Lists of Chemical Warfare Agents and Precursors from International Nonproliferation Frameworks: Structural Annotation and Chemical Fingerprint Analysis

Stefano Costanzi,*[,a] Charlotte K. Slavick,[a] Brent O. Hutcheson,[a] Gregory D. Koblentz,[b] Richard T. Cupitt[c]

[a]*Department of Chemistry, American University, 4400 Massachusetts Avenue, NW, Washington, DC 20016, USA* [b]*Schar School of Policy and Government, George Mason University, 3351 Fairfax Drive, Arlington, VA 22201, USA* [c]*Stimson Center, 1211 Connecticut Ave, NW, Washington, DC 20036, USA*

*** Corresponding author**

Tel: +1-202-885-1722

Email: costanzi@american.edu

Supporting information available free of charge at: https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00896.

ABSTRACT

To support efforts to stem the proliferation of chemical weapons (CW), we have curated and structurally annotated CW-control lists from the three key international nonproliferation frameworks: the Chemical Weapons Convention (CWC), the Australia Group (AG), and the Wassenaar arrangement (WA). The curated lists are available as web tables at the Costanzi Research website (https://costanziresearch.com/cw-control-lists/). The annotations include manually curated 2D structural images, which provide a means to appreciate at a glance the similarities and differences between the different entries, as well as downloadable 2D structures, in two different formats, and three different structural identifiers, namely SMILES, standard InChI, and standard InChIKey, which are intended to provide a platform for cheminformatics analyses. The tables also include links to NCBI's PubChem and NIST's Chemistry WebBook cards, hence providing prompt access to a wealth of physicochemical, analytical chemistry, and toxicological information. To showcase the importance of structural annotations, we discuss a discrepancy in a CW-control list covering the defoliant Agent Orange, which we identified through our curation process, and propose a solution to address it. Moreover, we present the results of chemical fingerprinting analyses, through which we clustered the entries of the three CW-control lists under study into structurally related groups and studied the overlaps between the three lists. As an application of this study, we examine the recent updates of CWC Schedule 1 and the AG precursors list, highlighting the relationships between the two amendments and proposing the possible addition of further chemicals. Our research is intended to facilitate the communication between scientific advisors and policymakers as well as the work of chemists and cheminformaticians involved in the CW nonproliferation field. Ultimately, we seek to provide tools to bolster the control of CWs and support the global efforts to rid the world of this category of weapons.

INTRODUCTION

Chemical weapons (CWs) are weapons that exploit the toxicity of chemicals to bring about death or harm. The toxic chemicals on which chemical weapons are based are known as chemical warfare agents (CWAs). Based on their mechanism of action, CWAs can be divided into different classes, including

vesicants, choking agents, chemosensory irritants, blood agents, and central incapacitating agents.[1–3]

Notoriously, chemical weapons – chiefly irritants, choking agents, and vesicants – were widely used in World War I (WWI). More sophisticated chemical weapons, *e.g.* nerve agents and incapacitating agents, were subsequently developed prior and during World War II (WWII) and during the Cold War.[1–3]

Far from being a relic of the past, chemical weapons remain a current threat to national and international security. Historically produced in industrial amounts and regarded as Weapons of Mass Destruction (WMD), in recent years chemical weapons have been employed for operations that require significantly smaller quantities, such as counterinsurgency or targeted assassinations.[4,5] In particular, a variety of chemical weapons were used by state and non-state actors in Syria as well as non-state actors in Iraq.[6] A binary VX nerve agent was used for the assassination of Kim Jong-nam, the half-brother of North Korea's leader Kim Jong-un, at the Kuala Lumpur airport in Malaysia in February 2017. A Novichok nerve agent known as A-234 was used for the attempted assassination of Sergei Skripal and his daughter in Salisbury, England, in March 2018. A few months later, a woman resident of the nearby town of Amesbury, who came into contact with a discarded faux perfume bottle containing the same agent, tragically died.[1] A nerve agent belonging to the Novichok family, according to German authorities, was also used for the attempted assassination of Alexei Navalny in August 2020.[7] As new chemical technologies continue to be developed,[8] malicious actors will likely attempt to exploit them to acquire known agents or develop new ones.

To stem the proliferation of chemical weapons, *i.e*. to prevent actors from acquiring or developing them, and ensure the elimination of the currently existing weapons a number of international nonproliferation frameworks are in place, such as the Chemical Weapons Convention (CWC), the Australia Group (AG), and the Wassenaar Arrangement (WA). The CWC, which was opened for signature in 1993 and entered into force in 1997, poses a complete ban on chemical weapons. Importantly, state parties are prohibited not only from using or engaging in military preparations to use chemical weapons, but also from developing, producing, acquiring, stockpiling, retaining, or transferring them to other parties. Assisting anyone in engaging in activities that violate the Convention is also forbidden. The CWC, which is implemented by the Hague-based Organization for the Prohibition of Chemical Weapons (OPCW), enjoys

almost universal adoption, counting a total of 193 state parties.[1,9,10] The AG is an informal forum of countries that coordinate export control regulations for dual use items, *i.e.* items that can be used for civilian purposes as well as for the production of weapons, including precursors for the synthesis of CWAs. The group was established in 1984, out of concerns for the use of chemical weapons by Iraq in the Iran-Iraq War, and currently counts 42 participant countries, plus the European Union. The WA is an international framework intended to foster transparency in the transfer of conventional weapons and dual-use items, including chemicals with military applications. It was established in 1995 and, like the AG, it currently counts 42 participant countries. However, there is not a perfect overlap between the AG and WA membership, since Iceland and the Republic of Cyprus are AG but not WA members, while Russia and South Africa are WA members but not AG members. Generally, most members in the UN Western European and Others Group (WEOG) of countries are members of both AG and WA.[9]

To serve their purpose, these frameworks contain lists of chemicals that can be employed as CWAs and/or precursors for their synthesis (hereafter referred to as "CW-control lists"). Some of these lists are a compilation of discrete chemicals, where each chemical is individually enumerated. Other lists comprise both discrete chemicals as well as families of chemicals identified by a common scaffold with variable substituents. It is worth emphasizing that the purpose of these lists is to support efforts to control the proliferation of chemical weapons. None of these lists should be construed as an exhaustive compilation of CWAs and precursors. Indeed, the CWC includes within the scope of its definition of chemical weapons "*any chemical* (emphasis added) which through its chemical action on life processes can cause death, temporary incapacitation or permanent harm to humans or animals" (CWC, Article II, Paragraph 2).[11] Hence, any weapon designed to bring about death, temporary incapacitation, or permanent harm through the toxic properties of chemicals is to be considered a chemical weapon, and any toxic chemical at the basis of such weapons is to be considered a CWA.[1,9] For instance, chlorine, which has been extensively used as a chemical weapon in the Syrian Civil War, is not covered by the CWC schedules due to its extensive legitimate uses, which would make its control impractical. Nonetheless, the intentional employment of

chlorine to bring about harm or death undoubtedly qualifies as chemical weapon use and constitutes a violation of the CWC prohibitions.[6]

The CWC, in its Annex on Chemicals, comprises three tiered schedules of chemicals intended to support its declaration and verification regime. Schedule 1 comprises chemicals regarded exclusively or primarily as CWAs or precursors. Conversely, Schedules 2 and 3 comprise dual use items that, beyond their chemical weapon role, also have legitimate applications, on a smaller scale for Schedule 2 and on a larger scale for Schedule 3. Schedule 1 and Schedule 2 contain both families of chemicals as well as discrete chemicals, while Schedule 3 contains discrete chemicals only. Each schedule is divided into two parts: part A, which is reserved to CWAs, and part B, which is reserved to CWA precursors.[1,9,10] In the aftermath of the Salisbury incidents, the CWC schedules were revised for the first time after the treaty's entrance into force. The revision, which was approved by the Conference of State Parties in November 2019 and became effective in June 2020, resulted in the expansion of part A of Schedule 1 to include Novichok agents developed as CWAs in the Soviet Union during the Cold War as well as carbamates researched by the United States during the same period, although reportedly never developed into CWAs.[12,13] The AG comprises a list of precursors for the synthesis of CWAs, all of which are listed as discrete chemicals (hereafter referred to as "AG precursors list").[9] Following the addition of Novichoks to the CWC Schedule 1, precursors for their synthesis were added to the AG precursors list as well – for more information on the Salisbury events and the CWC and AG amendments, see *Recent amendments of the CWC Schedule 1 and the AG precursors list* in the **Conclusions** section. Lastly, the WA, in its Munitions List 7 (ML7), comprises a list of chemicals of proliferation concern. In particular, the WA ML7 mirrors the CWC Schedule 1, to which it adds one chemical from Schedule 2.[9] At the time of this writing, the Wassenaar arrangement has not yet been updated with the Novichok and carbamate agents recently added to CWC Schedule 1. However, it is reasonable to expect an update in this direction in the near future. Beyond the mentioned overlaps with the CWC schedules, the WA ML7 comprises a number of riot control agents and defoliants, none of which are listed in the CWC schedules, as they do not fit within the CWC definition of chemical weapons.[1,10] In particular, according to the CWC, riot control agents are defined as "any chemical *not listed*

in a Schedule (emphasis added), which can produce rapidly in humans sensory irritation or disabling physical effects which disappear within a short time following termination of exposure" (CWC Article II, paragraph 7) – for a recommendation from the Scientific Advisory Board to the OPCW regarding which chemicals can be legitimately regarded as riot control agents, see Timperley at Al.[14] Moreover, defoliants are not covered by the CWC, which focuses on chemicals toxic for humans or animals. The primary international framework covering defoliants is the Environmental Modification Convention.[1,10]

The above-mentioned CW-control lists are provided by the international frameworks as lists of chemical names and, for discrete chemicals, Chemical Abstract Service (CAS) registry numbers. To support these multilateral efforts to stem the proliferation of chemical weapons, we have curated and structurally annotated the CW-control lists from the three above mentioned frameworks, thus providing intuitive and rapid access to structural and chemical information. The curated lists are available as web tables at the Costanzi Research website (https://costanziresearch.com/cw-control-lists/), and are provided in the Supporting Information in PDF format (**Tables S1-S3**). This work is intended to facilitate the communication between scientific advisors and policymakers as well as the work of chemists and cheminformaticians involved in the CW nonproliferation field. Ultimately, we seek to provide tools to bolster the control of CWs and support the global efforts to rid the world of this category of weapons.

First, in our tables, all entries are annotated with chemical structures (exact structures for discrete chemicals and Markush structures for families of chemicals). This is important because chemicals are better described by structures than names. As we highlight in the **Discussion** section, a single letter difference in a chemical name can account for an important difference in the chemical structure, marking the watershed between chemicals that are controlled and those that are not. Hence, documents annotated with structures will make it easier for chemists and scientific advisors to communicate these differences to policymakers.

Our tables also are annotated with information that indicates the overlaps within the various CW-control lists, noting for every entry of each list whether that chemical is covered by one or more additional lists, either as a discrete chemical or as a member of a family of chemicals. Importantly, to further highlight

overlaps and differences between the lists, we also provide a synoptic table in which all the lists are provided side-by-side on one page (**Table S4**).

METHODOLOGY

Annotation of CW-control lists: entry type. The three international CW-control lists under study, consisting of chemical names and, for discrete chemicals, CAS registry numbers, were saved as comma-separated values (CSV) files and manually curated to add an "Entry Type" column. Specifically, to each entry, one of the following entry types was assigned, as appropriate: family (for entries describing families of related chemicals defined as a central scaffold with variable substituents), family example (for discrete chemicals provided as examples of a chemical family), family exception (for discrete chemicals that would fall within the scope a listed family, but are explicitly excluded from the CW-control list), individual chemical (for discrete small-molecule entries), proteins (for discrete protein entries), and mixtures (for entries defined as a mixture of discrete chemicals).

Automatic annotation of CW-control lists: chemical identifiers. Subsequently, the CSV files relative to the three CW-control lists were automatically curated following the algorithm shown in **Figure 1**. The curation was done through an in-house implemented R script based on the WebChem package (Version 1.0.0),[15,16] For all discrete chemicals (including individual chemicals, family examples, family exceptions, and mixtures), the automatic annotation script added columns with PubChem identifiers (CID) as well as structural identifiers, including SMILES, standard InChI, and standard InChIKey. As shown in **Figure 1**, the searches featuring a PubChem CID, either as input or output, were conducted through the NLM's PubChem website (https://pubchem.ncbi.nlm.nih.gov).[17] All other searches were conducted through the Chemical Identifier Resolver (CIR, https://cactus.nci.nih.gov/chemical/structure), a tool provided by the National Cancer Institute's (NCI) Computer-Aided Drug Design (CADD) Group Chemoinformatics Tools and User Services (CACTUS). The output of the automatic script was visually inspected and, whenever needed, manually corrected. Importantly, we ensured that all SMILES reflected the tautomeric form

indicated by the entry's CAS registry number. Conversely, this was not possible for InChI strings, as the standardization process does not allow designating specific tautomeric forms. Regarding stereochemistry, in the three CW-control lists object of this work, with one exception, all chemicals containing chiral centers are listed with unspecified chiral configuration. Thus, we ensured that all chemicals were kept with unspecified chirality in all SMILES and InChI strings. The lone exception is the natural compound saxitoxin, which is listed in CWC Schedule 1 as CWC 1A7 with defined stereochemistry. For this compound, we ensured that the SMILES and InChI string reflected the stereochemistry indicated by the CAS registry number provided in CWC Schedule 1.

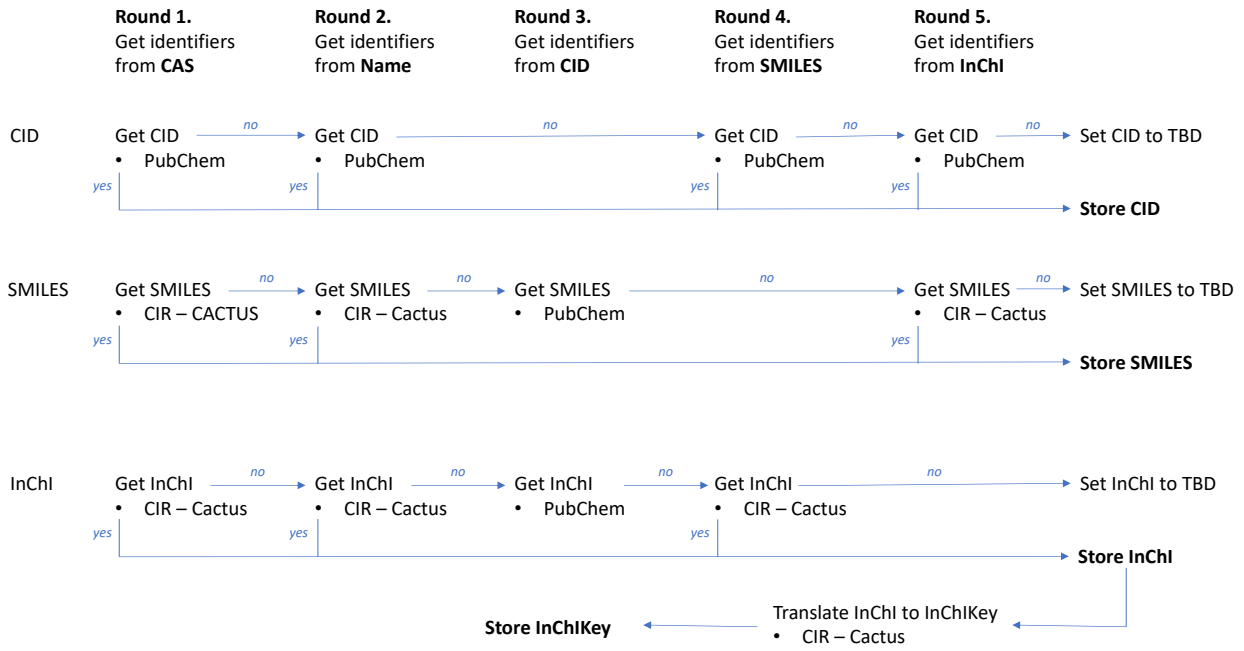


Figure 1. Algorithm for the automatic annotation of CW-control lists with PubChem identifiers (CID) as well as SMILES, standard InChI, and standard InChIKey structural identifiers. The annotation algorithm is implemented through an R script based on the WebChem package.

Drawing molecular structures in MarvinSketch. Each of the three CSV files was subsequently imported as a "query table" into ChemAxon's Instant JChem (Version 20.4.0 - https://chemaxon.com/products/instant-jchem), converting in batch the SMILES identifiers into 2D molecular structures. The structures were subsequently manually curated to ensure the clarity of the 2D rendering (e.g. avoiding overlapping atoms). Moreover, it was ensured that all molecules based on similar molecular scaffolds were represented with the same spatial orientation. Because the 2D structures were derived from the curated SMILES identifiers, they all reflect the tautomeric form and stereochemistry indicated in by CAS registry numbers provided in the CW-control lists. We verified the correspondence through visual inspection and manually corrected the structures whenever needed.

Chemical fingerprint analyses. Chemical fingerprint analyses were conducted with an in-house implemented R script based on the Rcdk package (Version 2.3).[15,18] First, a global CW-control list was assembled by appending, one after another, the three individual CSV files relative to the three CW-control lists under study. The only protein entry (CWC 1A8, ricin) was expunged from the list, as proteins are not suitable for chemical fingerprinting analysis. For entries referring to families of chemicals rather than discrete chemicals, a SMILES identifier was assigned corresponding to the molecular structure of the lowest molecular weight example provided in the CW-control list. In cases where examples are not provided, all alkyl substituents were set to methyl. Subsequently, the SMILES identifiers were converted into Rcdk molecular entities. From these, two chemical fingerprints were calculated, namely 166 bit MACCS keys and ECFP6 extended-connectivity circular fingerprints.[19,20] Subsequently, a pairwise similarity matrix among the generated fingerprints was calculated based on the Tanimoto metric, which can assume values ranging between 0 and 1, with 0 indicating entirely non-overlapping fingerprints and 1 indicating identical fingerprints. From the similarity matrix, a distance matrix was calculated through the formula: *distance = 1 – similarity*.

Hierarchical clustering and tree plotting. Through the *hcust* function of R, the fingerprint distance matrix was subjected to a hierarchical clustering analysis based on the Ward.D2 clustering method. The resulting output was plotted as a "fan plot" with the APE package of R.[21] The *use.edge.length* logical variable was set as "TRUE," thus resulting in the length of the tree branches being proportional to the fingerprint distance.

Identification of overlaps: identical discrete chemicals and identical families. Overlaps between discrete chemicals were identified by expunging from the distance matrix based on the ECFP6 fingerprints all entries not showing a distance of 0 with at least another entry. The process was done separately for all entries referring to discrete chemicals and all entries referring to protein families, thus identifying overlapping discrete chemicals (*i.e.* discrete chemicals found in more than one list) and overlapping families (*i.e.* families of chemicals found in more than one list). The ECFP6 fingerprints were chosen for this analysis because they yield identical fingerprints exclusively in the case of chemical identity. Identical compounds will always have the same standard InChI and InChIKey strings. Hence, to double check our results, we verified that: a) all the identified overlapping compounds had matching standard InChIKey strings; and b) no compounds with matching InChIKey strings were missing from the list of identified overlapping compounds

Identification of overlaps: discrete chemicals that fall within the definition of a family of chemicals. Discrete chemicals that fall within the definition of a family of chemicals were identified with ChemAxon's Instant JChem (Version 20.4.0 - https://chemaxon.com/products/instant-jchem), subjecting the three query tables corresponding to each of the CW-control lists to an overlap analysis. Families of chemicals were represented with Markush structures corresponding to the entry definition. Moreover, the query tables were standardized to clean aromatic rings, clear isotopes, neutralize charges, remove fragments, remove solvents,

and strip salts. The search was conducted in "superstructure" mode, and the results were visually parsed to ensure the accuracy of the identified overlaps.

Final curation of CW-control lists and building of HTML tables. The CSV files relative to the three CW-control lists were further manually annotated to add information regarding, CWA Class,[1] and overlaps with other schedules. Finally, the CSV files were converted into HTML tables with an in-house implemented R script. In the process, the rows were color-coded based on the value of the Entry Type column, and links downloadable molecular structures hosted in the Costanzi Research site as well as external web resources were added.

RESULTS

Annotated CW-control lists. By means of the automated process described in the Methodology section, followed by a thorough visual inspection and, whenever needed, manual rectification, we have annotated the CW-control lists encompassed by the three main international frameworks for the control of chemical weapons: the three CWC schedules, the AG precursors list, and the WA ML7. The annotated lists are available as web tables at the Costanzi Research website (https://costanziresearch.com/cw-nonproliferation/cw-control-lists/). They are also provided in the Supporting Information in PDF format (**Tables S1 – S3**).

To facilitate the work of chemists working in CW nonproliferation, our tables are annotated with structural identifiers – including simplified molecular-input line-entry system (SMILES), International Chemical Identifier (InChI), and hashed InChI (InChIKey) notations – as well as links to in-house generated downloadable 2D structures hosted in our website, and PubChem identification numbers. Finally, our tables feature links to interactive 3D models of the chemicals visualized through NLM's iCn3D tool, as well as links to web cards that give prompt access to a wealth of chemical information on the entries, including the National Library of Medicine's (NLM) PubChem cards and the National Institute of Standards and

Technology's (NIST) Chemistry WebBook cards. As evident from **Figure 2**, which shows a snippet of the annotated CW-control lists, each table is composed by 10 columns.

The first column provides the CWA class to which the entry belongs. Specifically, in alphabetical order, entries are divided into central incapacitants, choking agents, defoliants, nerve agents (comprising canonical nerve agents, novichok nerve agents, and non-organophosphorus carbamate nerve agents), vesicants (comprising lewisites, nitrogen mustards, and sulfur mustards), biological toxins, precursors, and riot control agents. By definition, all entries belonging to part B of the three CWC Schedules, as well as all entries belonging to the AG precursors list, are listed as precursors.

The second column provides the entry number, i.e. the number given to that entry in the specific CW-control list. For clarity, we added as a prefix to the entry number an acronym that identifies the CW-control list to which the entry number refers: CWC for Chemical Weapons Convention schedules, AG for the Australia Group precursors list, and WA ML7 for the Wassenaar Arrangement Munitions List 7. For the CWC schedules, as it is common practice, entry numbers are composed of a first number that refers to the schedule, followed by the letter A (for part A) or the letter B (for part B), followed by a second number that refers to the listing order of the entry within that list. For instance, CWC 2B4 indicates the fourth entry of part B of CWC Schedule 2.

The third and fourth columns provide the entry's chemical name and CAS registry number, both exactly as reported by the CW-control lists. Entries that refer to chemical families rather than discrete chemicals, do not have CAS registry numbers assigned to them, as CAS registry numbers are only assigned to specific chemicals. For classification purposes, we assigned to these entries unique CR (Costanzi Research) numbers. If two family entries listed by two different CW-control lists are equivalent to each other, the same CR number is assigned to both entries. For instance, entry CWC 1A1 and Wassenaar ML7 b.1.a, both of which are defined as "O-Alkyl (<=C10, incl. cycloalkyl) alkyl (Me, Et, n-Pr or i-Pr)-phosphonofluoridates" are both assigned the CR number CR0001.

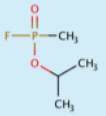
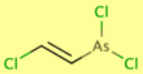
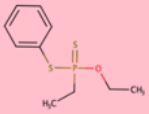
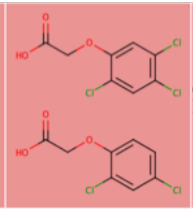
CWA Class	CWC Entry Number	CWC Entry Name	CAS Registry Number (or CR number)	PubChem ID (or UniProt ID)	Structure	SMILES	InChI and InChIKey	Overlaps	Links
Nerve agents	CWC 1A1	O-Alkyl (<=C10, incl. cycloalkyl) alkyl (Me, Et, n-Pr or i-Pr)-phosphonofluoridates	CR0001	---		---	---	WA ML7 b.1.a	---
Nerve agents	CWC 1A1 Example 1	Sarin: O-Isopropyl methylphosphonofluoridate	107-44-8	7871		CC(C)O[P](C)(F)=O	InChI=1S/C4H10FO2P/c1-4(2)7-8(3,5)6/h4H,1-3H3 InChIKey=DYAHQFWOVKZOOW-UHFFFAOYSA-N	WA ML7 b.1.a Example 1	Download 2D Structure (SDF) Download 2D Structure (MRV) Show 3D Structure NIST Chemistry WebBook
Lewisites	CWC 1A5-1	Lewisite 1: 2-Chlorovinyldichloroarsine	541-25-3	5372798		Cl/C=C/[As](Cl)Cl	InChI=1S/C2H2AsCl3/c4-2-1-3(5)6/h1-2H/b2-1+ InChIKey=GIKLTQKNOXNBNY-OWOJBTEDSA-N	WA ML7 b.2.b.1	Download 2D Structure (SDF) Download 2D Structure (MRV) Show 3D Structure NIST Chemistry WebBook
Toxins	CWC 1A8	Ricin	9009-86-3	P02879		---	---	---	Show 3D Structure
Exceptions	CWC 2B4 Exception 1	O-Ethyl S-phenyl ethylphosphonothiolothionate	944-22-9	13676		CCO[P](=S)(CC)Sc1ccccc1	InChI=1S/C10H15OPS2/c1-3-11-12(13,4-2)14-10-8-6-5-7-9-10/h5-9H,3-4H2,1-2H3 InChIKey=KVGLBTYUCJYMND-UHFFFAOYSA-N	---	Download 2D Structure (SDF) Download 2D Structure (MRV) Show 3D Structure NIST Chemistry WebBook
Defoliants	Wassenaar ML7 b.4.b Free Acid Form	2,4,5-trichlorophenoxyacetic acid mixed with 2,4-dichlorophenoxyacetic acid (Agent Orange) **Note:** this is the entry name given in ML7	8015-35-8 **Note:** this CAS Registry Number is not given in ML7 but can be inferred from the name given in ML7	24683		OC(=O)COc1ccc(Cl)cc1Cl.OC(=O)COc1cc(Cl)c(Cl)cc1Cl	InChI=1S/C8H5Cl3O3.C8H6Cl2O3/c9-4-1-6(11)7(2-5(4)10)14-3-8(12)13;9-5-1-2-7(6(10)3-5)13-4-8(11)12/h1-2H,3H2,(H,12,13);1-3H,4H2,(H,11,12) InChIKey=WWPVVPNDJBRXMS-UHFFFAOYSA-N	---	Download 2D Structure (SDF) Download 2D Structure (MRV) Show 3D Structure NIST Chemistry WebBook

Figure 2. Snippet of our annotated CW-control lists. Depending on the entry type, the rows are color-coded as follows: yellow = individual chemicals; red = mixture of chemicals; orange = proteins; blue = families of chemicals; cyan = family examples; pink = family exceptions.

The fifth column provides the PubChem identification number, with a hyperlink pointing to the corresponding PubChem card. For protein entries that do not have PubChem identification numbers, Uniprot identification numbers are given instead, with a hyperlink pointing to the corresponding Uniprot

card. Ricin, which is listed in CWC Schedule 1 as entry 1A8, is the only protein entry listed in the CW-control lists covered by this paper. PubChem cards contain a great deal of information, including, *inter alia*, physicochemical properties, biological activity, safety and toxicity data, chemical structures, structural identifiers, and patents.

The sixth column provides a 2D representation of each chemical structure. As mentioned, we manually curated these structures to ensure the clarity of the 2D rendering. In particular, we avoided any overlapping atoms, which would make the structures difficult to see. Moreover, to make the tables easier to interpret and facilitate the visual comparison of compounds, we ensured that all molecules based on similar molecular scaffolds were represented with the same spatial orientation.

The seventh and the eighth columns provide structural identifiers. In particular, the seventh column provides SMILES identifiers, while the eight column provides the International Union of Pure and Applied Chemistry (IUPAC) InChI and InChIKey identifiers, both in their standard form. These identifiers are strings that define the molecular structures. They can be used as input for a range of cheminformatic applications including, among many others, fingerprint analyses, database searches, calculations of chemical descriptors, and molecular rendering. Each of the three structural identifiers has its own peculiarities. The InChI identifiers are unique, i.e. there is only one standard InChI identifier that can be assigned to a given chemical through the available standardization algorithm. Conversely, different algorithms for canonicalization of SMILES have been produced. As a consequence, different SMILES can be associated with the same chemical – for instance, the chemical Ethylphosphonyl difluoride (AG 23) can be written as CC[P](F)(F)=O or CCP(=O)(F)F. Both InChI and SMILES identifiers can be used to algorithmically rebuild a chemical structure. However, InChI standardization can significantly alter the form of the chemical at hand, while SMILES always preserve all aspects of a chemical structure. Moreover, SMILES are generally more compact than InChI identifiers. Combined together, these characteristics make SMILES the input of choice for most cheminformatics platforms. InChIKey identifiers are generated from InChI identifiers through hashing. They are different from InChI and SMILES identifiers, in that they cannot be used to algorithmically reconstruct a molecule without a resolver that connects them back to the

InChI from which they were generated. However, InChIKey identifiers have the advantage of being compatible with search engines. Hence, they can be considered as "structure-based registry lookup identifiers."[22–24]

The ninth column provides information on the overlaps between the three CW-control lists object of this work. Specifically, when a given entry of a CW-control list overlaps with an entry of a different list, the overlap column reports the additional lists and associated entry numbers in which that particular chemical or family of chemicals is found. For more information, see the subsection **Overlap Analysis** below.

Finally, the tenth column provides a number of links to resources stored in our websites or elsewhere. The first two links allow the downloading of manually curated 2D structures from our website, as SDfiles (SDF) as well as Marvin documents (MRV). The SDF format is widely used in cheminformatics and computational chemistry. The MRV format is a format developed by the company ChemAxon (https://chemaxon.com/). MRV files can be opened with ChemAxon's freely available Marvin Sketch. In our downloadable MRV files,the 2D structures are given in exactly the same orientation and coloring as in the images provided in the sixth column. The third link allows the interactive visualization of the three-dimensional (3D) structures of the entries through NLM's iCn3D. Finally, the fourth link gives access to the NIST Chemistry WebBook cards, consisting in a collection of chemical and physical data compiled by NIST, including, *inter alia*, analytical spectra and physicochemical data.

Clustering of CW-control lists based on chemical structure. Based on 166-bit MACCS keys molecular fingerprints, we clustered the entries of the three CW-control lists according to their chemical structures. The MACCS keys fingerprints were chosen for this analysis because they afford an effective division of chemicals based on their structural features. The results are shown in **Figure 3** in the form of a fan plot, in which the branch lengths are proportional to the distance between the fingerprints of two chemicals – a larger format version of the figures is available in the **Supporting Information** (**Figure S1**).

We manually curated the plot annotating it with information on the chemical characteristics of the various clusters (large-font labels), assigning one label to each of the 11 main sections of the fan. Moreover, we annotated the plot with information on the CWA classes to which the members of the various sections belong (small-font labels).

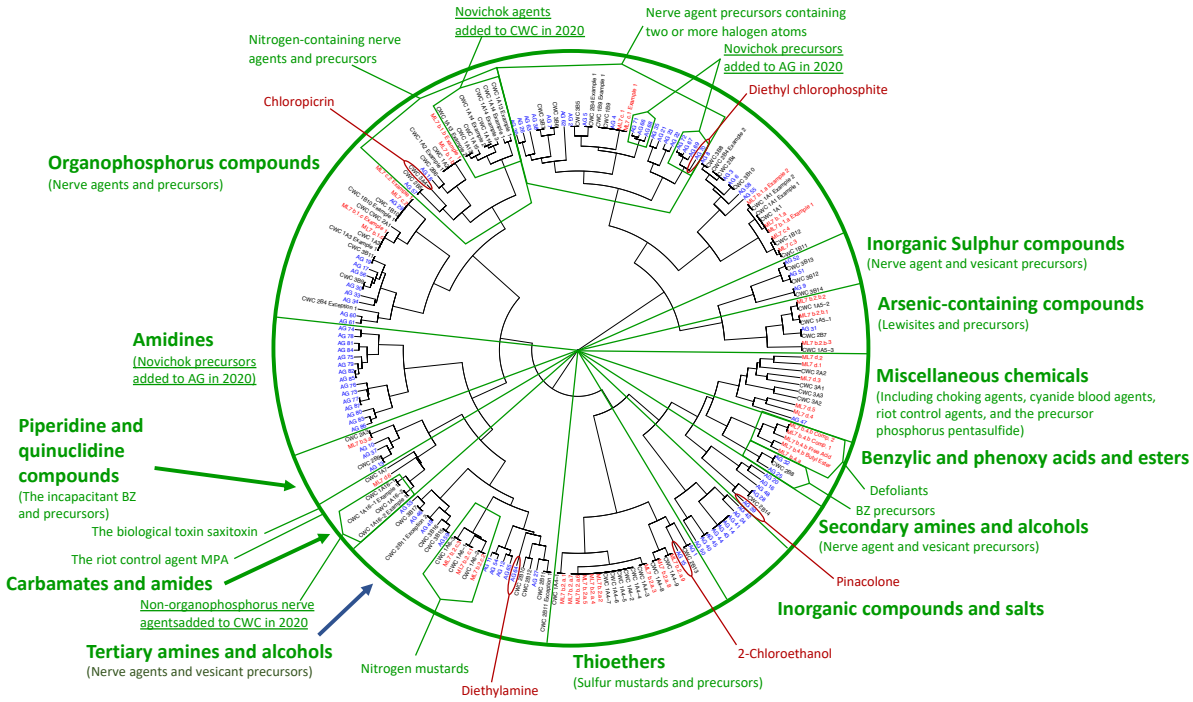


Figure 3. Fan plot showing the clustering of CW-control lists based on chemical structures. CWC schedule entries are shown in black, AG precursors list entries are shown in blue, and WA ML7 entries are shown in red. Large-font green labels indicate the chemical features characterizing the various clusters. Small-font green labels indicate the CWA class to which the entries comprised in the clusters belong. Subclusters within larger clusters are encased in pentagon shapes. The clusters that encompass the entries recently added to CWC schedule 1 and the AG precursors list are underlined. Outlier chemicals, i.e. chemicals that do not completely fit into the definition of the cluster to which they belong, are marked with dark red labels. A larger format version of the figure is provided in the **Supporting Information** (**Figure S1**).

The largest section comprises organophosphorus compounds, functionally classified as nerve agents and precursors for their synthesis. Within this section, two notable branches with common characteristics can be identified. The first one features nitrogen-containing nerve agents, including the Novichok agents added to CWC Schedule 1 in 2020. The second one features nerve agent precursors containing two or more halogen atoms, including the Novichok precursors added to the AG chemical precursors list in 2020.

Walking through the fan plot, in counterclockwise manner, the following groups of chemicals are found: a series of amidines, which were added to the AG chemical precursors list in 2020 as Novichok precursors; piperidine and quinuclidine compounds, including the incapacitant BZ and precursors for its synthesis; two amides (the biological toxin saxitoxin and the riot control agent N-Nonanoylmorpholine) and a family of carbamates, which are non-organophosphorus nerve agents added to CWC Schedule 1 in 2020; tertiary amine and alcohols, including nitrogen mustards as well as precursors for the synthesis of nerve agents and vesicants; thioethers, including sulfur mustards and precursors for their synthesis; inorganic compounds and salts used as precursors for a variety of CWAs; secondary amines and alcohols, used as precursors for the synthesis of nerve agents and vesicants; benzylic and phenoxy acids and esters, including defoliants as well as precursors for the synthesis of BZ; a section containing miscellaneous chemicals featuring a variety of chemical groups and belonging to a variety of CWA classes, including choking agents, blood agents, riot control agents, and precursors for the synthesis of nerve agents; arsenic-containing compounds, including lewisites and precursors for their synthesis; and inorganic sulfur compounds, employed as nerve agent and vesicant precursor.

In the fan plot there are only five outliers, which have chemical characteristics different from the compounds with which they are clustered. These are: the Novichok precursor diethyl chorophosphite (AG 70), which, although clustering with organophosphorus compound with two or more halogen atoms, has only one halogen atom; the choking agent chloropicrin (CWC 3A4), which, although clustering with nitrogen-containing organophosphorus compounds, does not feature phosphorus or nitrogen atoms; the nerve agent precursor diethylamine (AG 64), which, although clustering with tertiary amines, is a secondary amine; the vesicant precursor 2-chloroethanol (AG 15), which, although clustering with thioethers, does

not feature a sulfur atom; and the nerve agent precursor pinacolone (AG 39), which, although clustering with secondary alcohols and amines, is a ketone.

Overlap analysis: identification of CWAs and precursors covered by more than one CW-control list. As outlined below, we performed an overlap analysis to identify the intersections between the three CW-control lists. In particular, we identified discrete chemicals and families of chemicals found in more than one list. Moreover, we identified discrete chemicals of one list that fall within the coverage of a family of chemicals featured in another list. In our curated tables, the overlaps are annotated in the ninth column, where, for each entry of a CW-control list, the overlapping entries of the other two CW-control lists are provided. Moreover, in **Table S4**, the three CW-control lists are synoptically presented side-by-side in one document, hence providing an overarching view of their intersections.

Identical discrete chemicals and identical families. Based on ECFP6 extended-connectivity circular molecular fingerprints, we performed an overlap analysis in order to identify identical discrete chemicals and identical families found in more than one CW-control list. As mentioned, the ECFP6 fingerprints were chosen for this analysis because they yield identical fingerprints exclusively in the case of chemical identity. As a consequence, by removing from the distance matrix all entries that do not have a value of zero in at least two columns of the matrix (there will always be one zero, in the column where the entry is compared with itself, i.e. along the diagonal of the matrix), one is left with a subset of entries that show exact overlaps with others. The results of our fingerprint-based overlap analysis are shown in **Figure 4** in the form of two fan plots, one featuring overlapping discrete chemicals (panel A) and the other showing overlapping families of chemicals (panel B). As it can be seen from the figure, there are two discrete chemicals that overlap across the three CW-control lists. Namely, the two universally overlapping discrete chemicals are: methylphosphonyldifluoride (DF), which is listed as AG 4, CWC 1B9 Example 1, and ML7 c.1 Example 1, and O-ethyl O-2-diisopropylaminoethyl methylphosphonite (QL), which is listed as AG 29, CWC 1B10 Example 1, and ML7 c.2 Example 1. Moreover, there are 20 discrete chemicals are found in the CWC schedules and the AG precursors list, but not the WA ML7 list, and 22 discrete chemicals that are found in

the CWC schedules and the WA ML7 list, but not in the AG precursors list. Finally, there are 5 overlapping families, which can be found in CWC schedules as well as the WA ML7 list.

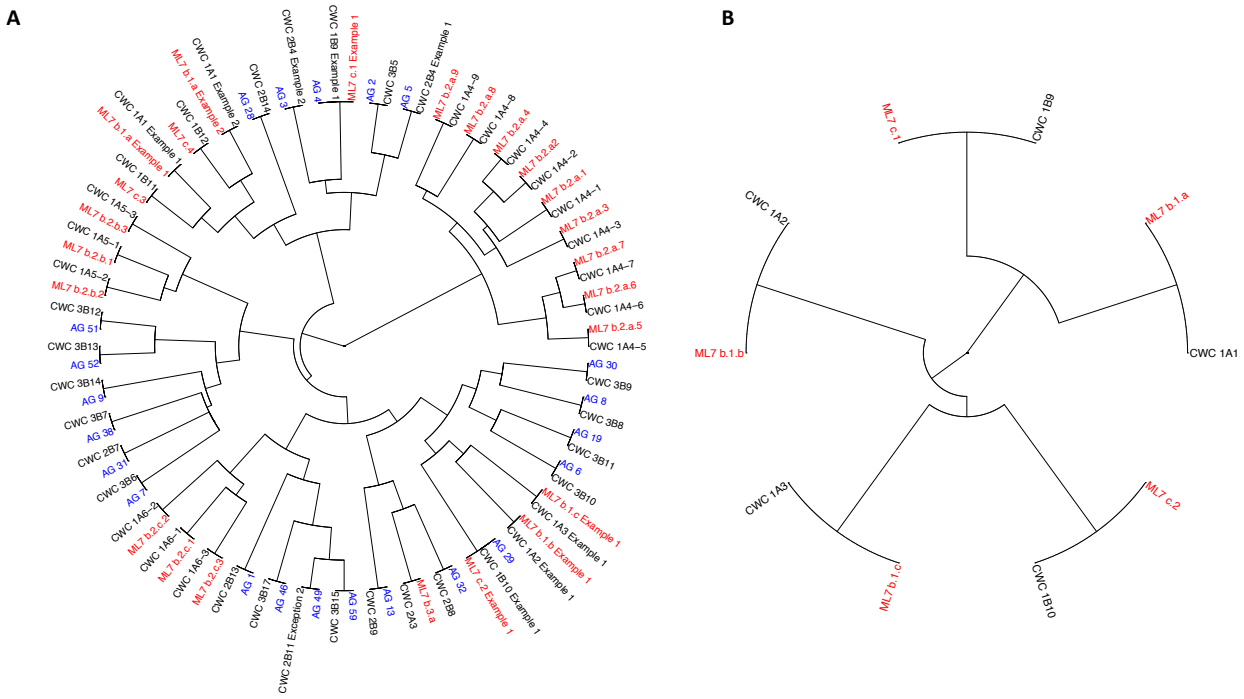


Figure 4. Fan plot showing the overlaps between identical discrete chemicals (panel A) and identical families of chemicals (panel B) in the three CW-control lists under study. CWC schedules entries are shown in black, AG precursors list entries are shown in blue, and WA ML7 entries are shown in red.

Discrete chemicals comprised within chemical families. Beyond overlaps between identical entries (identical discrete chemicals and identical families), a second type of overlaps can be found within CW-control lists: discrete chemicals of one list that are comprised within the scope of a family of chemicals featured in a second list. In order to identify them, we performed an overlap analysis with ChemAxon's Instant JChem database system. The results of this analysis are shown in **Figure 5, Panel A**. As it can be seen from the figure, there is one chemical that is listed as one discrete chemical in the AG precursors list that is also covered at the same time by CWC schedule and WA ML7 families. Namely, this chemical is ethylphosphonyl difluoride, which is listed as AG 23 and is also covered by family CWC 1B9 and family

ML7 c.1. Moreover, there are 18 additional discrete chemicals listed in the AG chemical precursors list that are also covered by a total of six families listed in part B of CWC Schedule 2.

A bar graph showing the number of chemicals found across multiple CW-control lists is provided in **Figure 5, panel B**. In the paragraphs below, we provide further qualitative details on the intersections between the three CW-control lists.

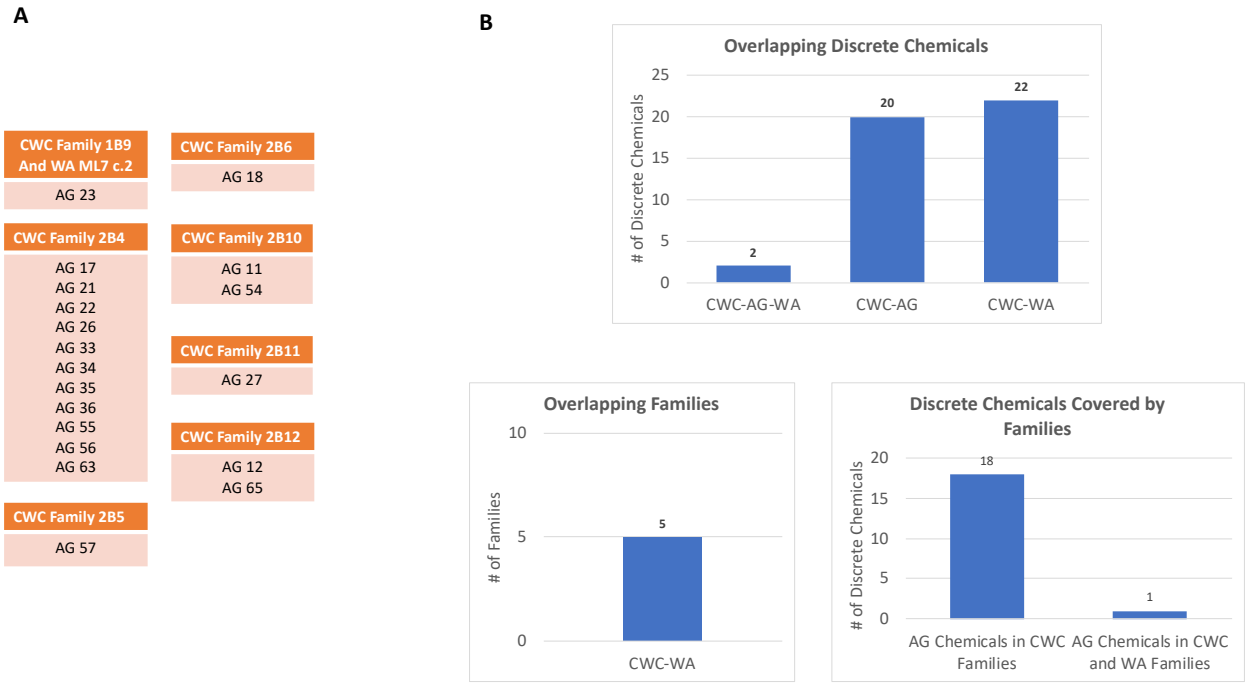


Figure 5. Panel A: discrete individual chemicals of the AG precursors list that fall within the scope of families of chemicals of the CWC schedules and/or WA ML7. Panel B: bar graphs summarizing the number of overlapping chemicals found across multiple CW-control lists.

Overlaps between the CWC schedules and the WA ML7. With the exception of saxitoxin and ricin (CWC entry numbers 1A7 and 1A8) and the new additions of chemical families and associated chemical examples (CWC entry numbers 1A13, 1A14, 1A15, 1A16-1, and 1A16-2), all CWC Schedule 1A entries, including those covering families of chemicals, are also covered by WA ML7 b.1 and WA ML7 b.2. All CWC Schedule 1B entries, including those covering families of chemicals, are covered by WA ML7 c. In

CWC Schedule 2A, the incapacitant BZ (CWC 2A3 and WA ML7 b.3.a) is the only overlapping chemical. There exist no other overlaps between the remaining CWC schedules and the WA ML7 entries.

Overlaps between the CWC schedule and AG chemical precursors list. Of the 87 entries of the AG precursors list, 41 are also listed in Part B of one of the three CWC schedules, either as individual chemicals, or in virtue of being comprised in one of the families of chemicals listed in the CWC schedules. The breakdown is as follows: 3 chemicals in the AG precursors list are also covered by CWC Schedule 1B; 25 chemicals in the AG precursors list are also covered by the CWC Schedule 2B; 12 chemicals in the AG precursors list are also covered by CWC Schedule 3B; 1 chemical in the AG precursors list, namely N,N-diethylaminoethanol (AG 49), is listed in CWC Schedule 2B as an exception to entry 2B11 – i.e., although the chemical would fall within the scope of entry CWC 2B11, it is excluded from its coverage, due to its extensive legitimate uses.

Overlaps across all three lists. There exist only three chemicals that are covered by all three lists, where they either are listed as individual chemicals or are comprised within the scope of one of the families of chemicals. In particular, as we have seen, the three overlapping chemicals are: methylphosphonyldifluoride (DF), which is listed as AG 4, CWC 1B9 Example 1, and WA ML7 c.1 Example 1; ethylphosphonyl difluoride, which is listed as AG 23 and is also covered by family CWC 1B9 and family WA ML7 c.1; and O-ethyl O-2-diisopropylaminoethyl methylphosphonite (QL), which is listed as AG 29, CWC 1B10 Example 1, and WA ML7 c.2 Example 1. Beyond these three universally overlapping chemicals, there aren't any other overlaps between the AG chemical precursors list and the WA ML7.

DISCUSSION

Chemicals are better described by structures than names or registry numbers. Indeed, by simply looking at a registry number, one does not garner any information about the chemical to which it refers. Systematic IUPAC names do contain all necessary information to reconstruct the molecular structure of the chemical that they describe. However, a great deal of familiarity with organic chemistry nomenclature is required to picture the molecular structure of a chemical just by reading its name. As molecules become larger and

more complex, this mental operation becomes increasingly more daunting. For these reasons, we have annotated the CW-control lists from three key international frameworks for the control of chemical weapons – namely the CWC Schedules, the AG precursors list, and the WA ML7 – with manually curated 2D structures and links to interactive 3D structures. This annotation is of particular importance for the CW-control field, which sits at the intersection of science, policy, and international security and involves the work of chemists as well as security and policy experts. Molecular structures are key instruments to facilitate the communication between chemists and policymakers. A single letter difference in a chemical name can account for an important difference in the chemical structure, marking the watershed between chemicals that are controlled and those that are not. As an example, the chemical N,N-diisopropyl-2-chloroethylamine is a precursor for the synthesis of the nerve agent VX and is covered by a family of chemicals listed in part B of CWC Schedule 2 (entry CWC 2B10) as well as the AG precursors list (entry AG 11). Conversely, the chemical N,N-diisopropyl-2-chloroethylamide is not a precursor for the synthesis of any known CWA agent and, as a consequence, does not appear in CW-control lists. The difference between the two chemicals is more clearly grasped from their chemical structures than from their names (**Figure 6**). The success of the Science for Diplomats initiative of the OPCW and its structurally annotated CWC schedules testifies to the importance of using molecular structures to facilitate the interaction between chemists and policymakers.[10]

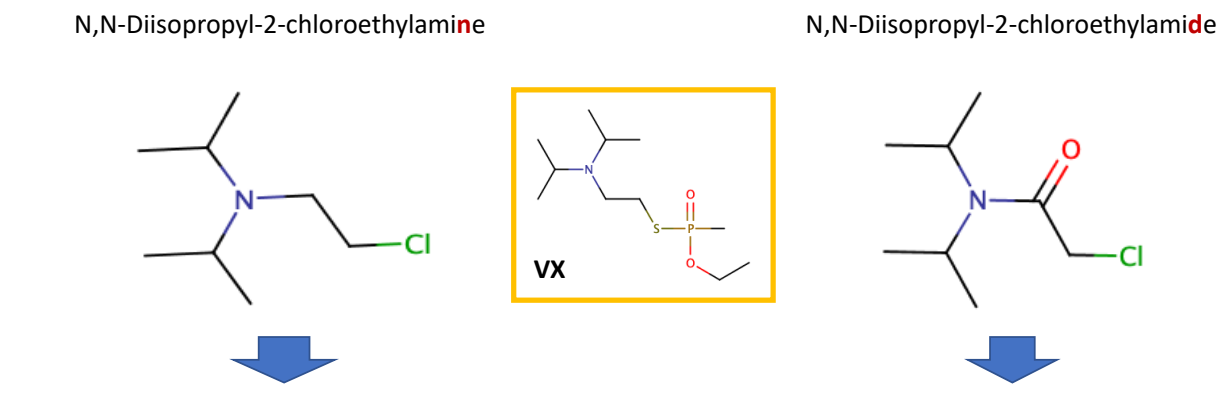


Figure 6. Two chemicals whose names differ for one letter only. One is precursor for the synthesis of the nerve agent VX and is covered by CWC Schedule 2 and the AG precursors list. The other one is not a known precursor to CWAs and is not found in CWC schedules or the AF precursors list.

Beyond featuring images of the chemicals, our annotated tables provide chemists and cheminformaticians with access to the 2D molecular structures via downloadable files, in both SDF and MRV format, as well as structural identifiers in three different formats, namely SMILES, standard InChI, and standard InChIKey – see the description of columns 7, 8, and 10 in the **Results** section for the differences between the various formats and the importance of including them in the annotated tables. These structures can be used as inputs for computational chemistry and cheminformatics software, thus enabling, *inter alia*, the calculation of molecular properties and descriptors, the compilation of molecular databases, and the conduction of molecular fingerprinting analyses such as those at the basis of our structural clustering of the chemicals featured in CW-control lists (**Figure 3**) and our overlap analyses (**Figures 4 and 5**). In the current CW proliferation landscape – which is characterized by a shift from the traditional use of CWs in large quantities as weapons of mass destruction to a use in smaller quantities for counterinsurgency and targeted assassinations[4,5] – it is becoming increasingly important to have an overarching view of all the CWA and precursors covered by international frameworks and national legislation. Emerging chemical technologies are being developed,[8] and chemical proliferators are likely to attempt to leverage them. The CW-control lists have been recently expanded to keep up with the current threat, and they will likely continue to be amended as chemical new chemicals of concern emerge. In this context, our clustering and overlap analyses provide chemists and policymakers with an encompassing view of the controlled chemicals, assisting in the identification of gaps that need to be narrowed and supporting a coordinated expansion of the chemical space covered by the various frameworks.

In the two sections below, first, to showcase the importance of structural annotation, we examine a discrepancy in the WA ML7 that we uncovered while annotating the tables and that needs addressing.

Subsequently, to illustrate an application of our structural clustering and overlap analyses, we will take a closer look at the recent updates of the CWC Schedule 1 and the AG precursors list.

A discrepancy in WA ML7: Agent Orange. The presence of CAS registry numbers in CW-control lists is of great help to overcome one of the main difficulties provided by chemical names: the existence of a myriad of synonyms. Although multiple CAS registry numbers are associated with different variants of a chemical (e.g. salts, isotopic variants, isomers, tautomers, etc.), each specific variant is associated with a unique registry number.[9]

However, as mentioned, registry numbers do not convey any information on a chemical at a glance. As an example of this, we can point out to a discrepancy in WA ML7 that we identified while curating the CW-control lists. The defoliant Agent Orange is listed in WA ML7 under entry b.4.b. The name given for this entry in WA ML7 is "2,4,5-trichlorophenoxyacetic acid mixed with 2,4-dichlorophenoxyacetic acid (Agent Orange)." The CAS registry number given to the entry in WA ML7 is 39277-47-9. However, as shown in **Figure 7**, the name and the CAS number do not match: while in the given name the two components of the mixture appear in their free carboxylic acid form, in the chemical associated with the given CAS registry number they appear in their butyl ester form. Below the mixture, WA ML7 also lists the two separate components. As shown in **Figure 7**, in the case of the components both names and registry numbers refer to the free carboxylic acid form. Hence, it seems evident that the intention was to have mixture in its free carboxylic acid form, as indicated by the name, rather than in the butyl ester form, as indicated by the CAS registry number. The issue can be easily addressed by changing the CAS registry number provided for the mixture with the one correctly associated with the free carboxylic acid form: 8015-35-8.

The detection of discrepancies while curating chemical databases is not unique. In fact, as outlined by Williams and coworkers, challenges are common in chemical database curation, and efforts are needed to improve the quality of the chemical data in the public domain.[25] The cruciality of cleaning chemical structures in databases prior to performing cheminformatics analyses and computational chemistry studies

has been clearly laid out by Tropsha and coworkers. Importantly, the authors underscored how rigorous computational analyses can help detect and correct inconsistencies in chemical databases, for instance erroneous biological activity data. Such situations bear a striking similarity with our detection of the mismatch between the name and CAS registry number attributed to Agent Orange in WA ML7.[26]

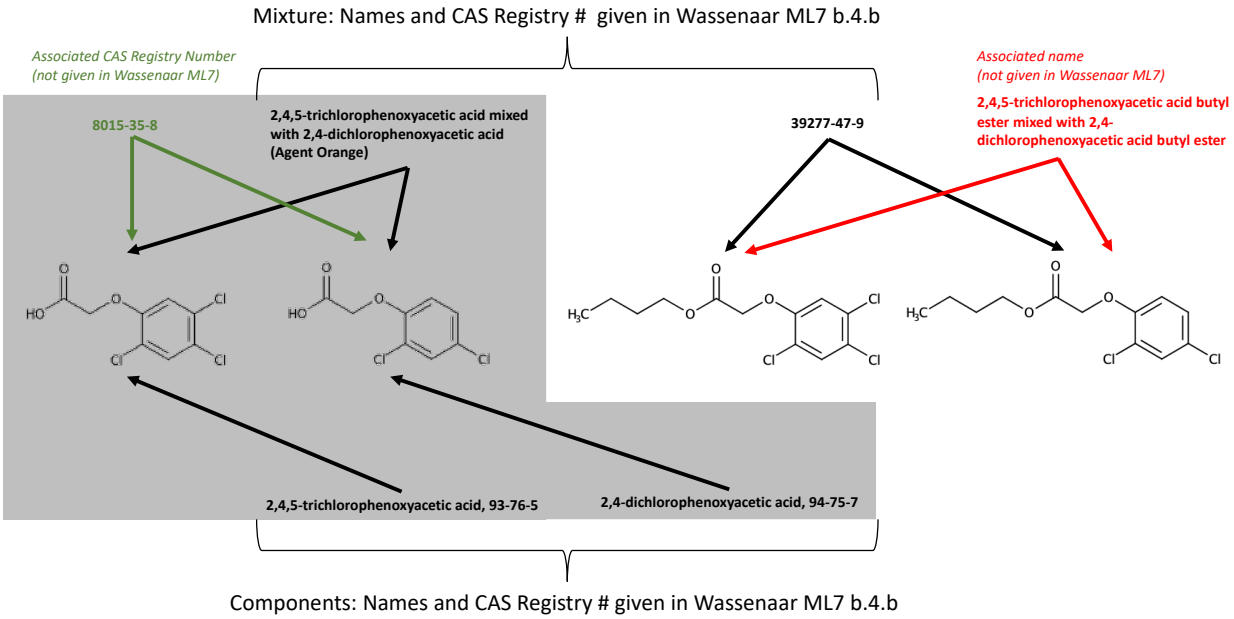


Figure 7. A discrepancy in WAML7. The name and CAS registry number given for the Agent Orange mixture do not match.

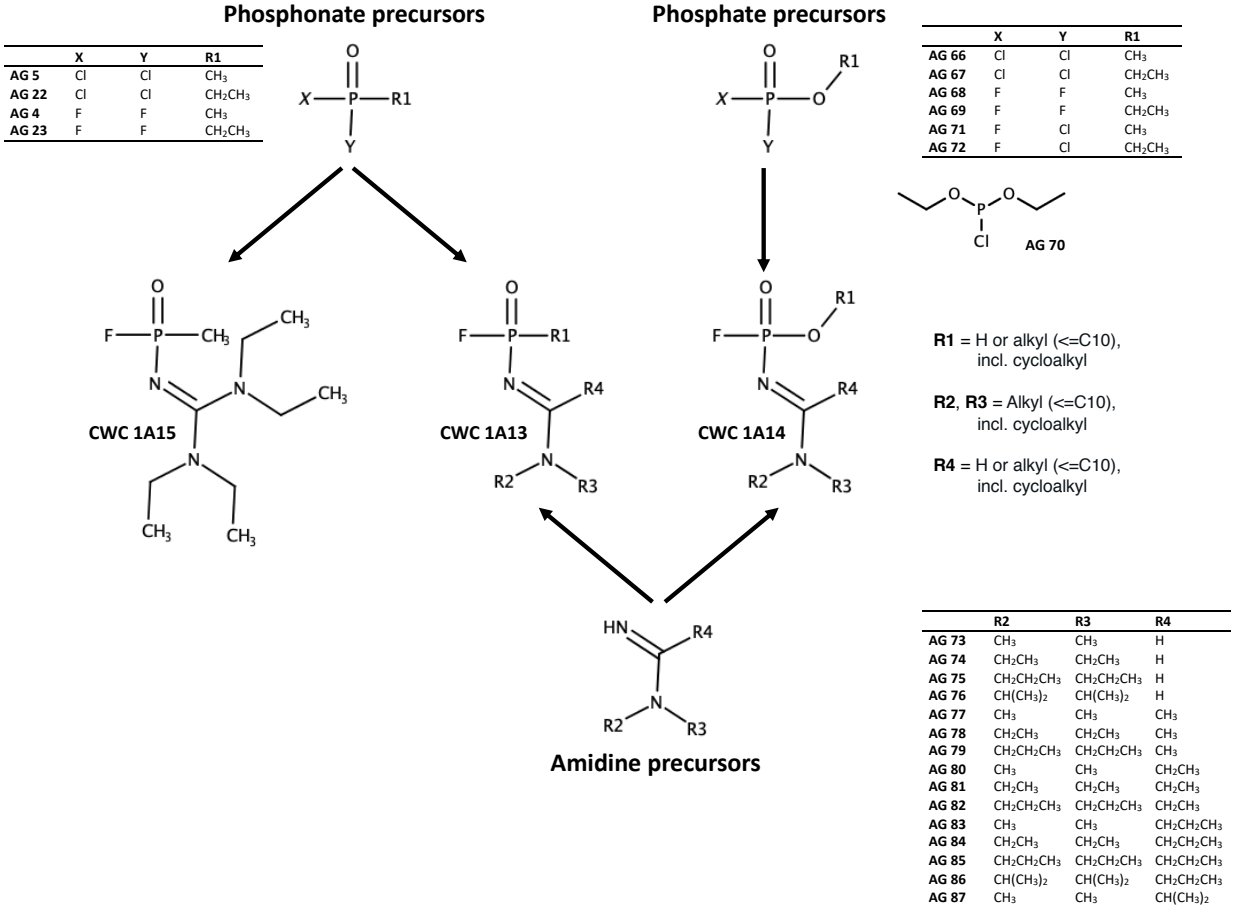
Recent amendments of the CWC Schedule 1 and the AG precursors list. On March 4, 2018, a nerve agent belonging to the Novichok class, developed in the Soviet Union during the Cold War, was deployed in Salisbury, UK. The agent was reportedly applied to a door handle in the attempt to assassinate Sergei Skripal, a former Russian intelligence agent who had defected to the West, and his daughter. Although the Skripals and the law enforcement officers that responded to the incident survived the attack, Dawn Sturgess, a resident of the nearby Amesbury who accidentally came into contact with the same agent after her boyfriend found a discarded faux perfume bottle with which it was filled, tragically died. The Salisbury facts set into motion a series of events that eventually led to the amendment of the CWC Schedule 1 and

the subsequent addition of chemicals to the AG precursors list. It also triggered a number of literature reviews and research studies on Novichok agents.[27–34]

The CWC Schedule 1 amendment was voted by CWC Conference of State Parties in November 2020, and it entered into force in June 2020. Its final form is the result of the combination of two proposals, a joint proposal submitted by the United States, Canada, and the Netherlands, which sought to cover two large families of Novichok nerve agents, and a proposal from the Russian Federation, which sought to cover a narrower set of specific Novichok agents (only one of which was not already covered by the joint proposal's families), as well as carbamates researched, although never developed, as CWAs by the United States during the Cold War.[12,13] Following the approval of the CWC amendment, in February 2020, the AG added a 22 Novichok precursors to its list.

The CWC amendment comprises the Novichok families identified by entries CWC 1A13 and CWC 1A14, with a number of discrete examples, and the single discrete chemical CWC 1A15 (**Figure 8**). As it is evident from **Figure 3**, all these Novichok agents form a well-defined cluster within a larger group of nitrogen-containing nerve agents. The addition of both of these families stems from the joint proposal. Specific examples are listed as well, some deriving from the joint proposal and other deriving from the Russian proposal. The examples added by the Russian proposal comprise three Novichok agents described by Vil Mirzayanov, an analytical chemist formerly involved in the FOLIANT chemical weapons program of the Soviet Union, in his memoir book *State Secrets*.[35] Namely, these are agent A-230 (CWC 1A13 Example 2) and the agents A-232 and A-234 (CWC 1A14 Examples 2 and 3), the latter of which is the chemical reportedly employed in the Salisbury incident. Chemically, the two families are phosphonates (CWC 1A13) and phosphates (CWC 1A14), both featuring amidine substituents. The presence of the amidine substituent is an element that sets these Novichok agents apart from other nerve agents already covered by the CWC schedules. Moreover, the chemicals covered by CWC 1A14 further differ from other nerve agents in that they feature a phosphate, not a phosphonate scaffold. The addition of a further Novichok agent to CWC Schedule 1, namely entry CWC 1A15 (described by Mirzayanov as agent A-242), stems from the Russian proposal. This chemical is a phosphonate-based Novichok, like those covered by family

CWC 1A13. However, CWC 1A15 features a guanidine, rather than an amidine substituent. An analogue of A-242, featuring the same guanidine substituent but endowed with a phosphate scaffold has been described by Mirzayanov with the name of code name A-262. It might be worth exploring the possibility of adding this chemical to CWC Schedule 1.[12]



	X	Y	R1
AG 5	Cl	Cl	CH_3
AG 22	Cl	Cl	CH_2CH_3
AG 4	F	F	CH_3
AG 23	F	F	CH_2CH_3

	X	Y	R1
AG 66	Cl	Cl	CH_3
AG 67	Cl	Cl	CH_2CH_3
AG 68	F	F	CH_3
AG 69	F	F	CH_2CH_3
AG 71	F	Cl	CH_3
AG 72	F	Cl	CH_2CH_3

	R2	R3	R4
AG 73	CH_3	CH_3	H
AG 74	CH_2CH_3	CH_2CH_3	H
AG 75	$CH_2CH_2CH_3$	$CH_2CH_2CH_3$	H
AG 76	$CH(CH_3)_2$	$CH(CH_3)_2$	H
AG 77	CH_3	CH_3	CH_3
AG 78	CH_2CH_3	CH_2CH_3	CH_3
AG 79	$CH_2CH_2CH_3$	$CH_2CH_2CH_3$	CH_3
AG 80	CH_3	CH_3	CH_2CH_3
AG 81	CH_2CH_3	CH_2CH_3	CH_2CH_3
AG 82	$CH_2CH_2CH_3$	$CH_2CH_2CH_3$	CH_2CH_3
AG 83	CH_3	CH_3	$CH_2CH_2CH_3$
AG 84	CH_2CH_3	CH_2CH_3	$CH_2CH_2CH_3$
AG 85	$CH_2CH_2CH_3$	$CH_2CH_2CH_3$	$CH_2CH_2CH_3$
AG 86	$CH(CH_3)_2$	$CH(CH_3)_2$	$CH_2CH_2CH_3$
AG 87	CH_3	CH_3	$CH(CH_3)_2$

Figure 8. Novichok agents and precursors for their synthesis recently added to the CWC schedules and the AG precursors list.

The chemicals added to the AG precursors list are all precursors of the aforementioned Novichok agents (**Figure 8**). Specifically, they are precursors to agents that fall within the scope of the families CWC 1A13 and CWC 1A14. They can be divided into two categories: phosphate precursors and amidine precursors (**Figure 8**). The phosphate precursors (AG 66 – AG72) are precursors for the synthesis of CWC 1A14 chemicals. In our clustering analysis, they formed two clusters found within a larger group of nerve agent

precursors endowed with two or more halogen atoms (**Figure 3**). The first cluster (AG 66, AG 68, and AG71) comprises ethyl phosphate derivatives. The second cluster (AG 67, AG 69, and AG 72) comprises methyl phosphate derivatives. A further chemical within the first category is the upstream precursor diethyl chlorophosphite (AG 70), which is clustered together with AG 67, AG 69, and AG 72. It has not been necessary to add precursors for the phosphonate moiety of family CWC 1A13 and the discrete chemical CWC 1A15, as these coincide with the precursors for the synthesis of canonical nerve agents and were already covered by the AG precursors list. As indicated in **Figure 8**, the phosphonate precursors AG 4, AG 5, AG 22, and AG 23 are the direct correspondent of the newly added phosphate precursors. Of note, these phosphonate precursors are also covered by CWC 1B9 (corresponding to WA ML7 c.1) and CWC 2B4. The only difference between the phosphonate and the phosphate precursors is that the latter also include two mixed chlorofluoro derivatives. It might be worth exploring the possibility of adding to the AG precursors list the corresponding chlorofluorophosphonates, which, in any case, are already covered by the CWC 2B4 family. Moreover, it might worth adding to part B of the CWC schedules the phosphate precursors added to the AG list.

The amidines added to the AG precursors list (AG 73 – AG 87) are precursors for the synthesis of both CWC 1A13 and CWC 1A14 chemicals. In our clustering analysis, they are all clustered together in a well-defined section of the fan plot (**Figure 3**). The guanidine precursor for the synthesis of CWC 1A15 (A-242) has also not been added to the AG precursors list. It might be worth exploring the possibility of adding this chemical to the list, although this might not be necessary, as the availability of the guanidine precursor alone is not sufficient to allow for the synthesis of the agent, given the fact that the phosphonate precursors are already controlled chemicals.

It should be noted that phosphates and amidines added to the AG precursors account for only a limited subset of the Novichok agents covered by the CWC amendment. In particular, while the two CWC families 1A13 and 1A14 feature alkyl groups with up to 10 carbon atoms, the AG precursors feature alkyl groups up to two carbon atoms on the phosphate moiety and three carbon atoms on the amidine moiety. It might be worth exploring the possibility of adding a wider coverage of alkyl groups to the AG precursors list,

although the AG additions indeed account for the synthesis of the amidine agents described by Mirzayanov (A-230, A-232, and A-234) and closely related analogs.

Beyond Novichok agents, carbamates researched in the United States during the Cold War were also added to CWC Schedule 1. These additions, which stemmed from the Russian proposal, comprise a family of quaternaries of dimethylcarbamoyloxypyridines and a family of bisquaternaries of dimethylcarbamoyloxypyridines, both listed within entry CWC 1A16, together with two examples. These chemicals form a small cluster, adjacent to two amide compounds covered by the CWC schedules, namely the biological toxin saxitoxin and the riot control agent N-nonanoylmorpholine. It might be worth exploring the possibility of adding precursors for the synthesis of these compounds to the AG precursors list and part B of the CWC schedules.

As always, adding chemicals to control lists requires a thorough assessment of the legitimate uses of the chemicals and the unintended consequences resulting from their addition to control lists. This is true for all dual-use technologies and items with legitimate as well as illegitimate applications, as the goal is to rich the right balance between maximizing security and minimizing the hindrance of progress.[12,36]

SUMMARY AND CONCLUSIONS

To support multilateral efforts to stem CW proliferation, we have curated and structurally annotated CW-control lists from the CWC, AG, and WA. Our research is intended to facilitate the communication between scientific advisors and policymakers as well the work of chemists and cheminformaticians involved in the chemical weapon non-proliferation field. Ultimately, we seek to provide tools to bolster the control of chemical weapons and support the global efforts to rid the world of them.

The curated lists are available as web tables at the Costanzi Research website (https://costanziresearch.com/cw-control-lists/) and are provided in PDF format in the **Supporting Information** (**Table S1 – S3**). The annotations include manually curated 2D structural images, which provide a means to appreciate at a glance the similarities and differences between the different entries, as well as downloadable 2D structures, in two different formats, and three different structural identifiers,

namely SMILES, standard InChI, and standard InChIKey, which are intended to provide a platform for cheminformatics analyses of CWA and precursors. The tables also include links to NCBI's PubChem and NIST's Chemistry WebBook cards, hence providing easy access to a wealth of physicochemical, analytical chemistry, and toxicological information. To showcase the importance of structural annotations, we discuss a discrepancy in a CW-control list covering the defoliant Agent Orange, which we identified through our curation process, and propose a solution to address it.

Moreover, we conducted chemical fingerprinting analyses, through which we clustered the entries of the three CW-control lists under study into structurally related groups (**Figure 3**) and studied the overlaps between the three lists (**Figure 4**, **Figure 5**, and **Table S4**). As an application of this study, we examined the recent updates of the CWC Schedule 1 and the AG precursors list, highlighting the relationships between the two amendments and proposing the possible addition of further chemicals.

Moving forward, we plan on using our structurally annotated tables as the starting base for the development of a cheminformatics database featuring discrete chemicals and Markush structures. Such database will support the control of CWA and precursors by allowing national and international authorities as well as chemical and shipping companies to easily identify, in an automated manner, whether a given chemical is covered by a CW-control list. For instance, through this tool, users will be able to assess whether a given chemical under scrutiny falls within the scope of one of the controlled families of chemicals. Moreover, the tool will automatically establish the equivalence of different variants of chemicals, including, *inter alia*, salts, stereoisomers, tautomers, and isotopically-labelled forms. All these variants have their own chemical names and CAS registry numbers. However, a structural database search, following a standardization of the query, will be able to associate all variants to the same chemical entity. Such an approach is in line with the OPCW Scientific Advisory Board's recommendation of extending the coverage of CW-control lists to all variants of the controlled chemicsals.[9] This is particularly important in the current CW proliferation landscape, in which even chemicals typically synthesized in smaller quantities constitute a threat, and in light of the advancement of chemical technologies.[4,5,8]

ACKNOWLEDGEMENTS

SC, CS, and BH acknowledge support for from American University. The authors declare no conflicts of interest.

SUPPORTING INFORMATION

The following material is provided as supporting information, in PDF format: (1) **Supporting Information Figures and Tables**, which include: **Figure S1**, a larger resolution version of **Figure 3**; **Table S1,** Annotated Chemical Weapons (CWC) Convention Schedules; **Table S2,** Annotated Australia Group (AG) Chemical Weapons Precursors List; **Table S3**, Annotated Wassenaar Arrangement (WA) Munitions List 7 (ML7); **Table S4**, Synoptic Table. (2) downloadable version of the tables, in SDF, CSV, and Microsoft Excel format.

This material is available free of charge via the Internet at https://pubs.acs.org/doi/abs/10.1021/acs.jcim.0c00896.

REFERENCES

(1) Costanzi, S. Chemical Warfare Agents. *Kirk-Othmer Encyclopedia of Chemical Technology* **2020**, 1–32.

(2) Ganesan, K.; Raza, S. K.; Vijayaraghavan, R. Chemical Warfare Agents. *Journal of pharmacy and bioallied sciences* **2010**, *2* (3), 166.

(3) Pitschmann, V. Overall View of Chemical and Biochemical Weapons. *Toxins* **2014**, *6*, 1761–1784.

(4) Hersman, R. K.; Pittinos, W. *Restoring Restraint: Enforcing Accountability for Users of Chemical Weapons*; Rowman & Littlefield, 2018.

(5) Hersman, R. K.; Claeys, S. *Rigid Structures, Evolving Threat: Preventing the Proliferation and Use of Chemical Weapons*; Center for Strategic & International Studies, 2019.

(6) Koblentz, G. D. Chemical-Weapon Use in Syria: Atrocities, Attribution, and Accountability. *The Nonproliferation Review* **2019**, *26*, 575–598.

(7) German Federal Government, Statement by the Federal Government on the Navalny case https://www.bundesregierung.de/breg-en/news/statement-by-the-federal-government-on-the-navalny-case-1781882 (accessed Sep 6, 2020).

(8) Kosal, M. E. Emerging Chemical and Biological Technologies: Security & Policy Challenges. In *Responsible Conduct in Chemistry Research and Practice: Global Perspectives*; ACS Publications, 2018; pp 51–68.

(9) Costanzi, S.; Koblentz, G. D.; Cupitt, R. T. Leveraging Cheminformatics to Bolster the Control of Chemical Warfare Agents and Their Precursors. *Strategic Trade Review* **2020**, *6*, 69– 92.

(10) Pontes, G.; Schneider, J.; Brud, P.; Benderitter, L.; Fourie, B.; Tang, C.; Timperley, C. M.; Forman, J. E. Nomenclature, Chemical Abstracts Service Numbers, Isomer Enumeration, Ring Strain, and Stereochemistry: What Does Any of This Have to Do with an International Chemical Disarmament and Nonproliferation Treaty? *Journal of Chemical Education* **2020**.

(11) OPCW. Chemical Weapons Convention, Article II, Definitions and Criteria https://www.opcw.org/chemical-weapons-convention/articles/article-ii-definitions-and-criteria (accessed Aug 6, 2020).

(12) Costanzi, S.; Koblentz, G. D. Controlling Novichoks after Salisbury: Revising the Chemical Weapons Convention Schedules. *The Nonproliferation Review* **2019**, 1–14.

(13) Costanzi, S.; Koblentz, G. D. Updating the CWC: How We Got Here and What Is Next. *Arms Control Today* **2020**, *50*, 16–20.

(14) Timperley, C. M.; Forman, J. E.; \AAas, P.; Abdollahi, M.; Benachour, D.; Al-Amri, A. S.; Baulig, A.; Becker-Arnold, R.; Borrett, V.; Carino, F. A. Advice from the Scientific Advisory Board of the Organisation for the Prohibition of Chemical Weapons on Riot Control Agents in Connection to the Chemical Weapons Convention. *RSC advances* **2018**, *8*, 41731–41739.

(15) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.

(16) Szöcs, E.; Stirling, T.; Scott, E. R.; Scharmüller, A.; Schäfer, R. B. Webchem: An R Package to Retrieve Chemical Information from the Web. *Journal of Statistical Software* **2020**, *93*, 1–17.

(17) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic acids research* **2019**, *47* (D1), D1102–D1109.

(18) Guha, R. Chemical Informatics Functionality in R. *Journal of Statistical Software* **2007**, *18*.

(19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of chemical information and modeling* **2010**, *50*, 742–754.

(20) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of chemical information and computer sciences* **2002**, *42*, 1273–1280.

(21) Paradis, E.; Schliep, K. Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R. *Bioinformatics* **2019**, *35*, 526–528.

(22) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of cheminformatics* **2015**, *7*, 23.

(23) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(24) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of chemical information and computer sciences* **1989**, *29*, 97–101.

(25) Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. *Drug discovery today* **2012**, *17*, 685–701.

(26) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of chemical information and modeling* **2010**, *50*, 1189–1204.

(27) Costanzi, S.; Machado, J.-H.; Mitchell, M. Nerve Agents: What They Are, How They Work, How to Counter Them. *ACS Chem. Neurosci.* **2018**, *9*, 873–885. https://doi.org/10.1021/acschemneuro.8b00148.

(28) Kloske, M.; Witkiewicz, Z. Novichoks–The A Group of Organophosphorus Chemical Warfare Agents. *Chemosphere* **2019**, *221*, 672–682.

(29) Franca, T. C.; Kitagawa, D. A.; Cavalcante, S. F. de A.; da Silva, J. A.; Nepovimova, E.; Kuca, K. Novichoks: The Dangerous Fourth Generation of Chemical Weapons. *International journal of molecular sciences* **2019**, *20*, 1222.

(30) Bhakhoa, H.; Rhyman, L.; Ramasami, P. Theoretical Study of the Molecular Aspect of the Suspected Novichok Agent A234 of the Skripal Poisoning. *Royal Society open science* **2019**, *6*, 181831.

(31) Imrit, Y. A.; Bhakhoa, H.; Sergeieva, T.; Danés, S.; Savoo, N.; Elzagheid, M. I.; Rhyman, L.; Andrada, D. M.; Ramasami, P. A Theoretical Study of the Hydrolysis Mechanism of A-234; the Suspected Novichok Agent in the Skripal Attack. *RSC Advances* **2020**, *10*, 27884–27893.

(32) Carlsen, L. After Salisbury Nerve Agents Revisited. *Molecular informatics* **2019**, *38*, 1800106.

(33) Harvey, S. P.; McMahon, L. R.; Berg, F. J. Hydrolysis and Enzymatic Degradation of Novichok Nerve Agents. *Heliyon* **2020**, *6*, e03153.

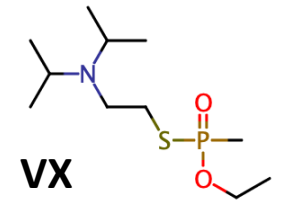
(34) Zeman, J.; Vetchỳ, D.; Pavloková, S.; Franc, A.; Pitschmann, V. Unique Coated Neusilin Pellets with a More Distinct and Fast Visual Detection of Nerve Agents and Other Cholinesterase Inhibitors. *Journal of Pharmaceutical and Biomedical Analysis* **2020**, *179*, 113004.

(35) Mirzayanov, V. S. *State Secrets: An Insider's Chronicle of the Russian Chemical Weapons Program*; Outskirts Press, Incorporated, 2009.

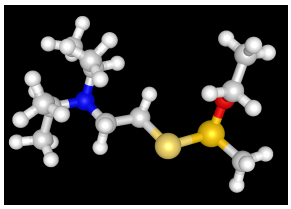
(36) Atlas, R. M.; Dando, M. The Dual-Use Dilemma for the Life Sciences: Perspectives, Conundrums, and Global Solutions. *Biosecurity and bioterrorism: biodefense strategy, practice, and science* **2006**, *4*, 276–286.

For Table of Contents Only

Structures



VX



Chemical Information Cards

- PubChem
- NIST Chemistry WebBook

Structural Identifiers
SMILES, InChI, InChIKey

Fingerprint Analyses