

This is a pre-print manuscript of an article published in the journal *Pure and Applied Chemistry*.

The published version of this article is available at the De Gruyter website:

<https://www.degruyter.com/document/doi/10.1515/pac-2021-1107/html>

**How to cite:** Costanzi, Stefano, Slavick, Charlotte K., Abides, Joyce M., Koblenz, Gregory D., Vecellio, Mary and Cupitt, Richard T. "Supporting the fight against the proliferation of chemical weapons through cheminformatics" *Pure and Applied Chemistry*, vol. , no. , 2022. <https://doi.org/10.1515/pac-2021-1107>

## Supporting the fight against the proliferation of chemical weapons through cheminformatics

Stefano Costanzi,<sup>a†</sup> Charlotte K. Slavick,<sup>a</sup> Joyce M. Abides,<sup>a</sup> Gregory D. Koblenz,<sup>b</sup> Mary Vecellio,<sup>c§</sup> and Richard T. Cupitt<sup>c</sup>

<sup>a</sup>*Department of Chemistry, American University, 4400 Massachusetts Avenue, NW, Washington, DC 20016, USA* <sup>b</sup>*Schar School of Policy and Government, George Mason University, 3351 Fairfax Drive, Arlington, VA 22201, USA* <sup>c</sup>*The Henry L. Stimson Center, 1211 Connecticut Ave, NW, Washington, DC 20036, USA*

**Abstract:** International frameworks have been put in place to foster chemical weapons nonproliferation and disarmament. These frameworks feature lists of chemicals that can be used as chemical weapons or precursors for their synthesis (CW-control lists). In these lists, chemicals of concern are described through chemical names and CAS Registry Numbers®. Importantly, in some CW-control lists, some entries, rather than specifying individual chemicals, describe families of related chemicals. Working with CW-control lists poses challenges for frontline customs and export control officers implementing these frameworks. Entries that describe families of chemicals are not easy to interpret, especially for non-chemists. Moreover, synonyms and chemical variants complicate the issue of checking CW-control lists through names and registry numbers. To ameliorate these problems, we have developed a functioning prototype of a cheminformatics tool that automates the task of assessing whether a chemical is part of a CW-control list. The tool, dubbed the Nonproliferation Cheminformatics Compliance Tool (NCCT), is a database management system (based on ChemAxon's Instant JChem) with an embedded database of chemical structures. The key feature of the database is that it contains not only the structures of the individually listed chemicals, but also the generic structures that describe the entries relative to families of chemicals (Markush structures).

**Keywords:** cheminformatics, chemical weapons, precursors, export controls, Australia Group, Chemical Weapons Convention, Syria, Wassenaar Arrangement, World Customs Organization.

---

<sup>†</sup> Corresponding author. Email: [costanzi@american.edu](mailto:costanzi@american.edu)

<sup>§</sup> Mary Vecellio has taken a position at another organization. Her contributions to this article were offered only during her employment at the Stimson Center.

## INTRODUCTION

Chemicals weapons (CW) are weapons intended to cause death or incapacitation through the toxic properties of chemicals. The Chemical Weapons Convention (CWC), a disarmament treaty that entered into force in 1997, poses a complete ban on chemical weapons, prohibiting not only their use, but also their development, production, and stockpiling [1–3]. As further explained in the “CW-control lists Implemented in the NCCT database” section, to support its verification regime and declaration requirements, the CWC features three schedules that list key toxic chemicals on which chemical weapons can be based as well as precursors for their synthesis.

However, far from being a relic from the past, chemical weapons have been used in recent years on several occasions, although the chemical weapons landscape has changed. Current chemical weapons attacks do not involve large quantities of chemicals like those that were deployed in World War I. Conversely, smaller quantities of chemicals have been used by Syria for counterinsurgency operations during its civil war and by the Islamic State against civilians and government forces in Iraq and Syria. [4–8]. Even smaller quantities of chemicals have been employed for targeted assassinations and assassination attempts, most recently for the attempted murder of political activist Alexei Navalny with a Novichok agent [9].

Due to the changing chemical weapons landscape, it is more important than ever to control even small quantities of toxic chemicals that can be used as chemical weapons and precursors for their synthesis. Hence, all possible variants of a given chemical, including those that are not typically produced on a large scale must be taken into account [10]. This is consistent with a recent recommendation from the Scientific Advisory Board (SAB) to the Organisation for the Prohibition of Chemical Weapons (OPCW), the international organization that supports the implementation of the CWC. Specifically, the SAB recommended that “the isotopically labelled compound or stereoisomer related to the parent chemical specified in the schedule should be interpreted as belonging to the same schedule” [11,12].

Several frameworks at the national and international level have been put in place to foster CW nonproliferation and promote chemical disarmament, including the CWC, Australia Group, Wassenaar Arrangement, the World Custom Organization’s Strategic Trade Control Enforcement Programme, and European Union export controls. To support their missions, these frameworks contain lists of chemicals that can potentially be employed as chemical weapons or precursors for their synthesis. In all these lists, chemicals are identified through chemical names and, whenever available, registry numbers from the American Chemical Society’s Chemical Abstract Service (CAS) – CAS Registry Numbers®, also referred to as CAS RNs® [10,13]. This is well illustrated by the snippet of a CW-control list provided in **Fig. 1** [14].

Very importantly, in some of the CW-control lists, some entries are not individual chemicals. They are families of related chemicals based on a common scaffold with variable substituents. These entries do not have CAS RNs associated with them, as CAS RNs only pertain to individual chemicals, not families of chemicals. For instance, the first entry of the first of the three schedules featured in the Annex on Chemicals of the CWC (Schedule 1) defines a large family of nerve agents that includes sarin and soman, which are listed as specific examples, as well as a large number of additional agents that are not explicitly listed. The entry is defined as “O-Alkyl ( $\leq C_{10}$ , including cycloalkyl) alkyl (Me, Et, n-Pr or i-Pr)-phosphonofluoridates”, where Me, Et, n-Pr and i-Pr stand for methyl, ethyl, normal-propyl, and isopropyl, respectively. As explained by Pontes and coworkers, depending on the definition accepted for the word

cycloalkyl, that entry may encompass in excess of 2 million different chemicals. Other listed families are so large that they cannot be enumerated [15–17].

Working with lists of chemicals and identifying what chemicals are covered by them poses some challenges for frontline officers implementing these frameworks, such as export control officers and customs officials, as well employees of chemical manufacturing, shipping, and logistics companies. In particular, entries that describe families of chemicals are not easy to interpret, especially for non-chemists. Moreover, synonyms and chemical variants complicate the issue of checking control lists through names and CAS RNs [10]. By way of example, as illustrated by Pontes and coworkers, more than 80 different synonyms can be found for the nerve agent sarin in chemical databases such as CAS, the Royal Society of Chemistry's ChemSpider, and the National Library of Medicine's PubChem [15]. Similarly, the CAS contains 17 different variants of sarin with different CAS RNs. Beyond the plain version of the molecule, these include isotopically labelled variants, specific isomers, hydrates, and cyclodextrin-enclosed variants [15].

Through this work, we are endeavoring to make it easier for a wide range of relevant stakeholders, many of whom do not have extensive training in chemistry, to identify whether a chemical under scrutiny is covered by a given list of chemicals that are controlled for security concerns. Among others, possible stakeholders include customs officials, export control officers, and employees of chemical manufacturing, shipping, or logistics companies. To simplify the task of assessing whether a chemical is part of a CW-control list, we propose the development and adoption of a cheminformatics tool, consisting of a database management system with an embedded database of relevant CW-control lists, that can automate this process. We call this system, of which we have developed a working prototype, the Nonproliferation Cheminformatics Compliance Tool (NCCT).

## RESULTS AND DISCUSSION

### General description of the IJC-based NCCT prototype

The NCCT prototype is a desktop-based database management system with an embedded database of CW-control lists (the NCCT database). We have developed the NCCT prototype through Instant JChem (IJC), a chemical database management system developed by ChemAxon (**Fig. 2**) [18].

The IJC-based NCCT prototype allows a user to input a query chemical, use the entered query to search the embedded NCCT database, and retrieve a list of NCCT Database entries that match the query chemical, if any (**Fig. 3**). Specifically, through the Instant JChem graphical user interface (GUI), the user can input a query chemical into the IJC-based NCCT prototype in a variety of ways, including CAS RNs, chemical names, and structural identifiers such as SMILES or the IUPAC-developed InChI codes.<sup>1</sup> Alternatively, the user can draw the 2D structure of the query chemical through the Instant JChem GUI. The Instant JChem engine then converts the chemical, no matter which way it has been entered, into a chemical structure. Subsequently, it establishes the equivalence of chemical variants by standardizing the query (**Fig. 4**). The standardized query structure is then checked against the standardized NCCT database, and the database entries that match the query, if any, are retrieved.

A key feature of the IJC-based NCCT prototype is that the NCCT database contains not only the structures of the individually listed chemicals, but also the generic structures that describe the entries

---

<sup>1</sup> Structural identifiers are text strings that provide a machine readable representation of molecular structures [19].

relative to the families of chemicals, encoded as Markush structures.<sup>2</sup> Another key feature of the IJC-based NCCT prototype is the mentioned standardization process applied to the NCCT database tables as well as to the query structures, which establishes the equivalence of different variants of the same chemical, including stereoisomers, protonation states, charges, radicals, salts, and isotopically labelled forms (for further details, see the **Methods** section). The standardization is particularly important considering the changing chemical weapons landscape, which makes it imperative to control all variants of the chemicals of concern. This approach is consistent with the mentioned SAB recommendation according to which isotopically labelled or stereoisomeric forms of a scheduled chemical should be considered as belonging to the same schedule [11,12].

### **CW-control lists implemented in the NCCT database**

As illustrated in the paragraphs below, to date, we have implemented into the NCCT database the CW-control lists from five international frameworks (**Fig. 5**). Each CW-control list was implemented as one or two separate NCCT database tables, depending on whether their entries required different standardization procedures (for further details, see the **Methods** section). In total, the NCCT database comprises a total of 11 database tables, one for each of the three CWC schedules and two for each of the remaining four international frameworks. The total number of entries, including the 3 exceptions listed in the CWC Schedules, is 588. The CW-control lists featured in the NCCT database are also available, in curated and structurally annotated form, at the Costanzi Research website (<https://costanziresearch.com/cw-nonproliferation/cw-control-lists/>) [13].

*Chemical Weapons Convention (CWC) Schedules.* The CWC is an international disarmament treaty that seeks the complete elimination of chemical weapons. As mentioned, the treaty poses a complete ban on chemical weapons, prohibiting not only their use, but also their development, production, and stockpiling. With its 193 State Parties, it enjoys almost universal embracement and is the most prominent international framework for chemical disarmament and nonproliferation. To support its verification regime and declaration requirements, in its Annex on Chemicals, the CWC features three schedules of chemicals: Schedule 1, Schedule 2, and Schedule 3. Going from Schedule 1 to Schedule 3, the schedules contain chemicals that, beyond their chemical weapons-related role, have increasingly larger legitimate commercial applications (with Schedule 1 listing chemicals with minimal or no commercial applications, and Schedule 3 listing chemicals with large industrial applications). Each schedule is divided into two parts: toxic chemicals are listed in Part A and precursors are listed in in part B. The CWC Schedules comprise 78 entries. Some of these entries describe individual chemicals, while others describe families of related chemicals, defined as a central chemical scaffold bearing a number of attached variable chemical groups [13,21].

*Australia Group (AG) Chemical Weapons Precursors list.* The AG is an informal arrangement between 43 like-minded states committed to preventing the proliferation of chemical and biological weapons through the harmonization of export controls. The AG Chemical Weapons Precursors list comprises a total of 87 dual-use chemicals that can be used as precursors for the synthesis of chemical weapons. All of these 87 entries are explicitly listed as discrete chemicals [13,22].

*The Wassenaar Arrangement (WA) Munitions List 7 (ML7).* The WA is an international framework, adhered to by 42 countries, that was established with the objective of “promoting transparency and greater

---

<sup>2</sup> A Markush structure is a chemical a structure composed of a central scaffold with one or more variable chemical groups attached to it. Markush structures are commonly used in patents to define families of related chemicals [20].

responsibility in transfers of conventional arms and dual-use goods and technologies.” Within its ML7, the WA features chemical agents, biological agents, riot control agents, radioactive materials, related equipment, components, and materials. In particular, with regards to chemical agents, ML7 comprises all of the CWC Schedule 1 chemicals, the central incapacitating agent BZ, which is a CWC Schedule 2 chemical, as well as a list of defoliants and a list of riot control agents, neither of which is in the CWC schedules. ML7 comprises 39 entries. Just like for the CWC Schedules, some of these entries describe individual chemicals while others describe families of related chemicals [10,23].

*European Union Council Regulation 36/2012 (Syria-related list).* The European Union tightly regulates the export of dual-use chemicals to Syria, within the scope of the restricting measures imposed by Council Regulation 36/2012. Beyond posing additional restrictions on the chemicals already present in the general EU export control lists, in Annex Ia and Annex IX, EU Council Regulation 36/2012 lists additional dual-use chemicals that, although being widely used in chemical industry for non-military purposes, are of security concern when exported to Syria. An example of such chemicals is isopropanol, a dual-use chemical that, beyond its legitimate uses in chemical synthesis or as a disinfectant, is also a precursor, in a highly pure form, for the synthesis of the nerve agent sarin. Annex Ia and Annex IX of EU Council Regulation 36/2012 comprise a total of 80 entries, all of which are explicitly listed as discrete chemicals [10,24].

*The strategic chemicals identified in the World Customs Organization (WCO) Strategic Trade Control Enforcement Implementation Guide (STCE).* The WCO is “an independent intergovernmental body whose mission is to enhance the effectiveness and efficiency of Customs administrations.” The WCO maintains a document – the Strategic Trade Control Enforcement Implementation Guide (or STCE) – intended to provide WCO members with “practical assistance related to enforcing strategic trade controls.” In Annex V, the STCE provides a list of strategic chemicals that goes beyond chemicals of CW-proliferation concern and includes chemicals of nuclear, missile, explosive, and military concern. The WCO list of strategic chemicals comprises a total of 304 entries, all of which are explicitly listed as discrete chemicals [10,25].

## Description of the table fields

For each entry in the NCCT database, 12 fields are given (**Fig. 6**).

The *Markush structure* field is the field that is searched when the NCCT database is queried. This field is essential for the functioning of the IJC-based NCCT prototype. Despite the fact that Instant JChem labels it as “Markush structure,” this field can be populated with both discrete structures (for individually listed chemicals) and Markush structures (for families of chemicals).

Three fields directly reflect the information provided by the official CW-control frameworks for the listed entries. Specifically, these include the *Entry number* field, which reflects the number assigned to the entry in the official CW-control list, as well as the *Entry name* and *CAS Registry Number®* fields.

The remaining eight fields contain complementary useful information with which we annotated the database [13]. Specifically, these fields include:

- three structural identifier fields (namely a *SMILES*, an *InChI*, and an *InChiKey* field);
- a *PubChem ID* field, which reports the ID provided for that chemical by the National Library of Medicine’s PubChem database and a related *PubChem URL* field, which provides an active link to the relative PubChem database entry [26,27];
- an *overlap* field, which indicates whether the chemical in question is also covered by one or more of the other lists of controlled chemicals implemented in the NCCT database;

- a *category* field, which indicates whether the chemical is a chemical weapon agent (indicating the specific class), a precursor for the synthesis of chemical weapons, a defoliant, a riot control agent, or a chemical posing a different threat (explosive, nuclear, missile, or general military concern);
- an *entry type* field, which indicates whether the entry is a family of chemicals, an example pertinent to a family of chemicals, an exception to a family of chemicals (*i.e.*, a chemical that falls within the scope of the family definition but is explicitly excluded from controls in the listing framework), an individually listed small molecule, a polymer, a protein, or a mixture of chemicals.

### Search example 1: addressing the families of chemicals issue.

The paragraphs below illustrate how the IJC-based NCCT prototype helps an operator identify whether a chemical under scrutiny falls under the scope of one of the families of chemicals featured in the CW-control lists implemented into the NCCT database.

The chloroethylamine in **Fig. 7A** is a precursor for the synthesis of the nerve agent VX (shown in the yellow inset) [10]. The CWC does not list this chemical explicitly. Instead, in entry 10 of Part B of Schedule 2 (CWC 2B10), it lists a family of chloroethylamines that comprises the VX precursor. Specifically, the definition provided in CWC Schedule 2 for this entry is: N,N-Dialkyl (Me, Et, n-Pr or i-Pr) aminoethyl-2-chlorides and corresponding protonated salts (**Fig. 7B**) [28]. This family is characterized by a central scaffold common to all family members and two variable R groups (**Fig 7C**).

The CAS RN for the VX precursor in question is 96-79-7 and the primary name listed by CAS is N-(2-chloroethyl)-N-(1-methylethyl)-2-propanamine. Given this chemical name or this CAS RN, it is not straightforward to infer that this chemical is covered by CWC Schedules. In fact, a frontline officer who tried to match said name and CAS RN with those listed in the CWC Schedules would conclude that the VX precursor is not one of the covered chemicals.

The IJC-based NCCT prototype automates the task of checking whether the VX precursor in question is covered by the CWC Schedules or any of the other implemented international CW-control lists. The operator can enter either the chemical name or the CAS RN® shown in **Fig. 8A**, both of which will be converted by the Instant JChem engine into a structure (**Fig. 8B**). The operator can then launch the query and the IJC-based NCCT prototype will search the NCCT database and identify all the entries in the implemented CW-control lists that cover this VX precursor, either as an individual chemical or as part of a family of chemicals. The results will indicate that the VX precursor is covered by CWC Schedule 2, as part of family 2B10 (**Fig. 8C**). The results will also indicate that the VX precursor is covered as an individual chemical by the AG Chemical Weapons Precursors list, as entry AG 11, and the World Customs Organization STCE list, as entry STCE 19 (**Fig. 8C**) [22]. Of note, it would have been straightforward to match this VX precursor with entries AG 11 and STCE 19 based on the CAS RN 97-79-7, as these entries list the chemical individually and report its CAS RN. However, it would not have been possible to do the same for entry CWC 2B10, because, as mentioned, the entry covers the whole family of chloroethylamines without explicitly enumerating its members.

### Search example 2: addressing the synonyms and chemical variants issue.

The paragraphs below illustrate how the IJC-based NCCT prototype addresses the issues caused by synonyms and chemical variants, thus helping an operator determine whether a chemical under scrutiny is part of a CW-control list even if the supplied chemical name or CAS RN® are different from those found in the list.

The alcohol shown in **Fig. 9A** is a precursor for the synthesis of the nerve agent soman [29]. It is listed by the AG as entry 28, with the name of pinacolyl alcohol and the CAS RN 464-07-3 (**Fig. 9B**) [22]. However, the same chemical can be described with many different synonyms, some of which are listed in **Fig. 9C**. Hence, simply checking whether a chemical name is part of a list, is not sufficient to assess whether the chemical in question is covered by that list.

Using the CAS RN rather than the name makes things easier, as all these synonyms have the same CAS RN. However, an additional layer of complexity is added by the fact that different variants of the same chemical have different CAS RNs. A few variants of pinacolyl alcohol, including an ionized version of the molecule, two stereoisomers, and an isotopically labelled version of the molecule, all of which have different CAS RNs, are shown in **Fig. 9D**. Hence, simply checking whether a CAS RN is part of a list is not sufficient to assess whether the chemical in question is covered by that list, as the chemical in question could be a variant of a chemical whose canonical form is included in a CW-control list.

For the (*S*) stereoisomer variant pinacolyl alcohol, given the chemical name (*S*)-1-tert-butylethanol or the CAS RN 1517-67-5 it is not straightforward to infer that this chemical is covered by a CW-control list. In fact, a frontline officer who tried to match said name and CAS RN with those listed in the CW-control lists would conclude that this chemical is not controlled.

The IJC-based NCCT prototype automates the task of checking whether the variant of pinacolyl alcohol in question is covered by one or more of the international CW-control lists implemented into it. An NCCT operator can enter either the chemical name or the CAS RN shown in **Fig. 10A**, both of which will be converted by the Instant JChem engine into a structure (**Fig. 10B**). The operator can then launch the query and the IJC-based NCCT prototype will search the NCCT database and identify all the entries in the implemented CW-control lists that cover pinacolyl alcohol, either as an individual chemical or as part of a family of chemicals. The results will indicate that pinacolyl alcohol is covered as an individual chemical by the CWC Schedules, as entry CWC 2B14, the AG Chemical Weapons Precursors list, as entry AG 28, and the World Customs Organization STCE list, as entry STCE 49 (**Fig. 10C**).

### Current limits and future directions.

The IJC-based NCCT prototype currently has some limits which we will seek to address as the tool evolves into a mature product.

First, Instant JChem's interface is designed with chemists in mind, thus making the IJC-based NCCT prototype less intuitive to use for people who lack a certain level of training in chemistry. Going forward, endowing the tool with an easier to use interface will be a key aspect of future development efforts. In particular, the development of the next iteration of the NCCT could leverage existing commercial products that were developed with the primary goal of supporting the control of regulated substances, chiefly prescription and illegal narcotic and psychotropic drugs. These tools, which were developed within the scope of Substance Compliance Service Project of the Pistoia Alliance,<sup>3</sup> include Controlled Substances Squared (CS2), from Scitegrity, and Compliance Checker, from ChemAxon [31–33]. Both tools are available in a web-based version, with a streamlined interface. Notably, although targeting mainly controlled substances, both tools already have some CW-control lists in their databases [10].

Second, the IJC-based NCCT prototype relies entirely on Instant JChem's engine for the conversion of names and CAS RNs into structures, and there are several situations in which this does not work. It occurs

---

<sup>3</sup> The Pistoia Alliance is a global, not-for-profit organization of stakeholders in the life science domain, including pharmaceutical companies with the mission of lowering barriers to innovation [30].

rather often that chemical names cannot be interpreted and converted into structures. Similarly, although less frequently, some CAS RNs cannot be converted into structures. Among other situations, in Instant JChem, this occurs rather commonly for isotopically labelled compounds. For instance, neither the name nor the CAS RNs for the isotopically labelled version of sarin shown in **Fig. 11a** can be converted into a structure by Instant JChem. This limit is due to the fact that CAS RNs cannot be converted to structures by computers algorithmically, as the numbers do not have structural information embedded in them. For the conversion of CAS RNs to structures, a connection to a comprehensive, up-to-date relational database that links the registry numbers to the chemicals that they identify is always needed. Going forward, endowing the NCCT with a robust engine for the conversion of names and CAS RNs into structures will be a key aspect of future development efforts.

It is worth pointing out that, rather than using names or CAS RNs, it is always possible to enter the chemical as a structural identifier, such as a SMILES or an InChI code. As mentioned, structural identifiers are text strings that encode a molecular structure [13,19]. A structural identifier string contains all the information needed to infer the structure of the chemical to which it pertains. Hence, structural identifiers will be successfully converted into a molecular structure by cheminformatics software algorithmically. For instance, although, Instant JChem cannot convert the name or the CAS RN of  $^{32}\text{P}$ -labelled sarin, the chemical can be successfully entered as a SMILES string or an InChI code (**Fig. 11a**). The structural identifier will be algorithmically converted by the software into a structure, thus allowing the search to be performed (**Fig. 11b**). Although the input chemical is an isotopically labelled version of sarin, thanks to the standardization process, the IJC-based NCCT prototype will be able to establish that it is equivalent to non-labelled sarin. The results will indicate that sarin is covered by the CWC Schedules, as member of family CWC 1A1 (for which it is also listed as a specific example), by the Wassenaar Arrangement ML7, as a member of family b.1.a (for which it is also listed as a specific example), and by the World Customs Organization STCE list, as entry STCE 28 (**Fig. 11c**). Beyond structural identifiers, it is also worth noting that, in many cases, the IJC-based NCCT prototype can convert proper IUPAC names into structures.

Lastly, the Instant JChem function that is used to query the database in the IJC-based NCCT prototype only searches the *Markush Structure* field of the NCCT database, but not the associated metadata (see **Launching queries** in the **Methods** section). Going forward, enabling searches for entries that cannot be searched by structure will be a key aspect of future development efforts. In particular, it will be important to add to the NCCT the ability to query the metadata associated with each entry as well. Coupled with a thorough annotation of NCCT database entries with the synonyms listed in databases such as CAS, ChemSpider, and PubChem, and the CAS RNs associated with all the known variants of the chemical in question, this feature is expected to significantly enhance the robustness of queries based on chemical names and CAS RNs, enabling the identification of controlled chemicals in some of the cases where the conversion of names or CAS RNs to structures does not succeed. Importantly, this feature will also mitigate the current complete inability of the NCCT to query for substances endowed with larger structures, such as biological macromolecules and polymers. Of the CW-control lists implemented into the NCCT database tables, only nine such entries exist (seven organic polymers and two biological macromolecules). In the NCCT database tables, the *Markush Structure* field is left empty, as organic polymers have a variable length, thereby lacking a uniquely defined molecular structures, and biological macromolecules have chemical structures that are too large to be handled by the system. Hence, given that, as mentioned, the queries are currently confined to the *Markush Structure* field of the NCCT database, organic polymers and biological macromolecules are left out of the search.



## CONCLUSIONS

Summarizing, for frontline officers, such as export control and customs officers as well as employees of chemical manufacturing, shipping, and logistics companies, it can be difficult to assess whether a chemical under scrutiny is covered by the lists of chemicals embedded in one or more of the international frameworks for the control of chemical weapons and, more generally, chemicals of security concern. As discussed, this is mainly due to two issues: 1) matching a chemical under scrutiny with CW-control list entries that describe whole families of chemicals is a very complex task that cannot be undertaken by individuals who do not have a substantial training in chemistry; 2) matching a chemical under scrutiny with CW-control list entries that describe individual chemicals, although being an easier task, is significantly complicated by the many synonyms associated with a chemical name and the fact that different variants of the same chemical have different CAS RNs [10].

To ameliorate these issues, we have developed a working prototype of the NCCT, a cheminformatics tool that automates the task of checking whether a chemical under scrutiny is indeed encompassed by one or more CW-control lists either because it falls within the scope of one of the listed families or because it is listed as an individual chemical. As described above, the IJC-based NCCT prototype is a database management system, based on ChemAxon's Instant JChem platform, with an embedded database of chemical structures.

Through internal tests, we have identified the limits of the IJC-based NCCT prototype, which chiefly revolve around the complexity of the Instant JChem interface, its engine for the conversion of names and CAS RNs into structures, and the restriction of the federated search to the structural field of the database. To guide the development of the NCCT, we have also established a network of relevant stakeholders drawn from international organizations, government agencies, industry, civil society, and academia, including an advisory group, who will assist in the identification of the requirements for a more advanced version of the prototype. Input from stakeholders is key to ensuring that the NCCT adds significant value to current and future efforts to enhance chemical security and prevent the proliferation of chemical weapons. Going forward, we will work with select jurisdictions to subject the IJC-based NCCT prototype to field tests intended to verify the practical usefulness of the tool, further probe its limitations, and assist with identification of the requirements for future iterations of the tool.

Controlling the export of toxic chemicals that can be used as chemical weapons, precursors for their synthesis, and, more generally, chemicals of security concern is paramount when trying to prevent the proliferation of chemical weapons and support chemical security. It is key that the controls be implemented in a thorough and timely manner to prevent illegal transfers of chemicals, while at the same time facilitating legitimate transfers. By bolstering the ability of frontline officers to effectively perform such controls, the NCCT will contribute to ensuring that chemistry be only applied to serve peaceful purposes and support the progress of humanity [10].

## METHODS

### Instant JChem Platform

The IJC-based NCCT prototype was built and is managed through ChemAxon's Instant JChem software (IJC), version 21.8.0 [18]. Specifically, as described below, Instant JChem was used to build the NCCT database. Moreover, Instant JChem is also used to manage and query the NCCT database. The Instant JChem software is a desktop-based application that runs on Windows and MacOS computers. Installation

of Instant JChem is required to query the NCCT database. The NCCT database also resides locally on the desktop computers, in Derby format.

### NCCT tables in CSV format

A comma-separated values (CSV) file was created for each of the CW-control tables to be implemented into the NCCT database. The CSV file had a total of 12 columns that correspond to the NCCT table fields listed above in the **Results** section (**Description of the table fields**), with the exception of the *PubChem URL* field, which was created directly in Instant JChem after the import of the NCCT tables (see next section for more details). For each entry, we populated the “Structure” column with the SMILES string retrieved from the Chemical Abstract Service through the SciFinder<sup>n</sup> web tool [34]. When this was not available, the field was left blank. The columns “Entry Number”, “Entry Name”, and “CAS Registry Number®” were populated with the information found in the official version of the CW-control lists. The remaining seven columns were populated with complementary pertinent information (see “Description of the table fields” in the Results section for further details). Two of these seven columns were populated with InChI structural identifiers and InChIKey codes, which are their hashed version [13,19]. InChI and InChIKey codes were derived from the SMILES retrieved from SciFinder<sup>n</sup>. In particular, the SMILES strings were pasted into the “Draw structure” tool of the National Library of Medicine’s PubChem website [27], where they were then converted to their corresponding InChI and InChIKey codes.

### Implementation of the NCCT database

The NCCT tables in CSV format were imported into Instant JChem, where they were converted into searchable database tables in Derby format through the IJC-embedded Derby database management system (the NCCT database). Each CSV file was converted into an NCCT database table; each row in the CSV file became an entry in the corresponding NCCT database table; each column in the CSV file became a field in the corresponding NCCT database table. The NCCT tables in CSV format were imported as “Markush libraries.” This is a key feature of the NCCT database, as it allows handling of the families of chemicals contained in several CW-control lists. The “Empty structures allowed” feature was turned on. The Instant JChem software automatically converted the SMILES strings in the “Structure” column into bidimensional molecular structures, which were visually inspected, thoroughly checked for accuracy, and corrected whenever needed. For the individually listed chemicals for which the “Structure” field was left blank, the molecular structures were built with ChemAxon’s Marvin interface, as implemented in Instant JChem. For families of chemicals, the Markush structures were built using ChemAxon’s Markush Editor, version 21.8.0 [35]. The Markush structures were then added to the NCCT database in Instant JChem. Once the NCCT database was implemented, we added a twelfth field featuring clickable PubChem URL links. This was done by adding a static URL field through the Instant JChem interface.

### Separation of single component and multiple component substances

Most of the entries in the CW-control lists implemented in the NCCT database contain a single chemical species (single component substances). However, in some CW-control lists, some of the entries, e.g. salts and mixtures, contain multiple chemical species (multiple component substances). In the NCCT database, we split the CW-control lists that feature both single component and multiple component substances into two tables, one with single component substances only and one with multiple component substances only. This is due to the need to apply different standardizers to the single component and multiple component substances (see the paragraph below for further details).

## Standardization of tables and queries

Standardizers were applied to the NCCT database tables to account for the equivalence of different variants of the same chemical as well as different ways of representing the same chemical. The following standardizers were applied to all NCCT database tables: Aromatize, Remove Explicit Hydrogens, Clear Isotopes, and Neutralize. For the NCCT database tables featuring single component substances, the Remove Solvents and Remove Fragment standardizers were also applied to account for the equivalence of salts (these standardizers cannot be applied to lists featuring multiple component substances because they would cause the deletion of one of the components of the entries). The parameters for the Remove Fragment standardizer were set to “Remove smallest” and “Depends on number of heavy atoms” in all the NCCT database tables except for the one relative to CWC Schedule 2, where it was set to “Keep largest” and “Depends on the number of heavy atoms.” The different approach taken for Schedule 2 is due to the fact that, because of the size of one of the Schedule 2 families (entry 2B4), the “Remove smallest” option would not allow Schedule 2 queries to run.

Query structures are automatically standardized during the search process, so that their standardization always matches the one that has been applied to the NCCT table that is being searched. In particular, before a query is run against a specific NCCT database table, it is subjected to the same standardization process to which that specific database table was subjected when the NCCT database was created. Once the search moves to the next NCCT database table, the query is automatically re-standardized to match the standardization process to which the currently searched database table was subjected.

## Alternative structural representations

The IJC-based NCCT prototype was tested to verify whether it recognized different structural representations of the chemicals featured in the NCCT database tables. In particular, we ensured that it recognized the representation obtained by inputting into the Instant JChem interface: 1) CAS RN; 2) SMILES string, as found in CAS; 3) SMILES string, as found in PubChem. Whenever a structural representation not recognized by the prototype was found, this alternative structural representation was appended to the relevant NCCT database table.

## Launching queries

Query chemicals are searched against the NCCT database using the “Simple Federated Search” function of Instant JChem. The Simple Federated Search allows for the searching of all the NCCT database tables at the same time. The search can also be restricted to one or more of the NCCT database tables, if desired. Of note, the Simple Federated Search only queries Structure column of the NCCT database tables, not the metadata found in the remaining 11 columns. The search mode is set to “Full” (*i.e.*, the query has to match the entire entry to yield a hit). The search options are set as follows. Stereochemistry: off; Charges, Isotopes, Radicals, and Valence: Ignore; Vague Bond: Ambiguous aromaticity 5-membered rings; Markush: Homology broad translation; Tautomer: Off. Once configured to the above settings, the query can be entered into Instant JChem by inputting its chemical name, its CAS RN or a structural identifier, which are all converted by Instant JChem into a molecular structure. Alternatively, the structure can be sketched using Instant JChem’s ‘Sketch Structure’ feature.

## RESEARCH FUNDING

This work was supported by Global Affairs Canada under award CWC-2020-0001 (P009515).

## REFERENCES

- [1] S. Costanzi, in *Kirk-Othmer Encyclopedia of Chemical Technology* (Wiley Online Library, 2020), pp. 1–32.
- [2] V. Pitschmann, *Toxins* **6**, 1761 (2014).
- [3] K. Ganesan, S. K. Raza, and R. Vijayaraghavan, *Journal of Pharmacy and Bioallied Sciences* **2**, 166 (2010).
- [4] R. K. Hersman and W. Pittinos, *Restoring Restraint: Enforcing Accountability for Users of Chemical Weapons* (Rowman & Littlefield, 2018).
- [5] R. K. Hersman and S. Claeys, *Rigid Structures, Evolving Threat: Preventing the Proliferation and Use of Chemical Weapons* (Center for Strategic & International Studies, 2019).
- [6] S. Costanzi, J.-H. Machado, and M. Mitchell, *ACS Chem. Neurosci.* **9**, 873 (2018).
- [7] S. Costanzi and G. D. Koblentz, *The Nonproliferation Review* **26**, 1 (2019).
- [8] S. Costanzi and G. D. Koblentz, *Arms Control Today* **50**, 16 (2020).
- [9] D. Steindl, W. Boehmerle, R. Körner, D. Praeger, M. Haug, J. Nee, A. Schreiber, F. Scheibe, K. Demin, P. Jacoby, R. Tauber, S. Hartwig, M. Endres, and K.-U. Eckardt, *The Lancet* **397**, 249 (2021).
- [10] S. Costanzi, G. D. Koblentz, and R. T. Cupitt, *Strategic Trade Review* **6**, 69 (2020), <https://doi.org/10.1080/10736700.2021.2020010>.
- [11] OPCW, Report of The Scientific Advisory Board on Developments in Science and Technology for the Fourth Special Session of the Conference of the States Parties to Review the Operation of the Chemical Weapons Convention, **RC-4/DG.1**, 30 (2018).  
[https://www.opcw.org/sites/default/files/documents/CSP/RC-4/en/rc4dg01\\_e\\_.pdf](https://www.opcw.org/sites/default/files/documents/CSP/RC-4/en/rc4dg01_e_.pdf)
- [12] C. M. Timperley, J. E. Forman, M. Abdollahi, A. S. Al-Amri, I. P. Alonso, A. Baulig, V. Borrett, F. A. Cariño, C. Curty, and D. Gonzalez, *Pure and Applied Chemistry* **90**, 1647 (2018).
- [13] S. Costanzi, C. K. Slavick, B. O. Hutcheson, G. D. Koblentz, and R. T. Cupitt, *J. Chem. Inf. Model.* **60**, 4804 (2020).
- [14] OPCW, Annex on Chemicals - Schedule 1, <https://www.opcw.org/chemical-weapons-convention/annexes/annex-chemicals/schedule-1>.
- [15] G. Pontes, J. Schneider, P. Brud, L. Benderitter, B. Fourie, C. Tang, C. M. Timperley, and J. E. Forman, *Journal of Chemical Education* **97**, 1715 (2020).
- [16] C. Rücker, M. Meringer, and A. Wassermann, *Journal of Chemical Education* **98**, 1465 (2021).
- [17] C. M. Timperley and J. E. Forman, *Journal of Chemical Education* **98**, 1468 (2021).
- [18] ChemAxon, Instant IChem, <https://chemaxon.com/products/instant-jchem>.
- [19] W. A. Warr, *WIREs Computational Molecular Science* **1**, 557 (2011).
- [20] D. A. Cosgrove, in *Scaffold Hopping in Medicinal Chemistry* (Wiley, 2013), pp. 15–38.
- [21] OPCW, Annex on Chemicals, <https://www.opcw.org/chemical-weapons-convention/annexes/annex-chemicals/annex-chemicals>.
- [22] The Australia Group, Chemical Weapons Precursors, <https://www.dfat.gov.au/publications/minisite/theaustraliagroupnet/site/en/precursors.html>.
- [23] The Wassenaar Arrangement, Control Lists, <https://www.wassenaar.org/control-lists/>.
- [24] Council Regulation (EU) No 36/2012, <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02012R0036-20200217&from=EN#tocId2>.
- [25] World Customs Organization, Strategic Trade Control Enforcement Implementation Guide, <http://www.wcoomd.org/en/topics/enforcement-and-compliance/instruments-and-tools/guidelines/wco-strategic-trade-control-enforcement-implementation-guide.aspx>.
- [26] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, and B. Yu, *Nucleic Acids Research* **47**, D1102 (2019).
- [27] National Library of Medicine, PubChem, <https://pubchem.ncbi.nlm.nih.gov>.

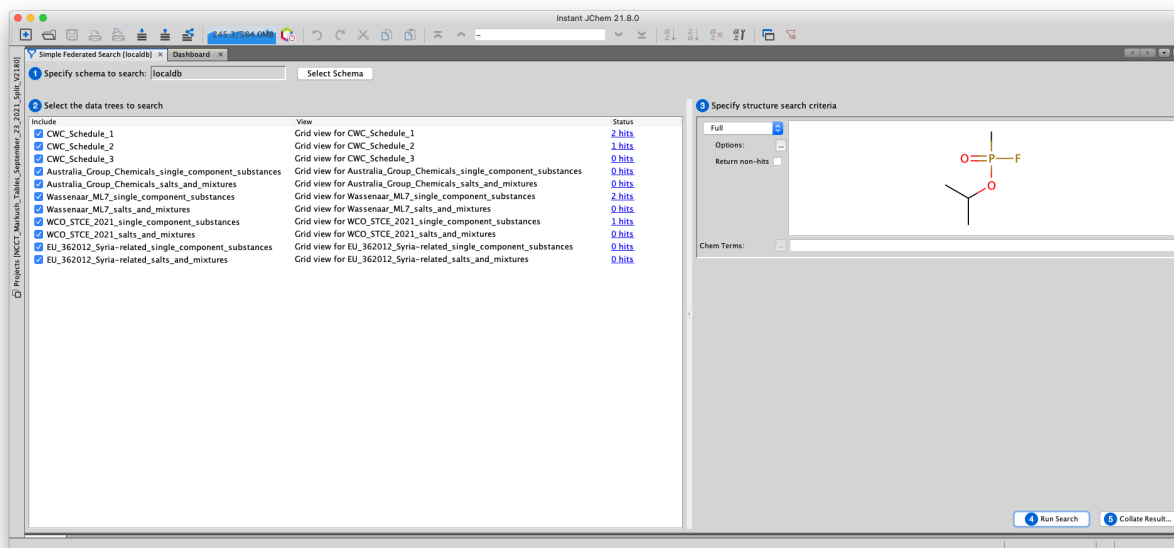
- [28] OPCW, Annex on Chemicals - Schedule 2, <https://www.opcw.org/chemical-weapons-convention/annexes/annex-chemicals/schedule-2>.
- [29] J. B. Tucker, *The Nonproliferation Review* **16**, 363 (2009).
- [30] Pistoia Alliance, <https://www.pistoiaalliance.org>.
- [31] Scitegrity, Controlled Substances Squared, <https://scitegrity.co.uk/index.php?page=cs2>.
- [32] ChemAxon, Compliance Checker, <https://chemaxon.com/products/compliance-checker>.
- [33] D. Taylor, S. G. Bowden, R. Knorr, D. R. Wilson, J. Proudfoot, and A. E. Dunlop, *Drug Discovery Today* **20**, 175 (2015).
- [34] Chemical Abstract Service (CAS), SciFinder-n, <https://scifinder-n.cas.org/>.
- [35] ChemAxon, Markush Editor, <https://chemaxon.com/products/instant-jchem>.

## FIGURES AND CAPTIONS

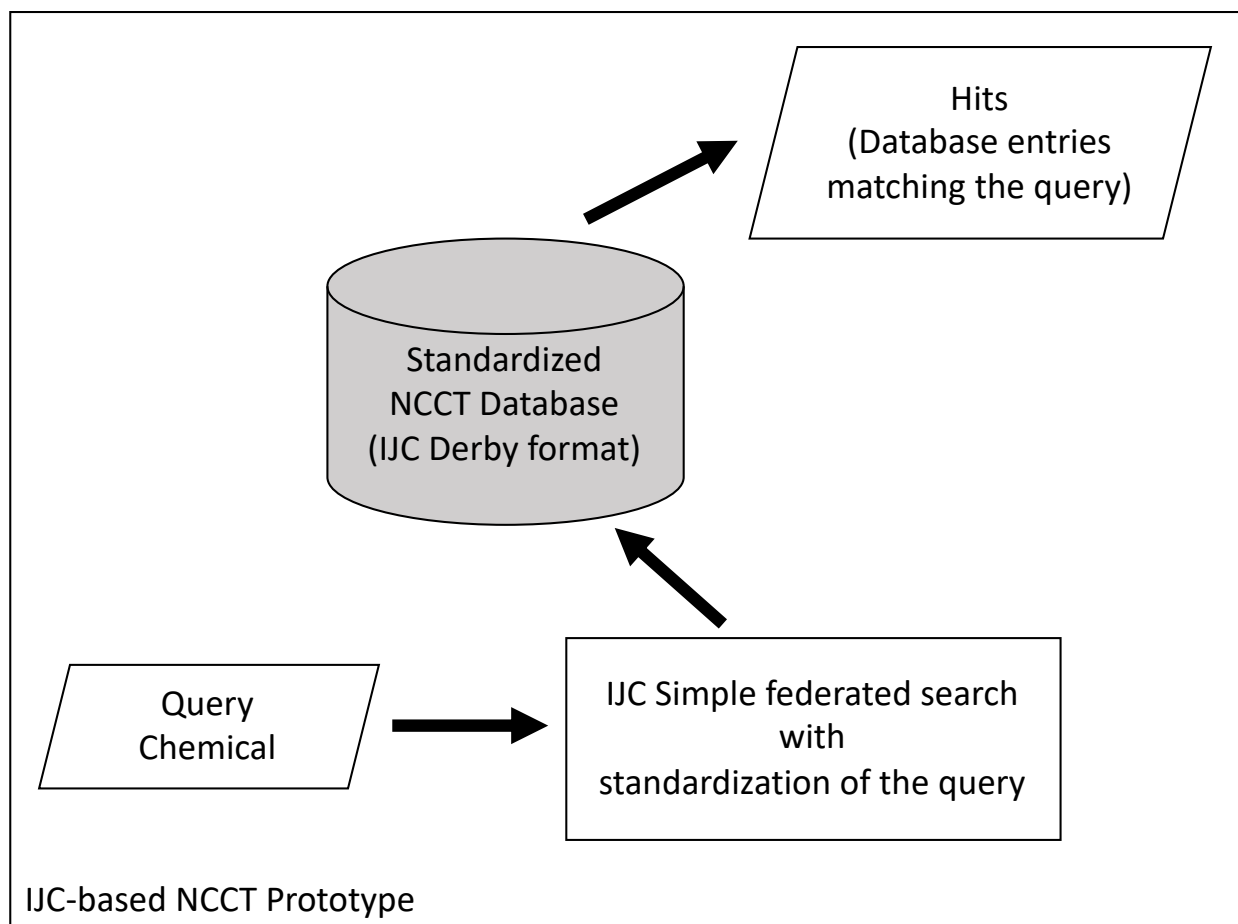
### A. Toxic Chemicals

	(CAS registry number)
(1) O-Alkyl (<=C10, incl. cycloalkyl) alkyl (Me, Et, n-Pr or i-Pr)-phosphonofluoridates	
e.g. Sarin: O-Isopropyl methylphosphonofluoridate	(107-44-8)
Soman: O-Pinacolyl methylphosphonofluoridate	(96-64-0)
(2) O-Alkyl (<=C10, incl. cycloalkyl) N,N-dialkyl (Me, Et, n-Pr or i-Pr) phosphoramidocyanidates	
e.g. Tabun: O-Ethyl N,N-dimethyl phosphoramidocyanidate	(77-81-6)
(3) O-Alkyl (H or <=C10, incl. cycloalkyl) S-2-dialkyl (Me, Et, n-Pr or i-Pr)-aminoethyl alkyl (Me, Et, n-Pr or i-Pr) phosphonothiolates and corresponding alkylated or protonated salts	
e.g. VX: O-Ethyl S-2-diisopropylaminoethyl methyl phosphonothiolate	(50782-69-9)
(4) Sulfur mustards:	
2-Chloroethylchloromethylsulfide	(2625-76-5)

**Fig. 1.** A snippet from of CWC Schedule 1 from the OPCW website (<https://www.opcw.org/chemical-weapons-convention/annexes/annex-chemicals/schedule-1>). The figure shows the first four entries of the schedule. The first three describe three families of chemicals, namely two families of nerve agents of the G-series and one family of nerve agents of the V-series. Specific examples are given for each of the three families. The fourth entry is an individually listed chemical belonging to the sulfur mustards family.

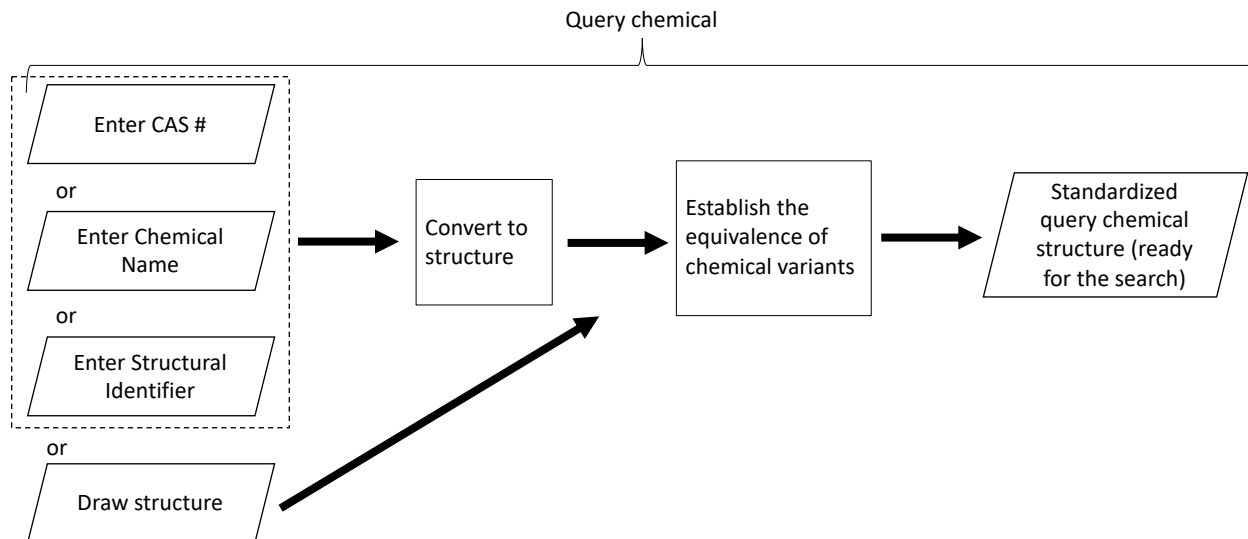


**Fig. 2.** A snapshot of the Simple Federated Search window of Instant JChem, showing the 5 CW-control lists, divided into 11 tables, implemented in the NCCT database.



**Fig. 3.** Flowchart showing the functioning of the IJC-based NCCT prototype. A query chemical, in various formats, is entered through the IJC interface (see **Fig.4** for further details). Through the IJC's Simple Federated Search function, the standardized query is checked against the standardized NCCT database, in Derby format. The database entries matching the query are retrieved, if any.

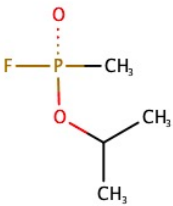




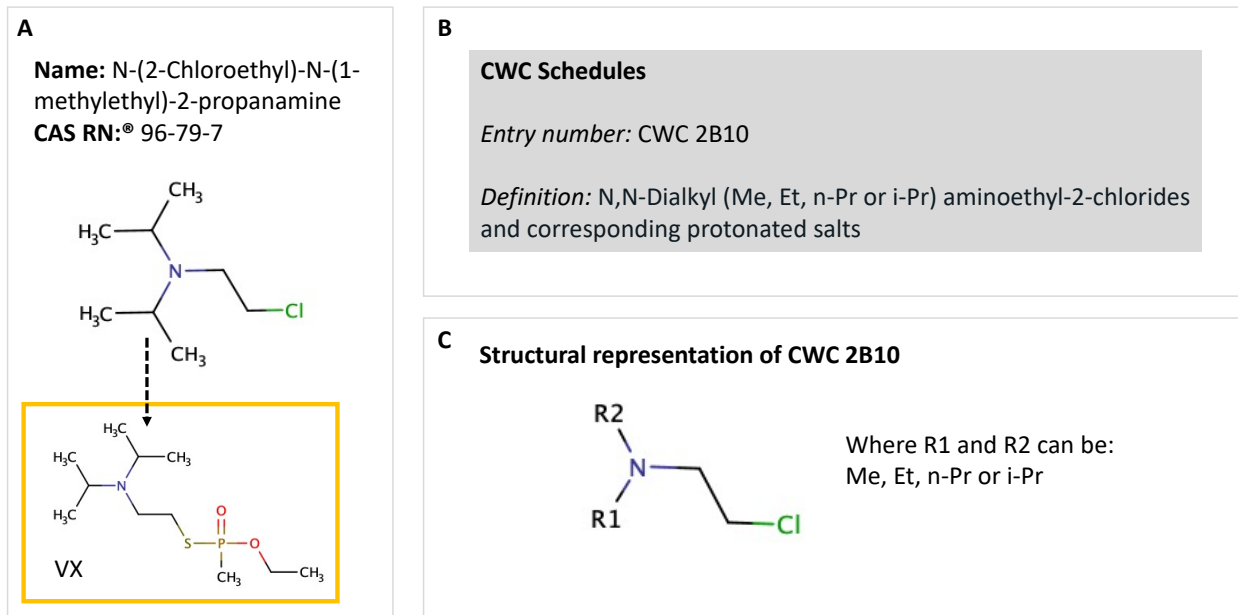
**Fig. 4.** A query chemical can be entered into IJC in a variety of ways, including by entering a CAS RN, a chemical name, or a structural identifier. The input will be converted into a 2D chemical structure. Alternatively, a 2D chemical structure can be sketched through the IJC interface. By subjecting the query to a standardization process, the equivalence of chemical variants is established. The standardized query is ready to be checked against the standardized NCCT database.

CW-Control Lists				
1	2	3	4	5
CWC Schedules (Schedule 1, Schedule 2, and Schedule 3)	Australia Group Chemical Weapons Precursors list	Wassenaar Arrangement Munitions List 7 (ML7)	European Union Council Regulation 36/2012 (Syria- related list)	Strategic chemicals identified in the World Customs Organization (WCO) Strategic Trade Control Enforcement Implementation Guide (STCE)

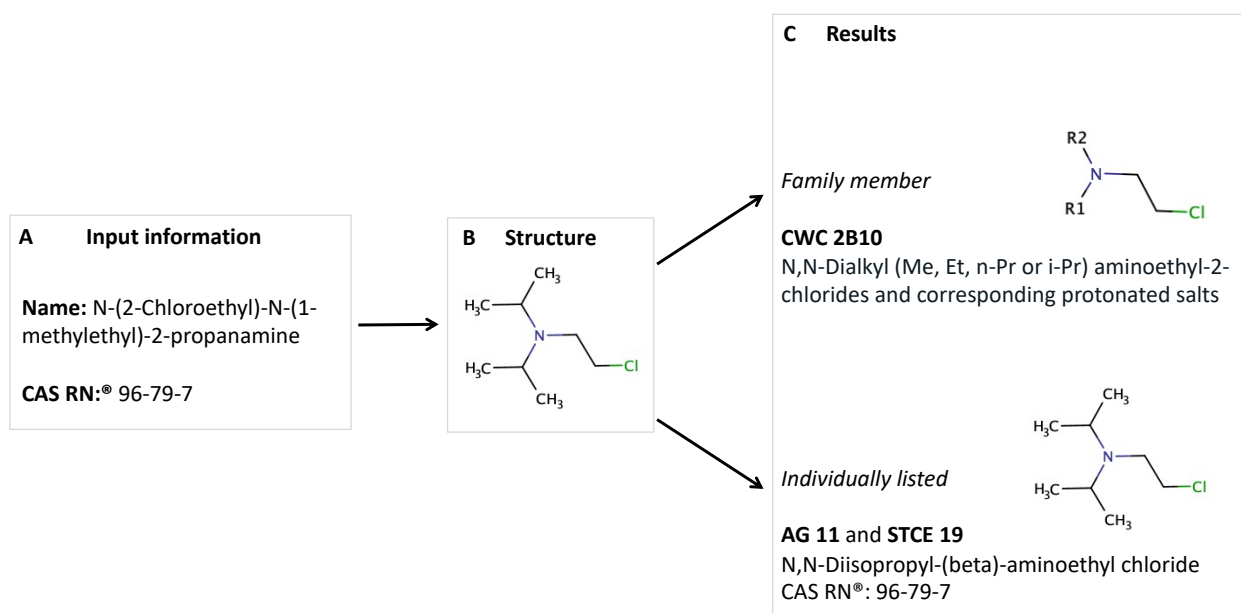
**Fig. 5.** The CW-control lists implemented into the NCCT database.

Markush structure		Entry Number	Entry Name	CAS Registry Number
		CWC 1A1 Example 1	Sarin: O-Isopropyl methylphosphonofluoridate	107-44-8
SMILES	InChI	InChIKey	PubChem ID	PubChem URL
CC(C)O[P](C)(F)=O	InChI=1S/C4H10FO2P/c1-4(2)7-8(3,5)6/h4H,1-3H3	InChIKey=DYAHQFWOVKZOW-UHFFFAOYSA-N	7871	<a href="https://pubchem.ncbi.nlm.nih.gov/compound/7871">https://pubchem.ncbi.nlm.nih.gov/compound/7871</a>
Overlaps	Category	Entry Type		
WA ML7 b.1.a Example 1; STCE 28	Nerve agents	Family_Example		

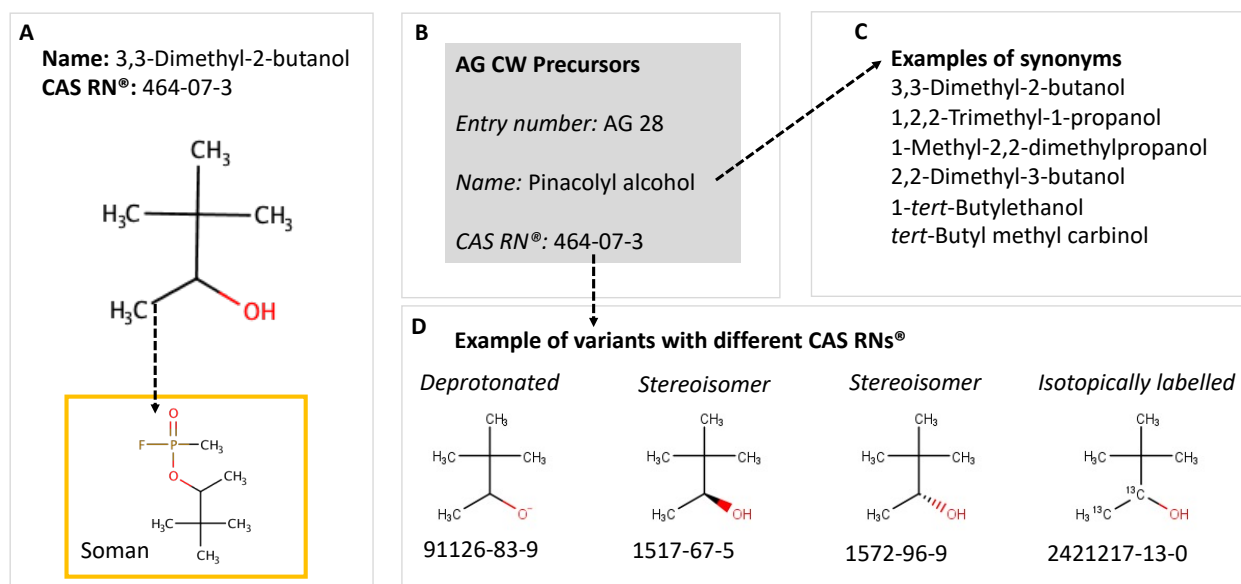
**Fig. 6.** The 11 fields featured in the NCCT database. Entry 1A1 of CWC Schedule 1 is given as an example.



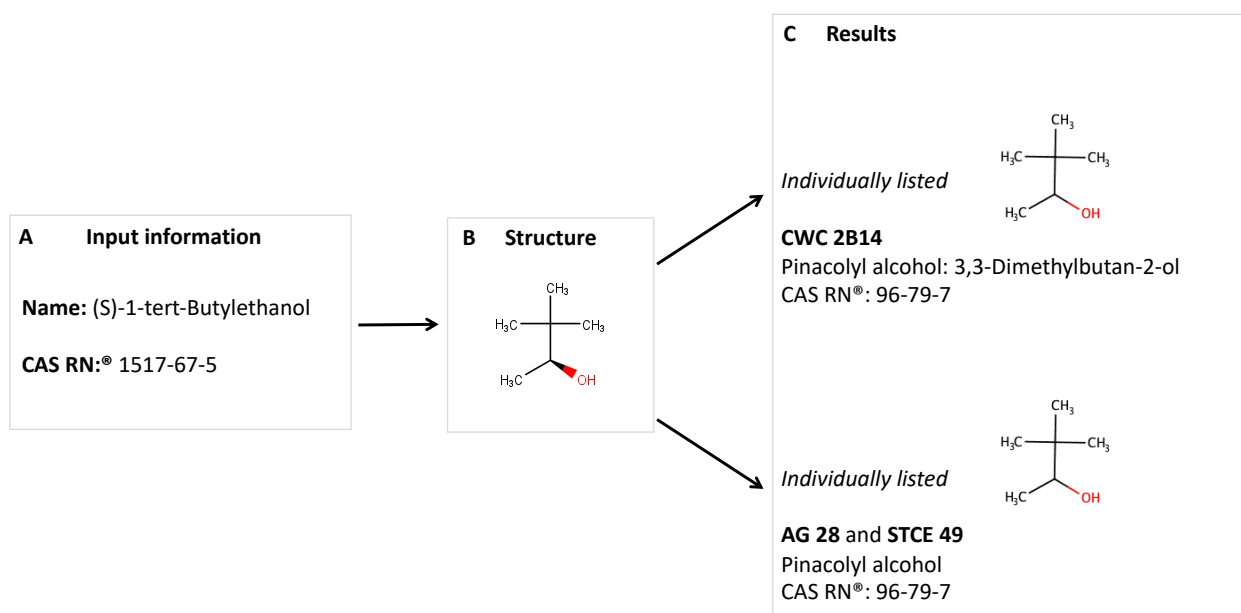
**Fig. 7.** Panel A: A chloroethylamine that can be employed as a precursor for the synthesis of the nerve agent VX. Panel B: the CWC Schedule 2 entry that describes the family of chemicals that encompasses the chloroethylamine shown in Panel A. Panel C: structural representation of the CWC Schedule 2 family shown in Panel B.



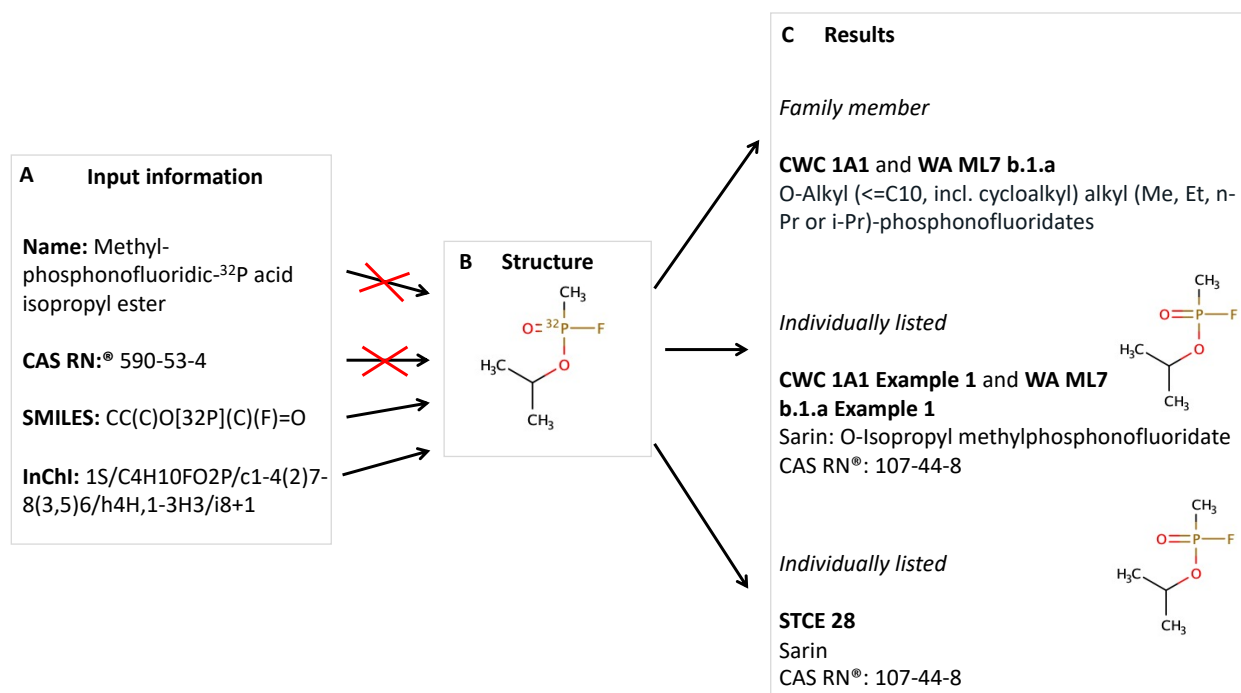
**Fig. 8.** A chloroethylamine VX precursor is entered into IJC either through its name or its CAS RN (Panel A). The query is converted into a structure (Panel B), is standardized, and is checked against the NCCT database. The results will indicate whether the entered chloroethylamine is covered is a member of the family described by CWC Schedule 2 entry 2B10 and is individually listed by the AG and WCO as entry AG 11 and STCW 19, respectively.



**Fig. 9.** Panel A: Pinacolyl alcohol, a chemical that can be employed as a precursor for the synthesis of the nerve agent soman. Panel B: the AG list entry that describes pinacolyl alcohol. Panel C: examples of chemical names synonymous of pinacolyl alcohol. Panel D: examples of chemical variants of pinacolyl alcohol with CAS RNs different from the one found in the AG list.



**Fig. 10.** The (S) isomer of pinacolyl alcohol is entered into IJC either through one of its synonyms or its CAS RN (Panel A). The query is converted into a structure (Panel B), is standardized, and is checked against the NCCT database. The results will indicate that pinacolyl alcohol is listed as an individual chemical by CWC Schedule 2, as entry 2B14, the AG list, as entry 28, and the WCO STCE list, as entry 49. Note how different names are provided for pinacolyl alcohol by CWC Schedule 2 on one hand and the AG and WCO STCE lists on the other hand.



**Fig. 11.** A <sup>32</sup>P-labelled version cannot be entered into IJC by name or CAS RN. However, it can be entered as a SMILES string or an InChI code (Panel A). The query is converted into a structure (Panel B), is standardized, and is checked against the NCCT database. The results will indicate that sarin is covered by the CWC Schedules, as member of family CWC 1A1 (for which it is also listed as a specific example), by the Wassenaar Arrangement ML7, as a member of family b.1.a (for which it is also listed as a specific example), and by the World Customs Organization STCE list, as entry STCE 28 .