



RaDaRR: Unlocking Historical Data for Rapid Threat Assessment

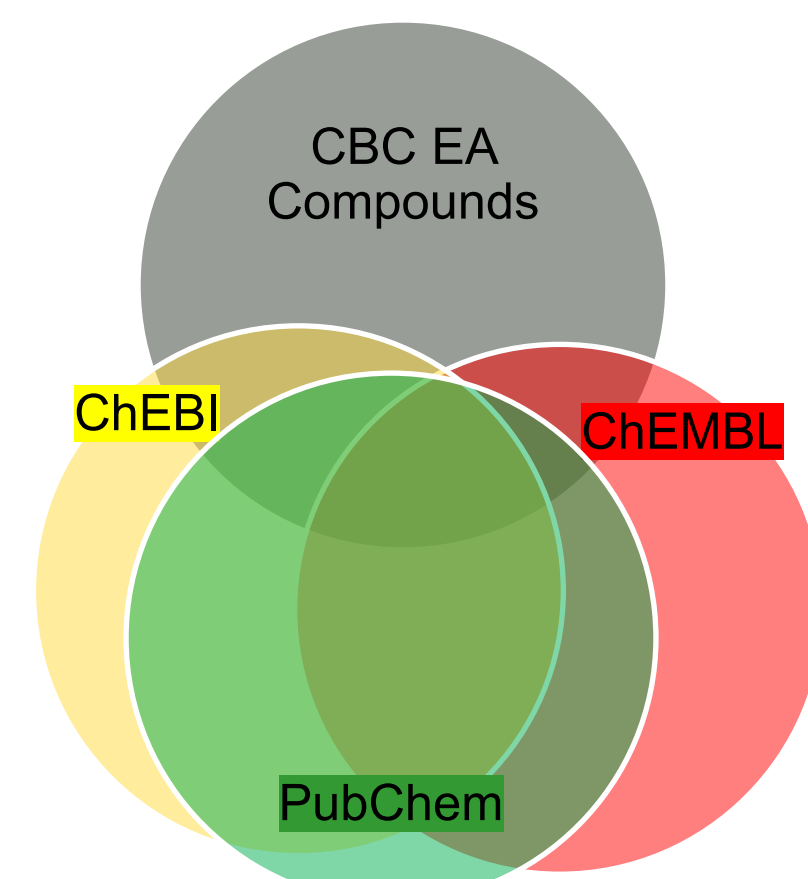
Kevin Ulmes¹, James Myslinski¹, Christopher Ellis¹, Andrew Walz¹, David Cullinan¹, Mark Winemiller¹, Kyle Glover¹

¹U.S. Army Combat Capabilities Development Command Chemical Biological Center, Aberdeen Proving Ground, MD



THE CHALLENGE

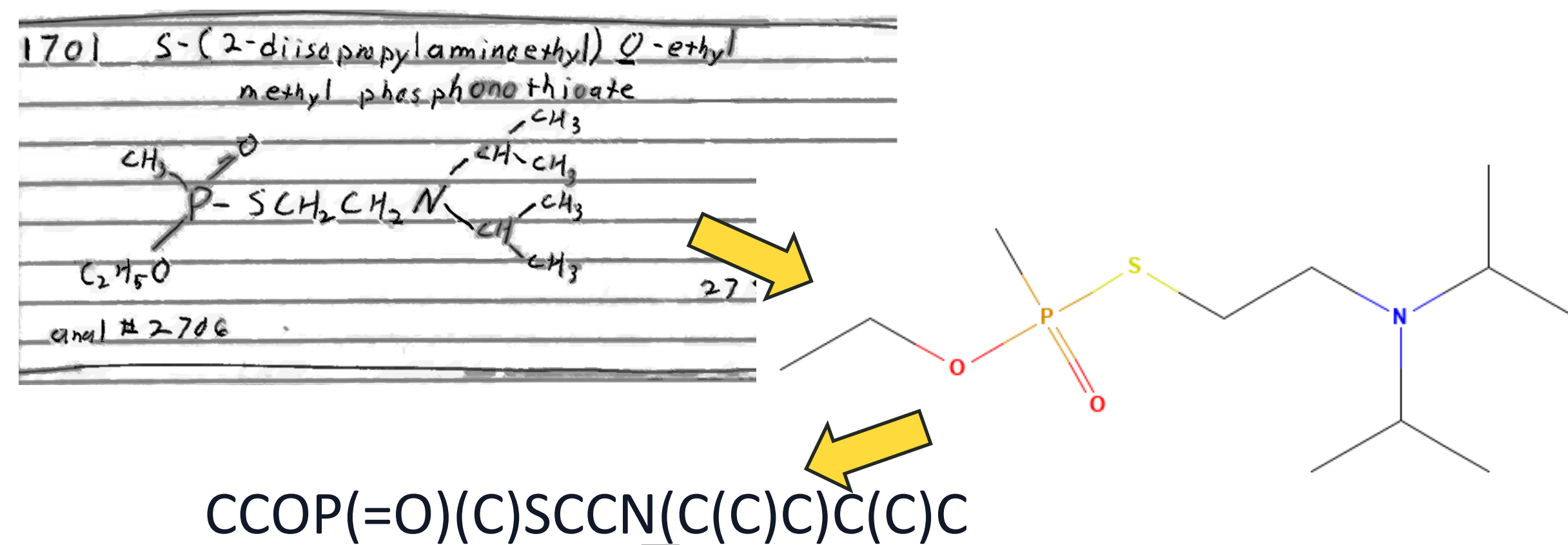
CBC has a unique set of data collected over more than 100 years of research. Much of this data is collected in technical reports dating back to the early 1900's or in lab notebooks in CBC's technical library. The RADARR effort is designed to organize this data from a chemical structure perspective and make this large historical data set more easily accessible for modern cheminformatic, big data, and AI/ML related projects.



PROJECT SMILES

To keep track of the many chemicals CBC has worked with historically the EA numbering system was developed in 1919 as an internal identification number. EA numbers are primarily linked to chemical structures but can also refer to mixtures or even products of a chemical reaction where the precise structure was unknown.

The official link between EA numbers and structures has been maintained in a series of composition notebooks kept as part of the Agent Chemistry Branch at CBC since the 1940's. Project SMILES was the initial effort of RADARR to digitize the chemical structures from hand drawn structures into Simplified Molecular Input Line Notation System (SMILES) strings into a database linking them to EA numbers.



EA Number	CASRN	Military C. Class	Chemical Name	IUPAC	Formula	MW	SMILES
EA1699							
EA1700							
EA1701	50782-69-9	VX	Nerve	Ethyl -S- 2-diisopropylaminoethyl methylphosphonothiolate	N-[2-[etho C11H26NO2]PS	267.37	CCOP(=O)(C)SCCN(C(C)C)C(C)C
EA1702							
EA1703							



DEVCOM CBC @ DTRA CBD S&T Conference

Scan the QR Code to view all of CBC's
2024 DTRA CBD S&T Conference materials
<https://cbc.devcom.army.mil/cbdst-conference/>

HURDLES

- Digitization of bound notebooks
 - Fragile notebooks from 1920's
 - Only source of structure for many compounds
 - Only available scanner was sheet fed
- Original data inconsistency
 - Entries flagged when
 - Name, structure, formula or molecular weight didn't match
 - Structure was unclear, or oddly drawn (7 bonds to a carbon)
 - Aromaticity was unspecified

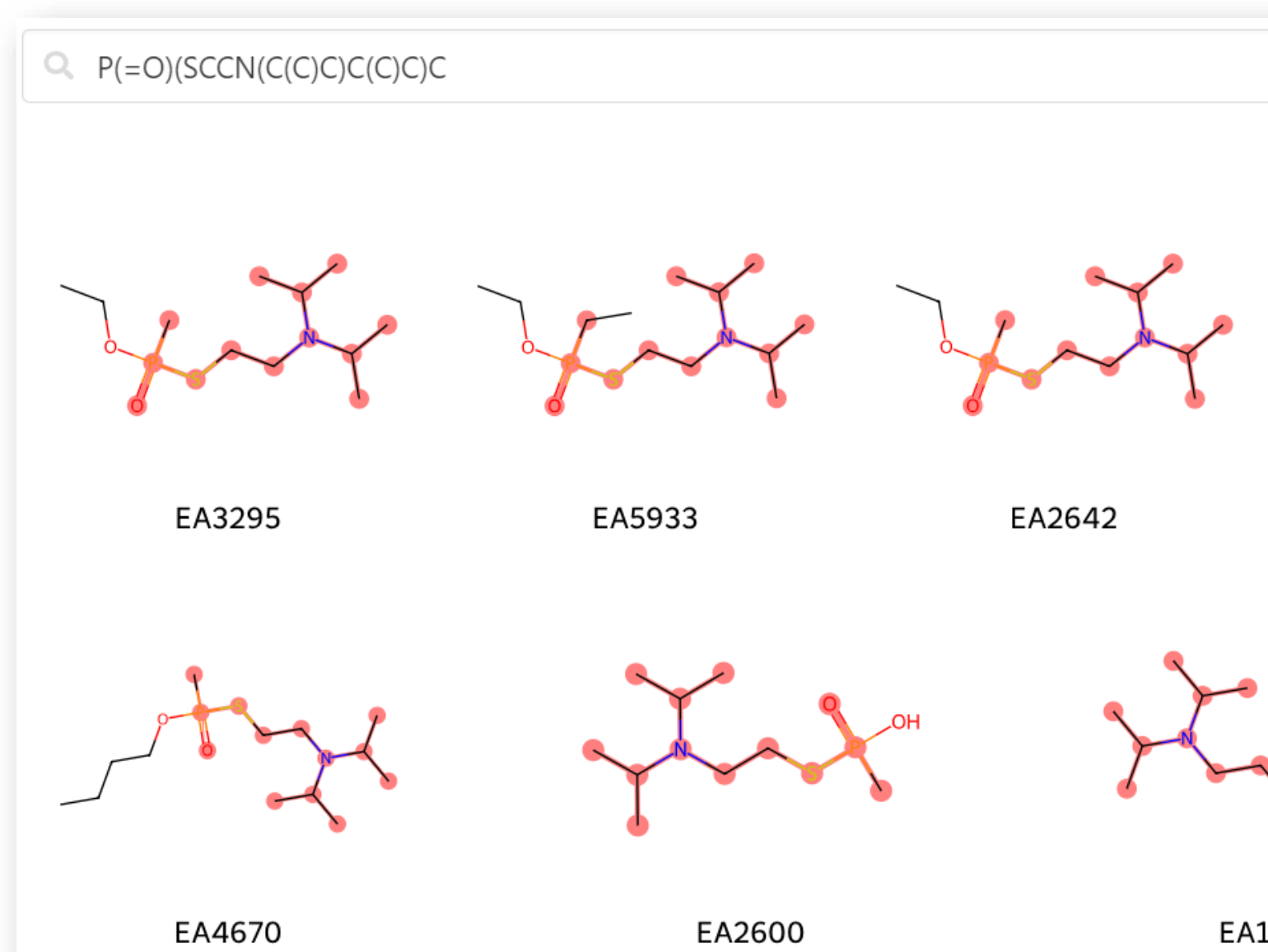
QA/QC

Independent "double entry" QA/QC effort ongoing to ensure RADARR dataset can be used with confidence and serve as an authoritative record of CBC's EA numbering system.

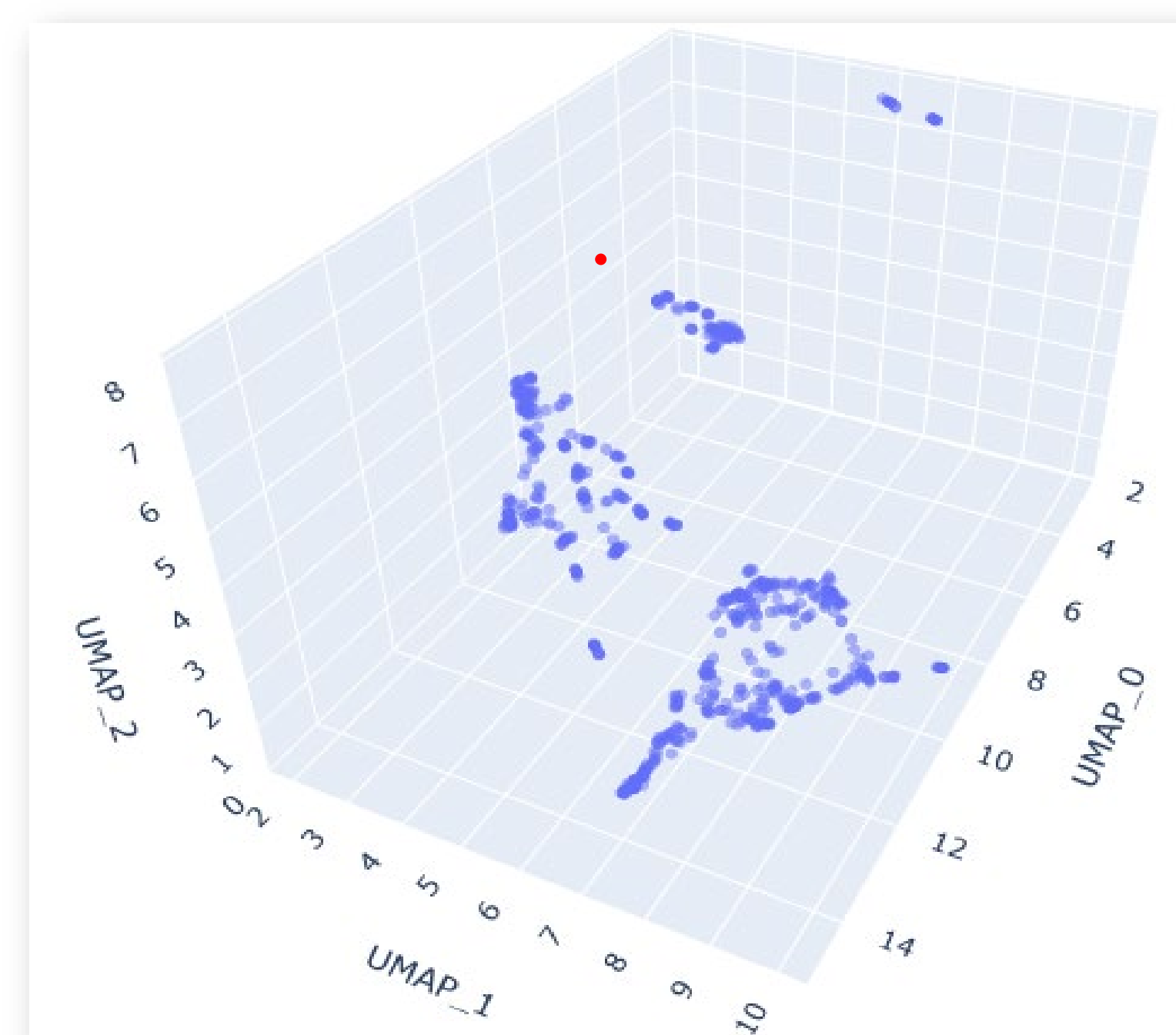
RADARR TOOLS AND USES

Online web-based JavaScript tool for quick searching of database by chemical similarity or sub-structure

- Results in seconds sorted by chemical structure similarity (multiple fingerprints and similarity score metrics available)
- Large decrease in EA chemical space searching time and chemical expertise required vs a chemistry SME paging through composition notebooks



*Note: EA Numbers and structures are randomly generated for example only

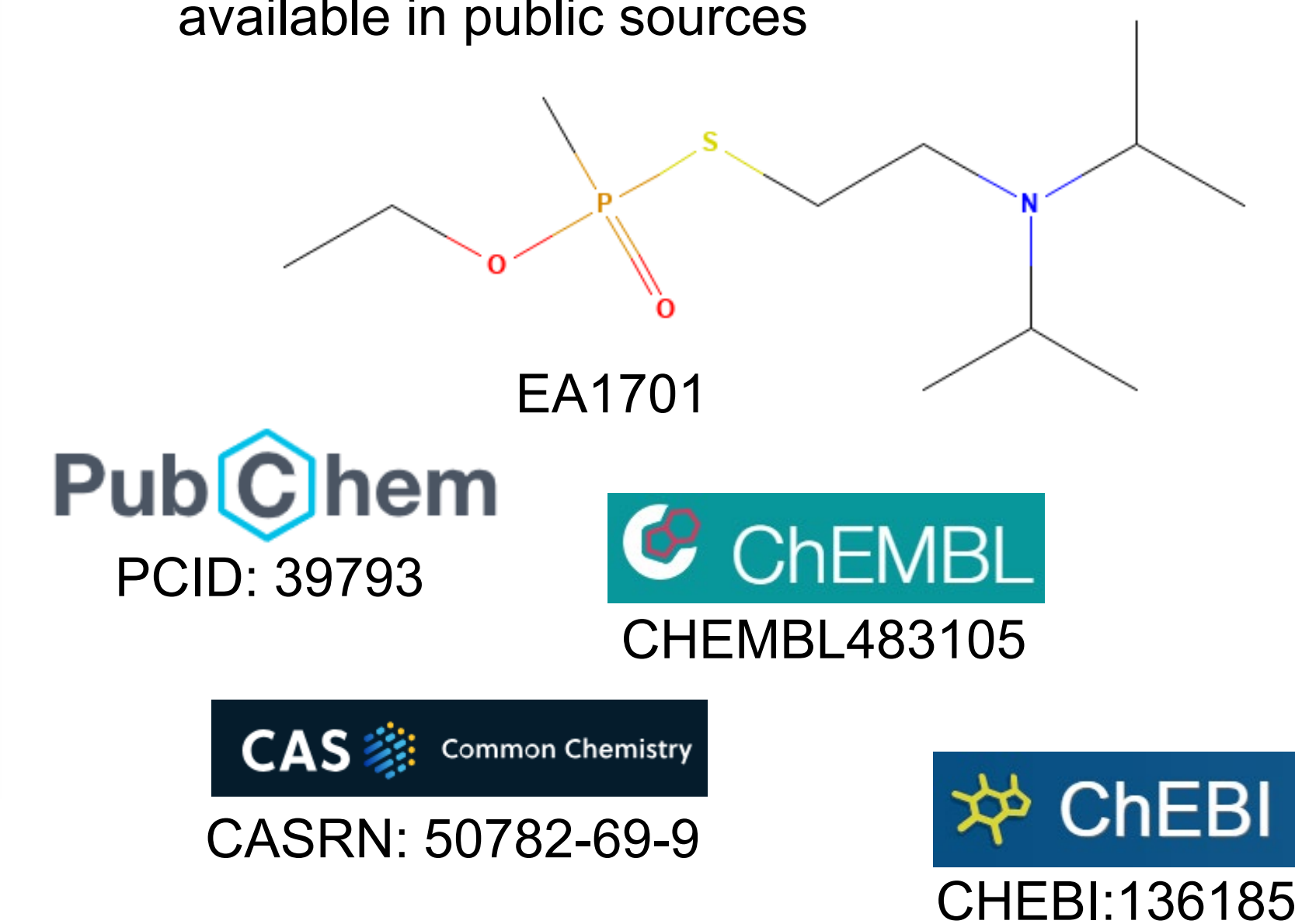


- Visualization of EA library space via multiple fingerprinting and dimensionality reduction methods
- Allows mapping of novel compound structures into EA Library "space" to identify most similar neighbors that CBC may have data on
 - Fingerprints: MACCS, Morgan, RDKit, AtomPair, Torsion, CACTVS, Saagar
 - Dimensionality reduction: PCA, t-SNE, UMAP

RESULTS

- 17 Composition notebooks scanned into PDF's
- ~5700 Compounds digitized over the course of 9 months
 - Molecular Weight: 1,684,718
 - Atoms: 106,342
 - Bonds: 107,830
 - Rings: 9,385
 - Salts/mixtures: 2,653
- ~1000 compounds flagged for further review
- Adjudication meetings for those 1000 compounds
 - 6 SME's looked at original source and agreed on most likely structure
 - Original lab notebooks consulted for further details

- Linkage of EA Numbers with external databases
- Linkage via SMILES structures ensures exact chemical matches
 - Allows CBC to leverage already curated data available in public sources



Future Uses:

- Training or fine tuning of QSAR or other ML models with CBC Historical data to better capture highly toxic chemical space
- Readily available structure database for use in molecular dynamics workflows

THE FUTURE

RaDaRR: Rapid Data Retrieval for Read-Across of Emerging Threats

Read-across methods require both identification of similar compounds and the data associated with those compounds for the properties of interest. Phase 1 involved digitization of chemical structures to enable similarity searching. Follow on projects will link reports and data to the EA structures to complete the ability to implement read-across methodologies for emerging threats.

Project Books

- Initial effort to link technical report to EA Compounds
- Enables searching and identification of technical reports based on structural similarity
- Potential use of AI (LLM's) to assist in identification of reports with actual data vs just mention of an EA number, and identification of types of data in report
 - For example: EA1701 mentioned in TR-12345 in a toxicological context
- Digitization and incorporation of "grey literature" available only in CBC technical library
 - Lab Notebooks
 - Unpublished Data
 - Historical reports

Project Data

- Effort to extract data from technical report into database format
- Provides large amount of tabulated data on unique compounds as a basis for future cheminformatic, AI/ML, or chemical data sheets
- Combine with external chemical databases to provide in house rapid access to digital structured information on chemicals
- Leverage previous work to provide data on chemical structure linked to original reports
 - Provides context of individual data points by linking back to original report
 - Allows CBC to maintain "provenance" of data to ensure that the original sources are traceable and the data is reliable

Acknowledgements: Funding for RADARR was provided by DTRA JSTO (Sara Peacock and Kristen Fulbright) The views expressed in this abstract are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense or the U.S. Government.