Received: 30 April 2021 | Revised: 15 July 2021 | Accepted: 19 August 2021

DOI: 10.1002/aic.17459

AIChE JOURNAL

FUTURES ISSUE: BIOMOLECULAR ENGINEERING, BIOENGINEERING, BIOCHEMICALS, BIOFUELS, AND FOOD

Data management schema design for effective nanoparticle formulation for neurotherapeutics

Hawley Helmbrecht[1] | Nuo Xu[1] | Rick Liao[1] | Elizabeth Nance[1,2,3,4]

[1]Chemical Engineering, University of Washington, Seattle, Washington, USA

[2]e-Science Institute, University of Washington, Seattle, Washington, USA

[3]Center for Human Development and Disability, University of Washington, Seattle, Washington, USA

[4]Department of Radiology, University of Washington, Seattle, Washington, USA

Correspondence
Elizabeth Nance, Chemical Engineering, University of Washington, Seattle, WA, USA.
Email: eanance@uw.edu

Funding information
Eunice Kennedy Shriver National Institute of Child Health and Human Development, Grant/Award Number: R21HD100639; National Institute of General Medical Sciences, Grant/Award Number: R35GM124677; National Science Foundation, Grant/Award Numbers: DGE-1633216, HDR1934292

Abstract

Translation of nanotherapeutics from preclinical research to clinical application is difficult due to the complex and dynamic interaction space between the nanotherapeutic and the brain environment. To improve translation, increased insight into nanoformulation-brain interactions in preclinical research is necessary. We developed a nanoformulation-brain database and wrote queries to connect the complex physical, chemical, and biological features of neurotherapeutics based on experimental data. We queried the database to select nanoformulations based on specific physical characteristics that enable effective penetration within the brain, including size, polydispersity index, and zeta potential. Additionally, we demonstrate the ability to query the database to return select nanoformulation characteristics, including nanoformulation methodology or methodological variables such as surfactant, polymer, drug loading, and sonication times. Finally, we show the capacity of our database to produce correlations relating nanoparticle formulation parameters to biological outcomes, including nanotherapeutic impact on cell viability in cultured brain slices.

KEYWORDS
database, nanoformulation, neurotherapeutics, polymer, query

1 | INTRODUCTION

Nanomedicine has achieved limited translation from preclinical research to clinical application for non-cancer neurological disease.[1] To improve translation, it is important to understand the dynamic interactions between nanomedicine platforms and the brain environment. Individually, nanomedicine and the brain environment are two independently complex entities: Nanomedicines are designed to control physical attributes that define the stability of the nanoformulation while allowing for drug incorporation and tailored drug release.[2,3] Nanomedicine physical and chemical characteristics such as size, charge, composition, and surface functionality impact interactions with cells, proteins, and extracellular components within the brain and ultimately the therapeutic effect.[4] Simultaneously, the brain microenvironment is heterogeneous from brain region to brain region, and dynamic during development, aging, and disease, which can affect nanomedicine delivery to the target destination.[5] The multitude of nanomedicine design parameters and neurobiological factors creates a large data space for which identification of key nanoparticle-brain interaction parameters is critical. To manage and query the large data space for nanoparticle-brain interactions, a nanoformulation database can assist in organizing and integrating key experimental variables that might influence the effectiveness of nanotherapeutics in the brain. The development of a nanoformulation database could improve understanding of nanoparticle-brain interactions and reduce bottlenecks in the preclinical to clinical nanotherapeutic translation pipeline.

Although there is limited or non-existent literature on effective database management for nanoformulations used preclinically, similar approaches for utilizing databases have existed for decades in computational cell biology, computational neuroscience, and clinical

applications.[6–9] For example, BioNumbers is a database that stores and organizes key quantitative features from cell biology.[10] However, the BioNumbers database and many others in the literature are composed mainly via natural language processing of published manuscripts and typically contain one specific value per property.

Rather than a database built from natural language processing of literature, experimental workflows are the foundation for the brain-nanoformulation database presented herein. Preclinical research databases need management schemas that logically connect nanoformulation experimental methodologies, which can be highly repetitive, vary in quality, are prone to rapid iteration and evolution, and are often research lab or facility-specific.[11] High repetition occurs because a statistically relevant experimental database must record duplicate experiments with identical or highly similar methodologies and possibly similar results.[12] Additionally, experimental data input may have incomplete information.[13] For example, a batch of particles may not have been characterized with every available methodology or even tested in a biological application.

A nanoformulation-brain database must also account for work from independent researchers with unique workflows.[14] Preclinical research in nanoformulation-biology interfaces is a dynamic endeavor; a successful database adapts to new methodologies and experiments so that independent researchers are free to follow expert-driven insight without being burdened by the database system.[15,16] Finally, an optimal nanoformulation-brain database connects biological outcome data to methodological data in an easily visualized way for assessing nanomedicine effectiveness and correlation to physical characteristics and methodological variables.

Therefore, in this study, we developed an entity-relationship diagram of neurotherapeutic research and then visualized and built a nanoformulation-brain database for preclinical experimental research. We loaded the database with experimental nanoformulation and biological application data, and then we developed hypothesis-driven queries that return insightful results from the organized data. Our queries serve four main roles: (1) return all nanoformulations in the database based on physical characteristics of the formulations, (2) return all nanoformulations in the database based on methodological information, (3) return all nanoformulations in the database from a specific researcher and with specific characteristics, and (4) return nanoformulation methodological variables, biological application information, and experiment information for drug screening applications. Finally, we visualized the results of each query to show the variety of insights gained by the database.

2 | MATERIALS AND METHODS

2.1 | Assessing and organizing data

We first compiled common laboratory procedures and data generated for formulating and testing nanotherapeutics to develop a process flow diagram. To begin the process flow diagram, we identified researchers that are nanoformulation experts in the lab ($n = 4$). Once we determined the researchers who would contribute data, we requested a copy of how they maintain overall formulation records and a basic description of their working methodology. All researchers provided independent .csv files of their formulation records and brief descriptions of workflow. From the .csv files, all variables, including methodological details, characterization results such as size and zeta potential (ZP), and researcher details such as name and education level, were pooled into a list. From the pooled list, we determined three main experimental variable categories: experimental set-up, nanoformulation methodology, and biological application with the researchers' provided descriptions of their working methodologies. The three main experimental categories became the main units of the nanoformulation-brain database.

Upon establishing a general unit-based structure for a flow diagram, we obtained data from any researcher in the lab who had performed a nanoformulation experiment, expanding our dataset from four researchers to 11. We received data in the same format as the original four independent researchers: a .csv file with all nanoformulations and related data and a text-based description of their workflow. The variables were extracted from each researchers' provided .csv files and sorted into a relevant category based on the provided descriptive workflow.

The three identified units, nanoformulations, biological application, and experimental components, became the main units of our process flow diagram. We then determined the major components of each unit by sorting common variables into specific experimental methodologies or protocols. For the nanoformulation unit, formulation methodologies from each researcher and characterization storage commonalities informed five main components: nanoprecipitation, single emulsion, double emulsion, reverse formulation, and nanoparticle characterization. The nanoparticle characterization components include data from two techniques, dynamic light scattering and drug loading assays.

The experimental unit contains information about researchers, collaborators, and a researcher-determined specific experiment. The information recorded about researchers includes first name, last name, a unique researcher id, and education level (e.g., graduate student, undergraduate student, high school student). Collaborator information includes first name, last name, a unique collaborator id, institution, and education/job level (e.g., faculty, research staff, graduate student, postdoctoral fellow).

Nanoformulations were evaluated for biological activity: nanoformulations were added to cultured organotypic brain slices and slices underwent a lactate dehydrogenase (LDH) activity assay. The data from the LDH assay and the information about the Sprague–Dawley rats used for producing brain slices are included within the ldh_assay table and the pup_info table accordingly. Information collected about the LDH assay includes a unique biological characterization id, an experiment id that correlates with the experiment table, a specimen id that correlates with the pup_info table, formulation id that correlates with the nanoformulation table, researcher id that correlates with an individual researcher in the researcher table, date of completed assay, a descriptive name for the treatment group, and results for timepoints 1 h, 2 h, 4 h, 8 h, and 1 day.

2.2 | Entity-relationship diagram

We used LucidChart, a web-based application for flowcharts and diagrams, to further organize our nanoformulation data from the process flow diagram into subcategories formatted into six tables: nanoformulations, double emulsion, single emulsion, nanoprecipitation, reverse formulation, and nanoparticle characterization tables (Figure 1). Then, connections between tables were determined for cardinality and ordinality, the maximum and minimum time that the row of one table can be related to the row of another table. Cardinality and ordinality decisions were based on laboratory practices and visualized in LucidChart with connecting lines between tables and with notation style for the appropriate cardinality and ordinality.

The developed entity-relationship diagram (Figure 1) connects 11 individual entity sets, tables of experimental information, and corresponding variables, across the experimental, biological application, and nanoformulation unit operations: researcher, collaborator, experiment, nanoformulations, single_emulsion. double_emulsion, nanoprecipitation, reverse_formulation, ldh_assay, pup_info, and np_charc (Table 1).

A three-column table represents each entity set. The three columns include the status of an attribute as a primary key (PK) or foreign key (FK), the attributes or variables of the table, and the data type of each attribute. There are also lines connecting each FK in an individual entity-set to the table that uses that FK as a PK. Every nanoformulation, researcher, collaborator, characterization result,

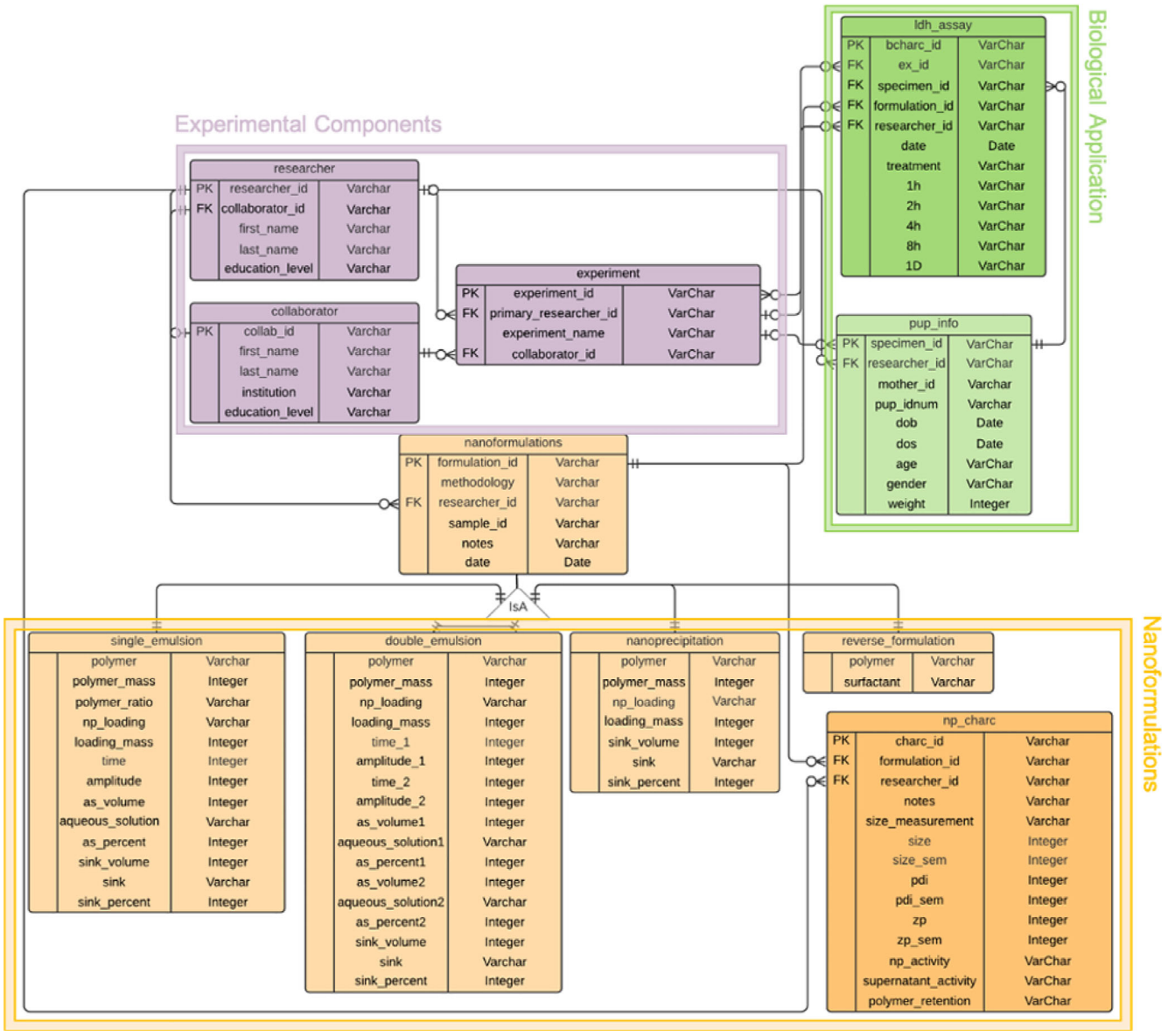

FIGURE 1 Entity-Relationship diagram for the database. The diagram relates experimental components with tables researcher, experiment, and collaborator (purple), biological components with tables pup_info and ldh_assay (green), nanoformulations components with nanoformulations (orange), including single_emulsion, double_emulsion, nanoprecipitation, reverse_formulations tables, and np_charc tables

TABLE 1 Entity-relationship diagram terms and definitions organized by the name of the table or entity-set, a description of the variables included within the table, a description of the connections to other tables, number of rows in the table, and file size of the table in the database

Table	Data description	Row count	File size
Researcher	A record of researchers, including their researcher id, names, and education level. A researcher may connect to zero or many formulations, characterizations, or overall experiments	11	336 bytes
Collaborator	A record of collaborators from the Nance Lab. A collaborator will have one and only one researcher id	1	110 bytes
Experiment	A record of experiments using formulated nanoparticles from the Nance Lab. May contain many nanoformulations and many biological specimens. Will only have one primary researcher	3	147 bytes
Nanoformulations	A record of all nanoformulations within the lab. A nanoformulation may be used in none, one, or many experiments. A nanoformulation may be characterized in none, one, or many ways, and characterization methods may be repeated for a specific nanoformulation	721	32 KB
Single_emulsion	A methodology for nanoformulation that includes only one emulsion step	171	16 KB
Double_emulsion	A methodology for nanoformulation that includes two emulsion steps	126	17 KB
Nanoprecipitation	A methodology for nanoformulation that uses solvent displacement for producing nanoparticles	392	29 KB
Reverse_formulation	A methodology for nanoformulation that allows higher control of specific physical features	31	2 KB
Ldh_assay	A record of characterization for biological specimens using catalase activity assay	88	8 KB
Pup_info	A record of characterization for littered animals used as biological specimens, including sex age, date of birth, date of sacrifice, and weight at sacrifice	9	434 bytes
Np_charc	A record of all nanoformulation characterizations, including dynamic light scattering and activity assays	717	53 KB

biological specimen, and biological characterization result is given at least one unique key. PKs signify that every key in that entity set is unique, whereas an FK for an entity set does not have to be unique. These lines end with cardinality and ordinality visualizations to symbolize the minimum and the maximum number of relationships each entity set can have with another.

Within the entity-relationship diagram, the experimental group contains the researcher, collaborator, and experiment tables. The experimental group connects developed nanoformulations, nanoformulation methodologies, and nanoformulation characterizations to the biological applications and their associated characterizations.

The nanoformulation group includes a table of basic information for every nanoformulation, specific information for the four main methodologies, and a characterization table with any characterization information from dynamic light scattering or drug loading assays. All of the information from the nanoformulation entity set for a specific nanoformulation is also included within the specific methodology table. Although this introduces some redundancy in the database, we decided to have separate tables to ease data standardization and querying. Each methodology has different, commonly manipulated variables for developing nanoformulations making it easier to standardize data if each methodology has an independent table. Additionally, for some queries, we only want basic nanoformulation information or only characterization information without all of the methodological variables. It is more efficient to search all nanoformulations in the basic nanoformulations entity set than through each specific methodology. Finally, the biological group contains two entity relationships that connect info about the rat pups for biological applications associated with specific brain slices or serum. These specimens then connect to the LDH assay variables and results.

2.3 | Data standardization and cleaning of database input

During the creation of the entity-relationship diagram, laboratory data informed the attributes for each entity set. All data were obtained in raw form and standardized for use in the database and to resolve downstream issues during querying. Data was tagged with researcher information and organized by nanoformulation methodology. For each methodology identified, all variables from the process flow diagram were listed and organized into table columns by order in which the variable occurs in the methodology. For example, polymer weight for dissolution in the organic phase is measured before measuring the volume of surfactant used in the aqueous sink condition, and therefore polymer weight occurs before surfactant volume in the table. Naming styles for individual samples were converted into a common name, and variables were standardized to have the same naming convention.

Additionally, attributes that contained semi-structured data, such as experiment notes, were cleaned of commas. Commas are the chosen delineator for .csv files imported into our database and cannot exist within variable entries. Once all of the data was cleaned, each table was loaded as an independent sheet in an Excel workbook and uploaded to Google Cloud for ease of access by multiple independent researchers.

2.4 | Local computer-based database

After the entity-relationship diagram was completed in LucidChart, we exported the schema as PostgreSQL commands to a local computer, with a 2.4 GHz Quad-Core Intel Core i5 processor and 8GB

2133 MHz LPDDR3 memory. PostgreSQL was installed using HomeBrew and accessed through the terminal. The database schema commands were imported from LucidChart, and a .csv for every individual entity set was imported into the database using the COPY functions of PostgreSQL with the top row of the .csv delineating the header for assigning columns to attributes of the entity set. The entire database is about 200 kilobytes. All developed queries were run through PostgreSQL on this created database. We turned on timing and ran each query 13 times following the leave-one-out rule to obtain average times for each query to run ($n = 12$). The resulting .csv files from each query are saved onto the same local computer using additional COPY functionality from PostgreSQL. The SQL code for creating the database is accessible on GitHub at: https://github.com/Nance-Lab/nanoformulations-database and specifically the file nancelabSchema.sql.

2.5 | Snowflake-based cloud database

We used the same PostgreSQL commands exported from LucidChart for developing the local database to also build a cloud-based database using Snowflake. In Snowflake, we created a database for the project. We copied all of our PostgreSQL commands into a worksheet and activated a size "X-small" data warehouse with 10-min auto-suspension for running our commands. We then imported all of the CSV data through the "load table" functionality within the Databases>"database name">"table name" window. To upload the data, the source file was selected from the local computer, and we created a custom file format for .csv files. The custom file format specifies using commas as column separators, a new line as row separators, and one line of a header to skip. With all data successfully uploaded to the database, we adjusted the WHERE statements of our developed queries to the Snowflake paths of each table and copied the queries into a worksheet. For evaluation, all queries were run 13 times following the leave one out methodology and leaving out the first run for a total of 12 points to evaluate the average time. The resulting .csv files from each query were downloaded onto the computer using the manual "Download or View Results" button in Worksheets.

2.6 | Statistics

Pairwise correlation of columns from the database was calculated using a Python package, Pandas, DataFrame.corr functionality. Pearson was input as the correlation method.

3 | RESULTS AND DISCUSSION

3.1 | Database evaluation

We evaluated the database for the six V's of "Big" Data, volume, variety, velocity, value, variability, and veracity,[17] to show the complexity of the data (Figure 2). Although the volume of the data is relatively small (Figure 2A), with about 160 kilobytes of data used for the results of this publication, there are other features of the data that support the treatment of our brain-nanoformulation data as a "big" data set.[18] First, there is a wide variety of data types, including structured, semi-structured, unstructured, and temporal data (Figure 2B). Regarding velocity, independent researchers update the data multiple times a week to incorporate new experiments and characterization techniques (Figure 2C). The value of the data is determined by scientific insight gained, animal lives used, and the amount of time for data collection (Figure 2D). For our nanoformulation-brain database, the nanoformulation data were collected over 6 years while the biological data were collected from nine different Sprague–Dawley rat pups over the course of 4 months and provides multi-faceted insight into nanoformulation features, biological features, and the interactions between nanoformulations and the brain. Variability for this database is increased because nanoformulations can be applied to many different brain environments that change according to treatment or injury, sex, age, and brain region (Figure 2E).[19] Additionally, veracity with neuroscience data is important both ethically and experimentally and is identified for this study in Figure 2F. Biological data has many ethical considerations, including minimization of life used, sampling bias, and perception bias.[20] Finally, experimental data can have errors and noise due to natural biological variability, human error, data obtained by an independent researcher during a period of training or method optimization, and input or databasing errors by how the data was added to the database or queried.[20]

In addition to evaluating the facets of "big" data, we tested the database both locally and with a cloud-based solution. The decision to host a database locally or in the cloud depends on multiple factors, including accessibility, economics, and time. This experimental database is approximately 160 KB and 2300 rows of data. While this is a significant amount of information to the lab collected over 6 years and from 11 independent researchers, volume-wise it is considered a small data set in database management. Cloud-based services offer access to multiple computers and locations at a higher economic cost than a local database. However, the price should be justified by a significant increase in performance. We tested our local database against the use of Snowflake. Snowflake is a popular, cloud-based data warehousing application with an online platform.[21] With large enough data sets, Snowflake offers an elastic, scalable, and secure system that significantly improves performance. However, with the size of the dataset used here, the Snowflake cloud platform took on average 10 times longer to return results than a local database (Figure 2G). The larger time per query run is likely because Snowflake was designed for large volume data from mainly transactional sources. Therefore, for experimental databases of this size, local hosting is a viable and efficient option. Accessibility can be improved by hosting the .csv files or spreadsheets in a cloud platform such as Google Cloud, with occasional imports to a local database.

3.2 | Hypothesis-driven query development

One specific goal of the database is to organize our nanoformulation data so that results are readily accessible to independent researchers

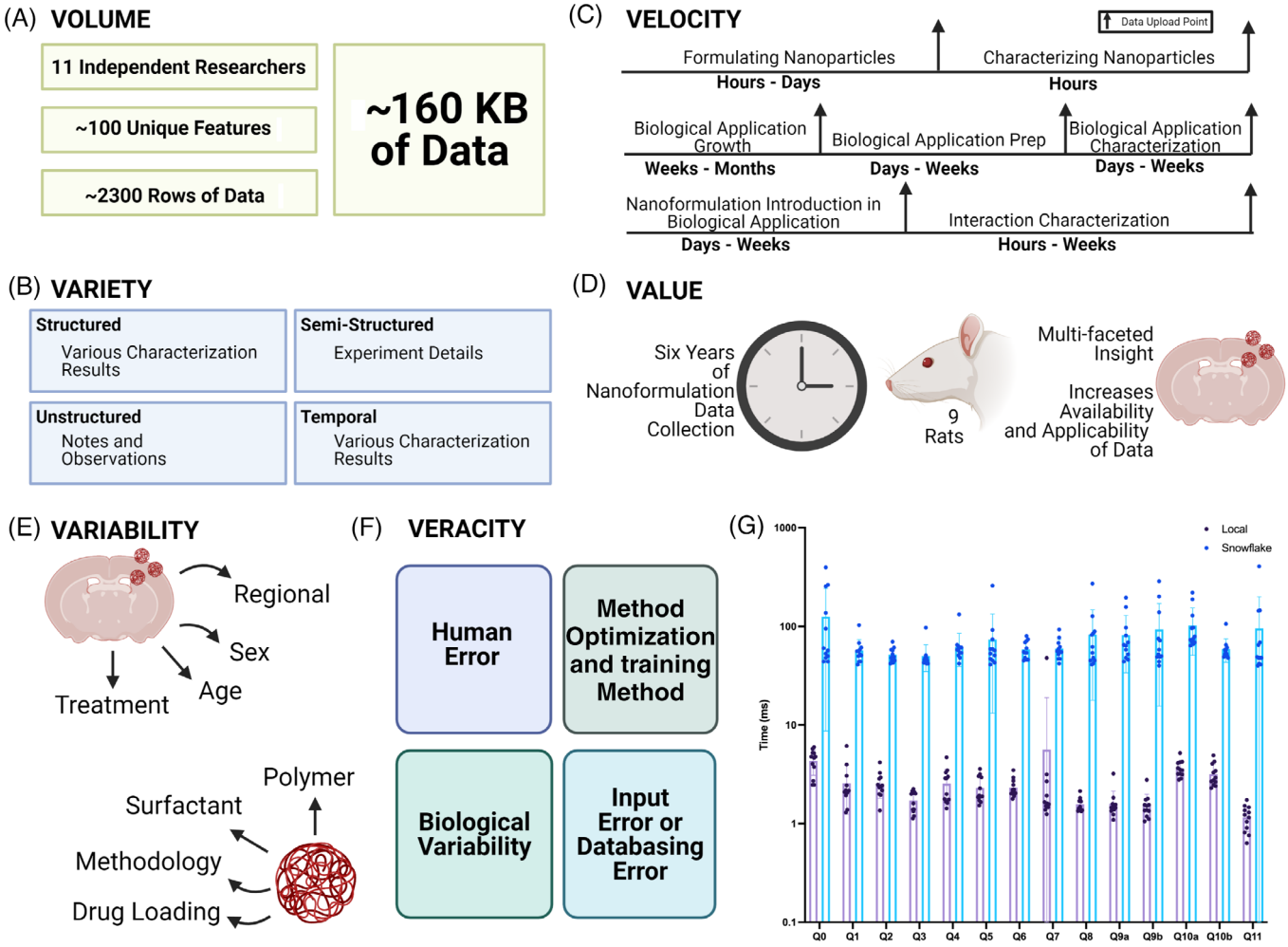

FIGURE 2 Database evaluation according to the six V's of big data. (A) Volume describes the number of independent researchers, features, rows, and quantity of data. (B) Variety describes the four major types of data with examples. (C) Velocity has example timelines with upload points for data and time scale values bolded below each major methodology step. (D) Value describes the time, life, and insight value of the database. (E) Variability shows the biological factors that affect the fate of nanoparticles and the methodological factors that affect the features of the nanoparticle. (F) Veracity describes the complexity of errors that can be introduced to the database. (G) Time for queries to run on a local server compared to Snowflake. Created with BioRender.com

using the database. To increase data accessibility for nanoformulation researchers, data must be searchable from various viewpoints, including biological outcome, nanoformulation methodological variables, and nanomedicine physical characteristics. The flexible searchability of a database is made possible through meaningful query development. The first step in meaningful query development was determining relevant results that researchers would consistently want from the database. Relevant results were determined by discussing the current hypotheses researchers have about their data but cannot answer efficiently without large data reorganization or increased experimentation. The variables the researchers needed to study to answer their hypotheses were recorded. With all variables for specific hypotheses outlined, we wrote queries that accessed the different units of the database and returned all specified hypotheses-related variables and data.

Our queries amplified the capabilities of the database by (1) writing a series of progressive queries (Queries 1–7) that result in all nanoformulations that have increasingly specific and scientifically informed constraints, (2) writing queries 8, 9a, and 9b for an experiment focused on nanoprecipitation effects on surface charge, and that can return nanoformulations made with specific polymers with or without drug loading, and (3) writing queries 10a, 10b, and 11 to test the database as a drug screening platform for the double emulsion and nanoprecipitation methodologies (Table 2). Each query has three parts: a SELECT, FROM, and WHERE portion. The SELECT portion outlines all variables that should be returned by the query. The FROM portion states all the tables that must be accessed for the information (Figure 3). The WHERE portion applies constraints and relationships between the tables. Queries for the local and Snowflake-based database are available on Github at: https://github.com/Nance-Lab/

TABLE 2 Descriptions of every query with the query number, query description, and the number of rows

Query	Description	Resulting number of rows
Query 0	Selects all nanoformulation, their characterization results, and related researcher information	704
Query 1	Selects all nanoformulations, their characterization results, and related researcher information with a size between 50 and 100 nm	334
Query 2	Selects all nanoformulations, their characterization results, and related researcher information with a size between 50 and 100 nm AND a ZP between −10 and 10 mV	262
Query 3	Selects all nanoformulations, their characterization results, and related researcher information with a size between 50 and 100 nm AND a ZP between −10 and 10 mV AND a PDI between 0 and 0.2	150
Query 4	Selects all nanoformulations made via single emulsion, their characterization results, their formulation methodologies, and related researcher information with a size between 50 and 100 nm AND a ZP between −10 and 10 mV AND a PDI between 0 and 0.2	4
Query 5	Selects all nanoformulations made via double emulsion, their characterization results, their formulation methodologies, and related researcher information with a size between 50 and 100 nm AND a ZP between −10 and 10 mV AND a PDI between 0 and 0.2	21
Query 6	Selects all nanoformulations made via nanoprecipitation, their characterization results, their formulation methodologies, and related researcher information with a size between 50 and 100 nm AND a ZP between −10 and 10 mV AND a PDI between 0 and 0.2	111
Query 7	Selects all nanoformulations made via reverse formulation, their characterization results, their formulation methodologies, and related researcher information with a size between 50 and 100 nm AND a ZP between −10 and 10 mV AND a PDI between 0 and 0.2	13
Query 8	Selects all nanoprecipitation nanoformulation methodologies from a specific researcher, based on researcher first name, that use '45 k PLGA' as the polymer and without loaded drug (Figure 3B)	10
Query 9a	Selects all nanoprecipitation nanoformulation methodologies from a specific researcher, based on researcher first name, that use 'P80' (Figure 3C)	6
Query 9b	Selects all nanoprecipitation nanoformulation methodologies from a specific researcher, based on researcher first name, that use 'DI Water' (Figure 3C)	6
Query 10a	Selects all nanoprecipitation formulations tested in slices with a related lactate dehydrogenase cytotoxicity assay completed from a specific publication along with the formulation methodologies, animal information, and researcher information (Figure 4A)	6
Query 10b	Selects all double emulsion formulations tested in slices with a related lactate dehydrogenase cytotoxicity assay completed from a specific publication along with the formulation methodologies, animal information, and researcher information (Figure 4B)	6
Query 11	Selects all double emulsion information that specifies both sonication time and nanoparticle activity characterization results along with the formulation methodology's variables and related researcher information	31

Notes: The description column provides a general description of all variables returned or "selected" by each query, which tables the variables were selected from, and any constraints. The resulting number of rows the total number of results retrieved by each query.
Abbreviation: PDI, polydispersity.

nanoformulations-database under files nanoformqueries_local.sql and nanoformqueries_snowflake.sql respectively.

3.3 | Querying nanoformulations for physical feature and methodological variables

The first main application of our database is to sort and query our nanoformulation data based on physical nanoformulation characteristics. To show the effectiveness of our database, we began by querying our nanoformulations without regard to biological information (Figure 4). Query 0 provides a baseline for the range of sizes, ZPs, and PDI values associated with the entire set of ~700 nanoformulations included within our database. The results from query 0 include 704 nanoformulations with sizes ranging from 1.7 to 1290 nm, ZPs ranging from −96.6 to 7.69 mV, and PDIs ranging from 0.01 to 0.96. Query 1 successfully constrains the list of nanoformulations from query 0 to 334 nanoformulations with sizes ranging from 50.42 to 99.89 nm. Query 2 further constrains the results to 262 nanoformulations with a size range from 50.42 to 99.47 nm and a ZP range from −10 to 7.69 mV. While query 3 imposes a third constraint on PDI, the query returns 150 results with a size range from 50.78 to 99.47 nm, ZP range of −10 to 0.47 mV, and a PDI range of 0.02–0.20.

Queries 0 through 3 show the ability of the database to select for specific features of formulated nanoformulations (Figure 4A). Being able to constrain nanoformulations based on size, ZP, and PDI is an important feature for experiments in probing and treating neurological diseases. Nanoparticle size and surface charge can influence nanoparticle passage across the blood–brain barrier and penetration within

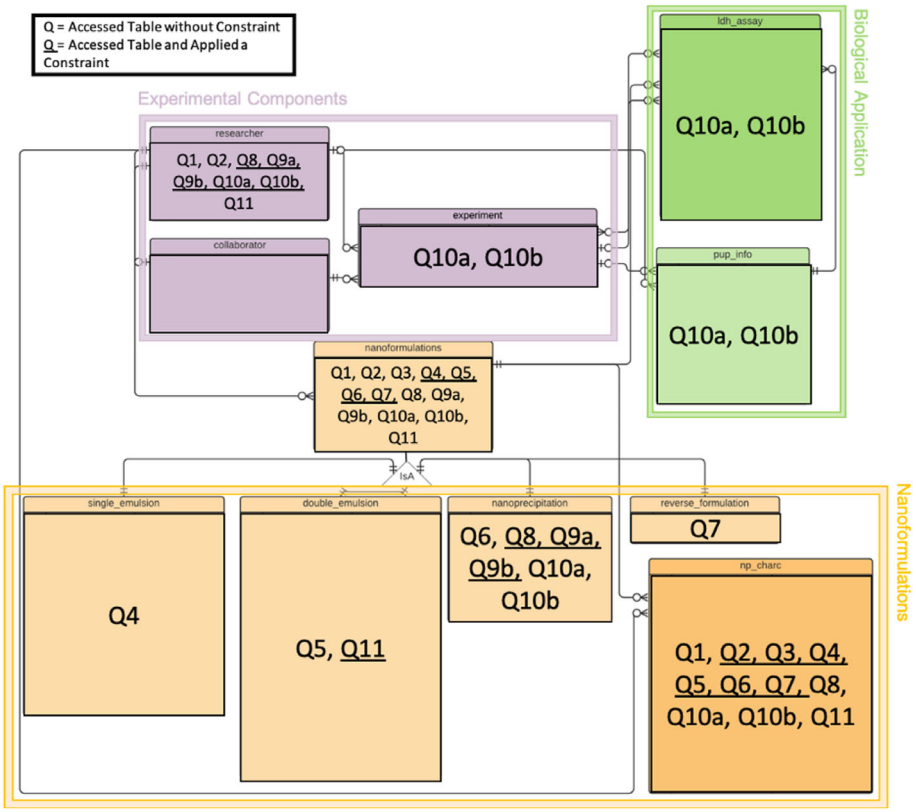

FIGURE 3 Entity-relationship diagram with query labels. The query numbers are included in every table they access. Underlined query labels designate that the query both accessed that table and placed some constraint on the results from that table

the brain parenchyma.[5] In comparison, PDI is an indicator of particle size uniformity. Each of these features provides key insight to the nanoformulations we have access to that are stable, uniform, and can theoretically transport through the brain to areas of interest.

We additionally are interested in querying for physical features of databased nanoformulations so that we may obtain the methodologies and methodological variables that obtained specific physical characteristics. We tested this application by developing queries 4 through 7, which applied the same constraints as query 3 and controlled for specific methodologies: single emulsion, double emulsion, nanoprecipitation, and reverse formulation (Figure 4A). We found the largest number of viable formulations with the query that specifically returned nanoprecipitation formulations followed by double emulsion, reverse formulation, and single emulsion methodologies. With the database, we can search for specific characteristics and then pick one of the four methodologies to get all variables that produced those nanoparticles. Each methodology has its strengths and weaknesses that allow for particular tailoring of nanoparticles for transport capacities or therapeutic loading capacities. Researchers can now efficiently find previously formulated and characterized nanoparticles and use them as a starting point for specific tailoring and further optimization.

To test the applicability of our database to nanoformulation analysis, we developed queries 8, 9a, and 9b. We developed query 8 to study the surface charge effects of surfactant on poly(lactic-*co*-glycolic acid) (PLGA) nanoparticles without drug encapsulated. All resulting formulations obtained through query 8 are nanoprecipitation methodologies. The query results show the effect of surfactant chosen for the nanoprecipitation methodology on the nanoformulations' sizes and ZPs (Figure 3B). Experimentally, we have shown CHA, F68, and PVA produce highly negative, slightly negative, and neutral nanoparticles, respectively (Figure 4B). However, to compare the effect of surfactant on nanoparticle surface presentation and subsequent nanoparticle interactions in the brain, nanoparticles made with these surfactants should have comparable hydrodynamic diameters, preferably within a range of ±10 nm in diameter (Figure 4B). Query 8 allowed us to visualize these results and determine the mass of polymer to use for each formulation to obtain a standard nanoformulation size with large ZP variation.

Queries 9a and 9b were written to find all formulations of a specific methodology, nanoprecipitation, that use surfactants and show the physical characteristics of nanoformulations made with different polymers. We showed physical characteristics of nanoformulations from two PLGA polymers—one copolymerized with PEG and the other without PEG—formulated with a specific surfactant polysorbate 80 (P80) while also specifying the presence or absence of a specific drug loaded into the nanoparticle (Figure 4C). The results from specific polymers, surfactants, and drug loading demonstrate the searchability of querying the database. Queries 9a and 9b can be efficiently changed to include different variable names or encompass fewer variables by quickly altering the strings in the WHERE clause. The hypothesis-driven queries leave the database exploration to the researcher's expertise without burdening the researcher with complex and time-consuming data science tasks.

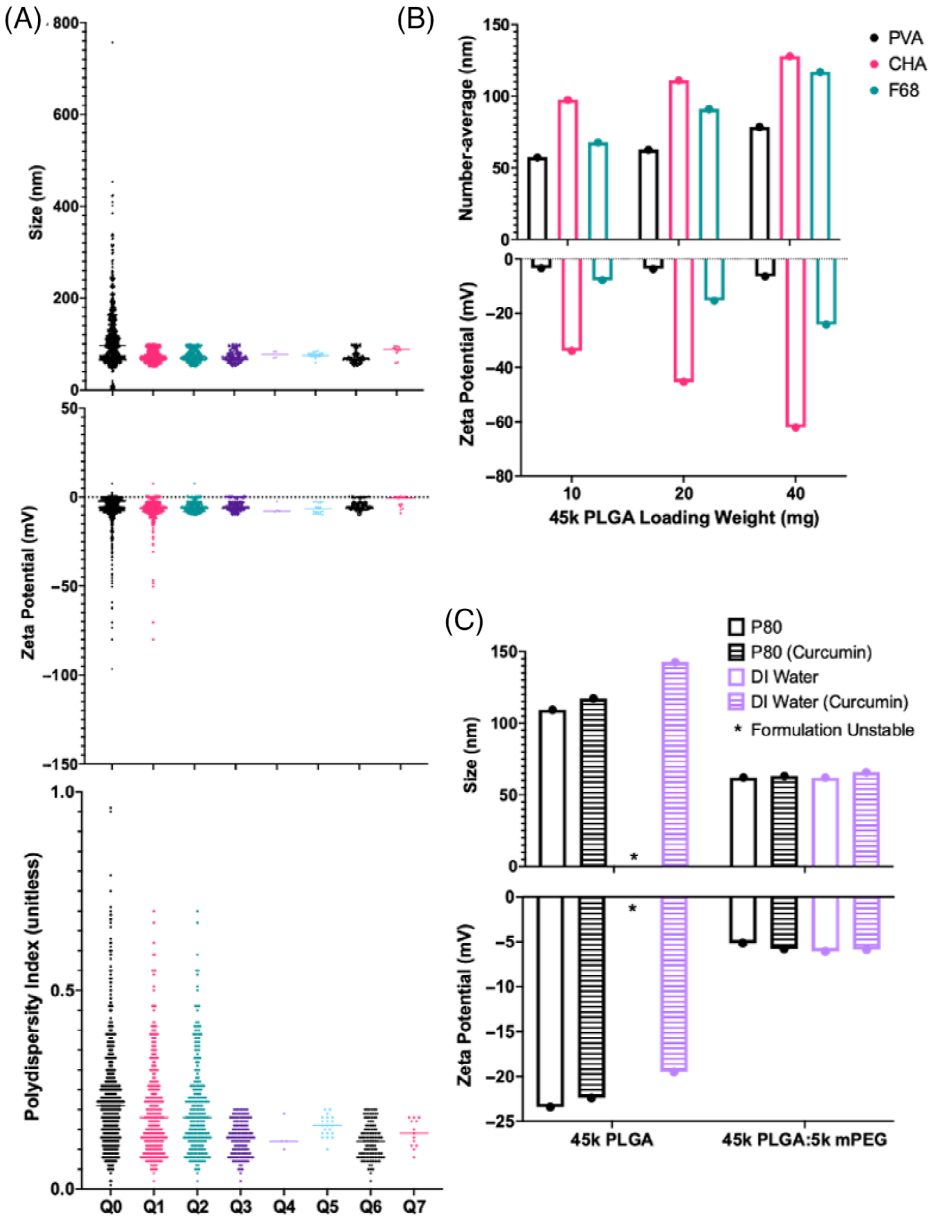

FIGURE 4 (A) Nanoformulation characteristic values for PDI, ZP, and size for the first seven queries developed. (B) Number-average size and ZP for nanoprecipitation formulations obtained with query 8 for varying surfactants: cholic acid sodium salt (CHA), polyvinyl alcohol (PVA), Pluronic F68 (F68), and varying polymer amounts. (C) Number-average size and ZP for two different polymer types with 1% P80 or DI Water as the nanoprecipitation sink

3.4 | Applying the database to enhance drug screening capabilities

The goal of designing a nanoformulations database for probing and treating neurological disease is to improve the screening of nanotherapeutics by providing insight between formulation variables, nanoformulation characteristics, biological applications, and visualizing the connections between them. Organotypic whole-hemisphere (OWH) brain slice models have been developed as high-throughput screening methods for nanoformulations.[22,23] Partnering OWH brain slice models with a nanoformulations database improves the connection between nanoformulation characterization and biological outcome.

We developed queries 10a, 10b, and 11 to obtain all nanoformulations from the database that were in OWH brain slices for an individual biocompatibility experiment (Figure 5). Queries 10a and 10b returned the nanoformulation methodological variables from the nanoprecipitation (Figure 5A) and double emulsion (Figure 5B) methods. The nanoprecipitation methodology has a weak negative correlation between the three physical characteristics—size, ZP, and PDI—of the nanoformulations and the LDH assay 1-, 2-, 4-, and 8-h times ranging from −0.27 to −0.38 (Figure 5A). The nanoprecipitation methodology has a moderate negative correlation between the nanoparticle physical characteristics—size, ZP, and PDI—for the 1-day LDH assay. In comparison, the double emulsion methodology shows increasing correlation strength between size, PDI, and ZP, as time for the LDH assay increases: a very weak correlation at 1-h LDH of −0.09, a strong correlation at 2-hours LDH of 0.63, and a very strong correlation at 4-h, 8-h, and 1-day LDH of 0.91, 0.94, and 0.95, respectively (Figure 5B).

The correlation strength differences between the nanoprecipitation methodology and double emulsion methodology highlight the complex effects of nanoformulation methodology on biological outcome. Meanwhile, the heatmaps show a visual representation of the capabilities of

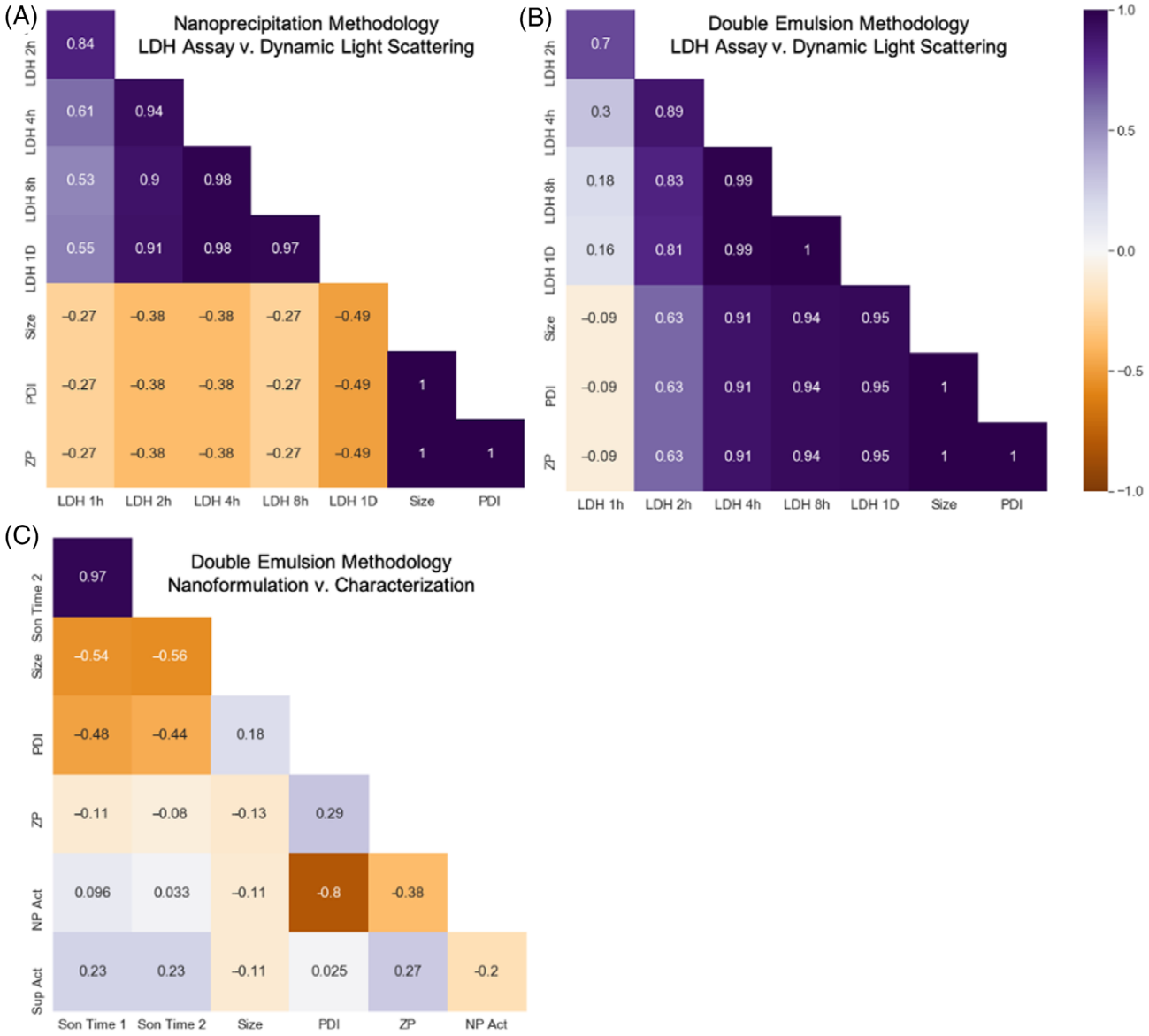

FIGURE 5 Heatmap of nanoprecipitation methodology variables, dynamic light scattering characterization results, and LDH assay results for (A) nanoprecipitation and (B) double emulsion experiments in brain slices with Pearson correlation as the colormap and explicitly stated in each box. (C) Heatmap of nanoformulation methodology data for double emulsions with varying sonication times versus dynamic light scattering characterization data, nanoparticle activity (NP Act), and supernatant activity (Sup Act) with color associated with Pearson correlation as colormap and explicitly stated in each box

the database for studying nanoformulation and biological outcome interaction. Specifically, the database enables a quick method to determine which methodological variables have the largest effect on a specific biological outcome or characterization. The effectiveness of the database enables researchers to tailor and alter nanoparticle-based nanoformulation-brain interactions efficiently.

Additionally, we used query 11 with the database to study the correlation between nanoparticle methodology, nanoparticle characteristics, and nanoparticle drug activity (Figure 5C). We found a strong negative correlation of −0.8 between PDI for double emulsion nanoparticles and nanoparticle activity. Additionally, there is a weak to moderate correlation between ZP and the nanoparticle activity of −0.38. We also found moderately negative correlations between sonication time, size, and PDI ranging from −0.44 to −0.56. The correlation between sonication time and nanoparticle characteristics shows that the choice of sonication time does affect the size and homogeneity of the samples without affecting their charge.

Interestingly, while there is a moderate correlation between the sonication time and both PDI and size, and a strong correlation between PDI and nanoparticle activity, there is no correlation between sonication time and nanoparticle activity. These results highlight the complexity of nanoformulation-brain interactions. The nanoformulation methodologies affect nanoparticle characteristics in a way that can be quantified by queries developed for the database. However, the relationship between the nanoformulation physical variables and nanoparticle activity is still difficult to understand. The

nanoformulation-brain database provides further insight into the complexity of the data via flexible queries such as queries 10a, 10b, and 11. The relationships between nanoparticle characteristics and biological outcomes quantified by the database allow researchers to select already developed nanoformulations with desired outcome in previously studied biological models for application in new models or alternative species. Additionally, the researcher may select a desired biological outcome for a specific nanoparticle batch and trace it back to the methodological variables that produced the nanotherapeutic to enhance or optimize the current outcome. For example, a researcher may determine that a specific nanoparticle batch produces the desired response from an LDH assay, so the researcher then reproduces the original nanoparticle batch while slowly altering ZP or size to determine an optimal range for nanoparticle characteristics of a specific methodology that still produce the desired outcome.

Utilizing a database for nanotherapeutic development and biological application improves data connectively across diverse and variable sets within academic laboratories. To successfully build and use formulation databases, traditionally experimental wet-labs will either need to commit to learning database fundamentals or outsourcing database development and maintenance. Fortunately, developing and building databases is a standard computer science and data science practice, and ample information is available for free or paid online for mastering SQL and database development within a few months. Once an experimental lab has database development knowledge, each lab can assess and connect their own methodologies through entity-relationship diagram design as we did with LucidChart or with similar visual graphic software. From there, laboratories partake in cleaning, standardizing, and implementing data upload practices for their own database according to the specifics of the data. The created database with SQL abilities enables each lab to use their scientific expertise to connect biological variables of interest with methodological variables that the lab can control or manipulate.

Alongside increasing data connectivity by improving the searchability of interconnected variables from biological and formulation methodologies, the database also highlights experimental research gaps. For example, the relationship between nanoformulation methodology and LDH assay differs for nanoparticles with similar characteristics such as neutral charge and sub 114 nm, but that are made via double emulsion or nanoprecipitation. Size, ZP, and PDI with DLS are not sufficient characterizations to fully elucidate the relationship between the chemistry or material composition of the nanoformulations with different formulation methodologies and resulting different biological outcomes. In this case, the database has helped identify a needed area of experimental exploration. With additional build-out of the database to include more features and data, increased visualizations and insights will shed light on the relationships between nanoformulation methodology, nanoparticle characterization, and biological outcome.

While the nanoformulation-brain database we developed is currently tailored to our lab, our methodology and applications are generalizable across research laboratories while being extendable to start-up and large pharmaceutical applications. Outside of academic laboratories, start-ups and larger pharmaceutical companies are likely familiar or already utilizing databases with best business practices for product tracking and quality control. However, these companies could consider building and integrating databases to enhance product development and information connectivity during the research and development to manufacturing process. Both start-ups and big pharmaceutical companies can follow the methodology applied here for connecting biological outcome variables to methodological features via each business's specific experimental metadata while extending database connectivity with federally approved drug and therapeutic product databases for enhanced drug development capabilities.

To improve the database as a tool to build translational capability, several current limitations of the nanoformulation-brain database are important to note and belong in two groups: software development and the nature of experimental work. Currently, the database does not include easy-to-use graphical interfaces or methodologies for automatic data upload. We need to develop an interface for data upload by researchers without data science or computer science expertise as well as procedures for regular system back-up for local data storage. Additionally, experimental data can have high noise, vary in quality depending on methodology or researcher, and is often in a continuous state of optimization or evolution. To improve the database, future work will develop a tagging system so researchers can annotate specific data with qualitative notes that may impact the integrity of the data such as contamination, user error, or data obtained during optimization or training of specific methodologies.

Therefore, we envision future work in three areas: robust database development, database-enabled drug screening, and experimental integration with molecular modeling. The first goal, to develop a robust database, will create a database that is open access, easy to contribute to, simple to query, and includes quality checks for data accuracy. Additionally, to apply the database as a nanotherapeutic screening method, we aim to integrate and validate the database-enabled OWH model results with additional biological characterization methods and nanoparticles with a wide range of physical and drug-loading characteristics. Finally, computational modeling for nanotherapeutics provides large volumes of information to inform better nanotherapeutic development. An experimental nanoformulation-brain database integrated with the wealth of information from computation modeling enabled by machine learning enables high-throughput insight between modeled and experimental nanotherapeutics and experiment-informed prediction of biological outcome. Databases are an organized way to store and access years of nanoformulation and biological data to increase insight and connectivity of physical, chemical, and biological characteristics.

4 | CONCLUSIONS

We put forward that a preclinical experimental database can facilitate translation for nano-based neurotherapeutics by connecting nanoformulation methodology and nanoformulation characterization with biological outcomes. Both nanotherapeutics and the brain have similarities in the level of complexity—the physical, chemical, and biological interdependencies and environment-dependent attributes create a

vast design and application space. Tailoring nanotherapeutic design such that nanotherapeutics can overcome biological barriers and reach target sites at therapeutically relevant concentrations remains a need in the nanomedicine field.[24] In addition, limitations in measuring multi-faceted in vivo interactions in clinically relevant models of brain disease can limit nanotherapeutic translation for non-cancerous brain diseases.[25,26] Our nanoformulation-brain database connects nanoformulations to biological applications through experimental details. The database builds a platform that provides additional insight into 6 years of nanoformulation development, including the ability to query nanoformulations based on PDI, ZP, and nanoparticle size. We also developed queries that return all nanoformulations made with specific methodologies, loaded with a specific drug, or formulated with certain surfactants or polymers. We demonstrated the capacity of our database for drug screening using heat maps of double emulsion and nanoprecipitation methodologies and their effects on a measure of brain cell viability. Our querying results indicate different strengths and patterns of correlations for nanoparticle physicochemical properties and formulation methodologies with biological outcome. Based on this initial investigation, well-designed nanoformulation-brain databases have the potential to improve preclinical neurotherapeutic insight and alleviate bottlenecks in clinical translation.

ACKNOWLEDGMENTS

We would like to thank Andrea Joseph, Chih-Chung Chen, Norah Alhindi, Georges Motchoffo Simo, Jessica Pon, and Michael Chungyoun for providing experimental formulation and biological datasets. We also thank Professor Dan Suciu at the University of Washington for mentorship on database development. This work was partially supported by the National Institute of General Medical Sciences Grant #R35GM124677, the National Institute of Child Health and Human Development Grant #R21HD100639, the National Science Foundation Grant #HDR1934292, and the Data Intensive Research Enabling Clean Technology (DIRECT) NSF National Research Traineeship Grant #DGE-1633216.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Elizabeth Nance https://orcid.org/0000-0001-7167-7068

REFERENCES

1. Anselmo AC, Mitragotri S. Nanoparticles in the clinic: an update. *Bioeng Transl Med*. 2019;4(3):e10143.
2. Moore TL, Rodriguez-Lorenzo L, Hirsch V, et al. Nanoparticle colloidal stability in cell culture media and impact on cellular interactions. *Chem Soc Rev*. 2015;44(17):6287-6305.
3. Singh Y, Meher JG, Raval K, et al. Nanoemulsion: concepts, development and applications in drug delivery. *J Control Release*. 2017;252:28-49.
4. Jong WHD, Borm PJA. Drug delivery and nanoparticles: applications and hazards. *Int J Nanomedicine*. 2008;3(2):133-149.
5. Helmbrecht H, Joseph A, Mckenna M, Zhang M, Nance E. Governing transport principles for nanotherapeutic application in the brain. *Curr Opin Chem Eng*. 2020;30:112-119.
6. Loew LM, Schaff JC. The virtual cell: a software environment for computational cell biology. *Trends Biotechnol*. 2001;19(10):401-406.
7. Nadkarni P, Marenco L. Database architectures for neuroscience applications. *Methods in Molecular Biology™*. Humana Press; 2007:37-52.
8. Wetheim S. The brain database: a multimedia neuroscience database for research and teaching. *Proc Annu Symp Comput Appl Med Care*. 1989;399-404.
9. Halavi M, Polavaram S, Donohue DE, et al. NeuroMorpho.Org implementation of digital neuroscience: dense coverage and integration with the NIF. *Neuroinformatics*. 2008;6(3):241-252.
10. Milo R, Jorgensen P, Moran U, Weber G, Springer M. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res*. 2010;38(suppl 1):D750-D753. https://doi.org/10.1093/nar/gkp889
11. Ioannidis JPA, Kim BYS, Trounson A. How to design preclinical studies in nanomedicine and cell therapy to maximize the prospects of clinical translation. *Nat Biomed Eng*. 2018;2(11):797-809.
12. Aban IB, George B. Statistical considerations for preclinical studies. *Exp Neurol*. 2015;270:82-87.
13. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406.
14. Steward O, Balice-Gordon R. Rigor or mortis: best practices for preclinical research in neuroscience. *Neuron*. 2014;84(3):572-581.
15. Cai P, Zhang X, Wang M, Wu Y-L, Chen X. Combinatorial Nano–Bio Interfaces. *ACS Nano*. 2018;12(6):5078-5084.
16. Design preclinical studies for reproducibility. *Nat Biomed Eng*. 2018;2(11):789-790.
17. Kitchin R, Mcardle G. What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc*. 2016;3(1):205395171663113.
18. Thomas DG, Klaessig F, Harper SL, et al. Informatics and standards for nanomedicine technology. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*. 2011;3(5):511-532.
19. Karp NA. Reproducible preclinical research—is embracing variability the answer? *PLoS Biol*. 2018;16(3):e2005413.
20. Fothergill BT, Knight W, Stahl BC, Ulnicane I. Responsible data governance of neuroscience big data. *Front Neuroinform*. 2019;13. https://www.frontiersin.org/article/10.3389/fninf.2019.00028
21. Dageville B, Cruanes T & Zukowski M et al. The Snowflake Elastic Data Warehouse. Proceedings of the 2016 International Conference on Management of Data. 2016.
22. Liao R, Pon J, Chungyoun M, Nance E. Enzymatic protection and biocompatibility screening of enzyme-loaded polymeric nanoparticles for neurotherapeutic applications. *Biomaterials*. 2020;257:120238.
23. Feigin VL, Nichols E, Alam T, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. *Lancet Neurol*. 2019;18(5):459-480.
24. Wu L-P, Wang D, Li Z. Grand challenges in nanomedicine. *Mater Sci Eng C Mater Biol Appl*. 2020;106:110302.
25. Hua S, De Matos MBC, Metselaar JM, Storm G. Current trends and challenges in the clinical translation of nanoparticulate nanomedicines: pathways for translational development and commercialization. *Front Pharmacol*. 2018;9. https://www.frontiersin.org/article/10.3389/fphar.2018.00790
26. Mitchell MJ, Billingsley MM, Haley RM, Wechsler ME, Peppas NA, Langer R. Engineering precision nanoparticles for drug delivery. *Nat Rev Drug Discov*. 2021;20(2):101-124.

How to cite this article: Helmbrecht H, Xu N, Liao R, Nance E. Data management schema design for effective nanoparticle formulation for neurotherapeutics. *AIChE J*. 2021;67(12):e17459. doi:10.1002/aic.17459