

Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. Journal of Biomolecular NMR 32:13–22. https://doi.org/10.1007/s10858-005-1717-0

Weiss C, Perrone B, Catone N, et al (2026) Co-sedimentation is the key to the structural investigation of wild-type FAT10. J Biomol NMR 80:17. https://doi.org/10.1007/s10858-026-00495-0

Wishart DS, Sykes BD (1994) The 13C Chemical-Shift Index: A simple method for the identification of protein secondary structure using 13C chemical-shift data. Journal of Biomolecular NMR 4:171–180. https://doi.org/10.1007/BF00175245

Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651. https://doi.org/10.1021/bi00121a010

Wojdyr M (2022) GEMMI: A library for structural biology. JOSS 7:4200. https://doi.org/10.21105/joss.04200

Yang Y, Fritzsching KJ, Hong M (2013) Resonance assignment of the NMR spectra of disordered proteins using a multi-objective non-dominated sorting genetic algorithm. J Biomol NMR 57:281–296. https://doi.org/10.1007/s10858-013-9788-9

Yokochi M, Kobayashi N, Ulrich EL, et al (2016) Publication of nuclear magnetic resonance experimental data with semantic web technology and the application thereof to biomedical research of proteins. J Biomed Semant 7:16. https://doi.org/10.1186/s13326-016-0057-1

Yu T, Zhang R, Yang K (2018) Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: Proc 2018 Conference on Empirical Methods in Natural Language Processing. pp 3911–3921