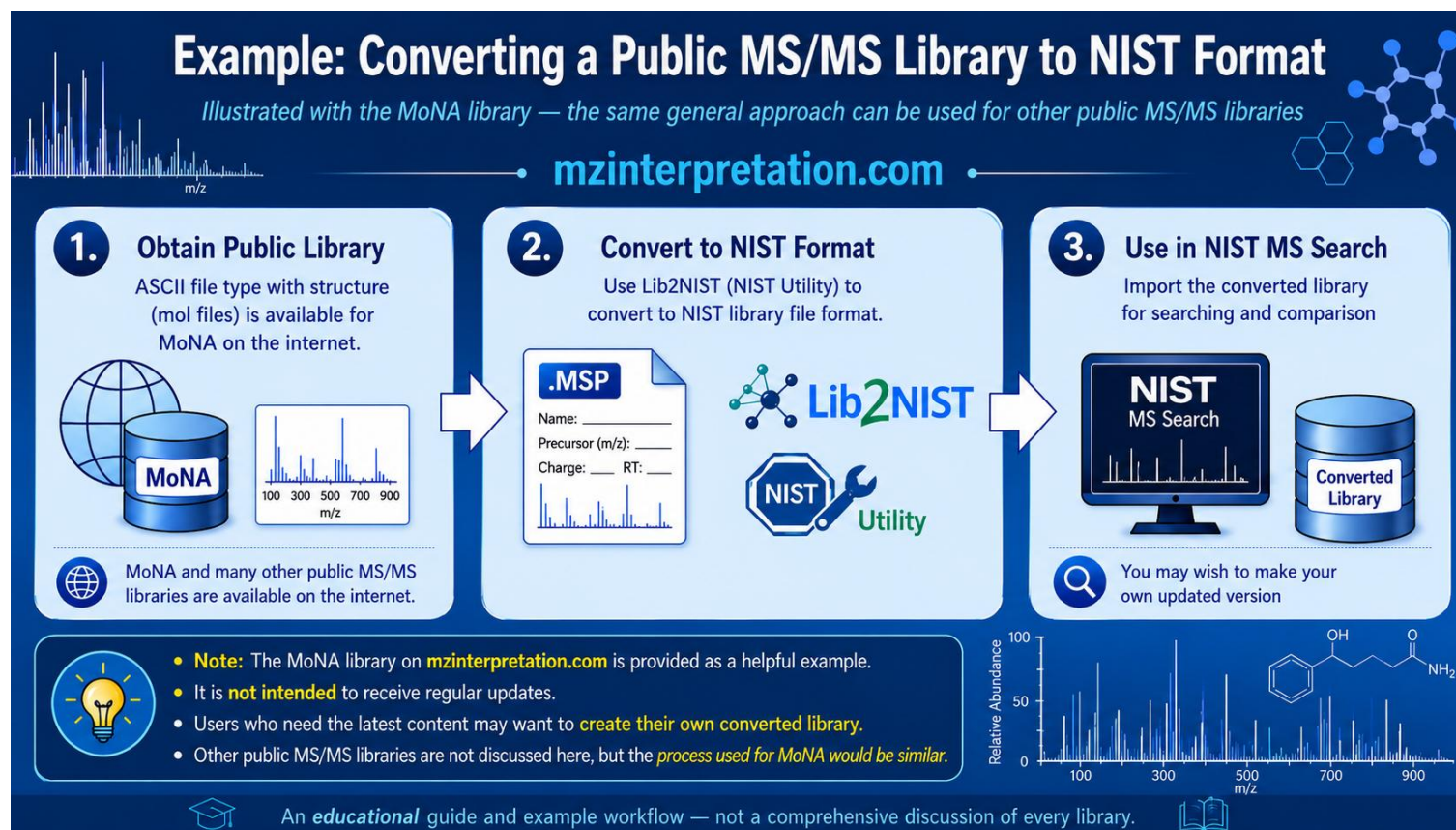


Creating Public MS/MS Libraries in NIST Format*

Video and Handout

James Little
Mass Spec Interpretation Services
8/21/2026

*Written for NIST26, but also
substituted for NIST23 Pt. VII
Video/Handout



See **Full Course** on NIST26 with new **Integrated** Deconvolution/Library Searching for **LC-MS/MS**!

[link to course](#)

Introduction

- Two public libraries, MoNA and MS-DIAL, currently available on my website in NIST format
- MS-DIAL-pos, MS-DIAL-neg, and MoNA LC-MS/MS
- Together, >500K spectra with >150K associated structures, [see link](#)
- Libraries are described in this [link](#)
- Made available as a basic resource
- These libraries will not be routinely updated
- Thus, user might want to create their own
- A MoNA library conversion demonstrated in this video/handout as a general example
- Also, NIST has a list of others that could be converted into NIST format
- See NIST full “List of Sources” of Free Public MS/MS Libraries on my website resources, [see link](#)

Status of Public Libraries Employed in My Work

- Originally, I downloaded MoNA LC-MS/MS Spectra as a group of ~153,000, see screenshot below
- Unfortunately, MoNA have now combined with Enveda-180, change made July 2026
- See a ChatGPT summary on next slide
- The file is very large, 2.7 GB, and cannot be opened by most Windows program
- One might be able to create a NIST formatted library from the SDF file using Lib2NIST
- But, **Enveda are Enamine REAL-space drug-like compounds**, so it may contain many synthetic structures that would rarely appear in routine environmental, metabolomics, pharmaceutical, or forensic samples.
- Thus, I decided not to try to convert this file
- On Windows, the easiest approach is to use **PowerShell** to stream through the file one record at a time and write a new SDF without Enveda-180. That way Windows never loads the entire 2.7 GB file into memory
- This approach described from my prompt in ChatGPT, but I decided not to attempt
- The MS-DIAL could be created with Lib2NIST, but with no structures
- MoNA [Link](#)
- MS-DIAL [Link](#)



ChatGPT Description of New MoNA Combined Library

Enveda-180 is a **real experimental MS/MS library**, not predicted spectra. It contains about **1.16 million spectra from 184,330 different small molecules**. The compounds were measured in positive and negative modes and at multiple collision energies, which is why there are several spectra per compound. It also includes retention time and ion-mobility/CCS information.

For library searching, its biggest advantage is **coverage**. The Enveda authors point out that existing open libraries such as GNPS, MassBank and MSnLib together cover roughly 150,000 unique compounds; Enveda-180 alone adds another ~184,000 structurally diverse drug-like compounds.

There are some practical downsides for what you are doing, though:

- **Huge size:** 1,158,293 spectra; the SDF is about **2.4 GB**.
- Many compounds have **multiple spectra/adducts/collision energies**, so the raw spectrum count sounds much larger than the unique-compound count.
- It was generated from **Enamine REAL-space drug-like compounds**, so it may contain many synthetic structures that would rarely appear in routine environmental, metabolomics, pharmaceutical, or forensic samples.
- Adding 1.16 million spectra will make your converted NIST library substantially larger and may make searching/indexing slower.
- MoNA offers a **separate download** of just the Enveda-180 (1,158,293 spectra)

Current LC-MS/MS Library on MoNA (see [link](#))

LC-MS/MS Spectra (1,317,257 spectra)

🔍 Browse associated spectra

Download

📄 JSON — Internal MoNA Format (3.0 GB)

📄 MSP — NIST Compatible (2.6 GB)

📄 SDF — NIST Compatible (2.7 GB)

📄 Enveda-180 (1,158,293 spectra)

Google AI Information on Enveda-180

See [Link](#) to paper.

The Enveda-180 library consists of **experimentally measured mass spectrometry data**, not *in-silico* simulations.

Why ChatGPT's Conclusion is Correct

The evidence from the primary data repositories explicitly confirms the experimental nature of the dataset:

- **Raw Instrument Files:** As seen in the Enveda-180 GitHub repository layout, the processing pipeline begins with raw Bruker .d files. A .d folder is the proprietary raw data format output directly by Bruker mass spectrometers. You cannot possess these files unless an injection was physically run on an instrument.
- **The Physical Instrument Used:** The public deposition entry on MassIVE (ID: **MSV000102122**) explicitly documents that the library was generated using reference standards and synthetic/isolated compounds analyzed on a physical **timsTOF Pro 2** mass spectrometer.
- **Data Scale:** The library captures roughly 1.158 million spectra derived from 184,330 physical drug-like compounds, tracking various experimental parameters like diverse collision energies, polarities, and adduct formations.

Where the "Computational" Misunderstanding Happens

The confusion typically stems from how the project was designed, as it relied heavily on computational workflows in three specific areas:

1. **The Target Chemical Space:** The selection of molecules was driven by the **Enamine REAL space**. This is a massive, computationally generated universe of billions of virtual, synthesizable compounds. Enveda utilized advanced computational selection algorithms to decide *which* actual compound pools to procure and physically test out of that massive space.
2. **Pooled Deconvolution:** To efficiently analyze 184,330 compounds without running 184,330 separate, costly experiments, Enveda ran them in **mixtures/pools**. They then used a computational package called `ms_prepper` to digitally extract, unpack, and assign the raw spectral peaks back to the individual molecules in the mixture.
3. **Defensive Structural Annotation:** Computational tools like **SIRIUS** were layered on top *after* the fact to annotate the chemical formulas of the fragments, but the code explicitly states that the pipeline completes independently of these predictions.

What This Means for Your NIST User Library

Because these are authentic, experimentally derived spectra measured on a widely used high-resolution instrument (timsTOF Pro 2), **Enveda-180 is highly eligible and structurally valuable for inclusion in a NIST user library**. Unlike *in-silico* libraries (which can suffer from artificial fragment intensities), this library contains real-world instrument responses, peak shapes, and real collision-energy fragmentation behavior.

Comparison of NIST and Several Public Libraries

Library source	What it tends to represent	
Enveda-180	Large collection of experimentally measured synthetic, drug-like medicinal-chemistry compounds from Enamine REAL chemical space	
MoNA traditional entries	Broad mix of metabolites, drugs, natural products, environmental chemicals, standards, and miscellaneous compounds contributed from many sources	
MS-DIAL	Strongly oriented toward metabolomics and lipidomics , with many metabolites, lipids, natural products, plant compounds, drugs, and related biological molecules	
NIST MS/MS	Broad curated reference collection covering drugs, metabolites, environmental chemicals, natural products, toxicology compounds, and many other small molecules	

MS-DIAL = metabolomics/lipidomics-heavy

MoNA = very broad community collection

Enveda-180 = synthetic drug-like chemical space

NIST = broad, curated reference library

User Could Consider Creating Separate Libraries

These Could be Downloaded Separately be Generally Useful

MoNA subset	Spectra	SDF size	General-purpose value
All LC-MS/MS Agilent QTOF	24,961	43.0 MB	Good, but relatively small
All LC-MS/MS Orbitrap	159,463	243.5 MB	Much broader coverage

Current Sub-List of LC-MS/MS Library on MoNA

📁 LC-MS Spectra (1,328,049 spectra)

📁 LC-MS/MS Spectra (1,317,257 spectra)

📄 All LC-MS/MS Agilent QTOF (24,961 spectra)

📄 LC-MS/MS Negative Mode (313,537 spectra)

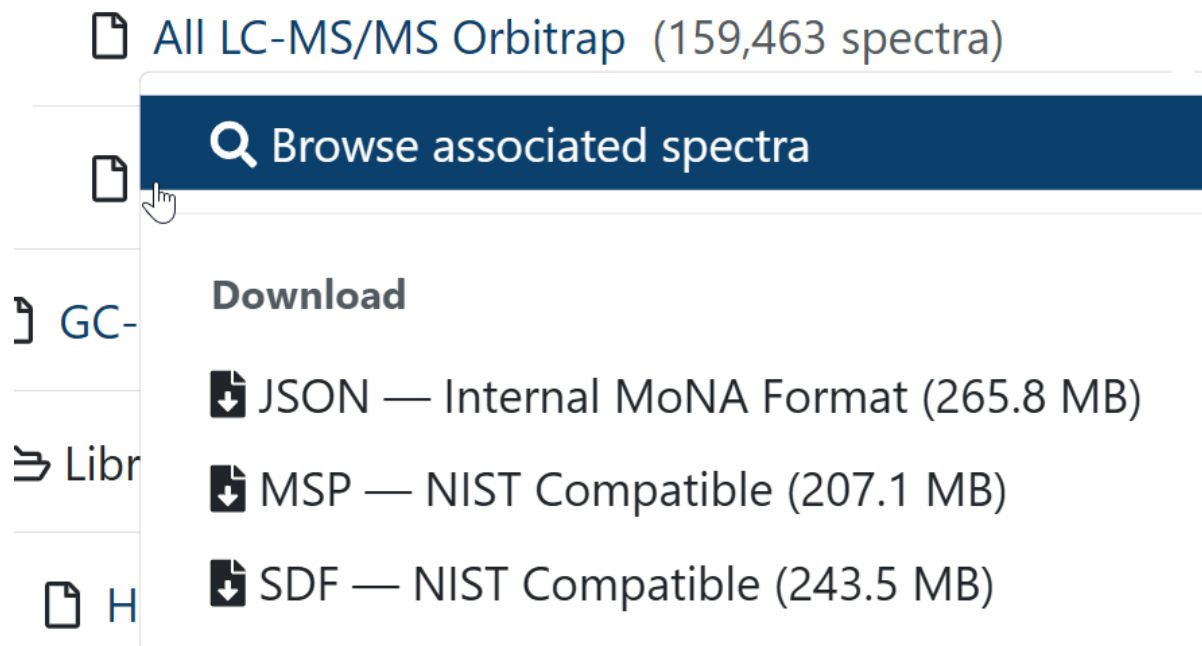
📄 All LC-MS/MS QTOF (1,210,513 spectra)

📄 All LC-MS/MS Orbitrap (159,463 spectra)

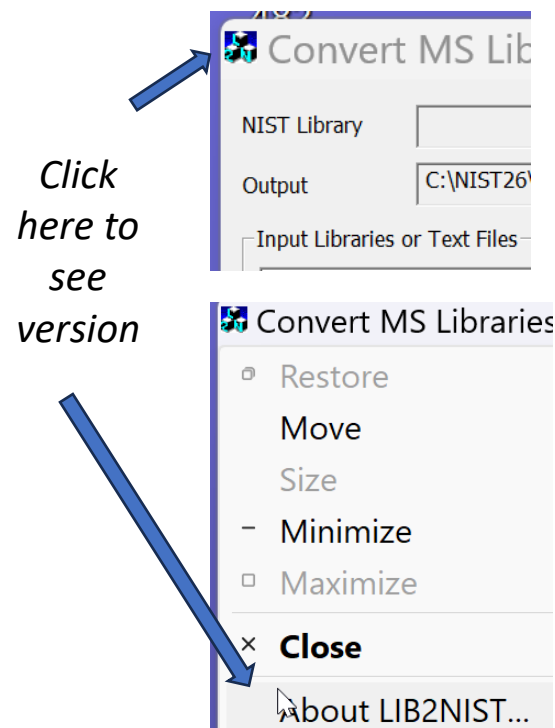
📄 LC-MS/MS Positive Mode (1,000,332 spectra)

Steps in Conversion Process

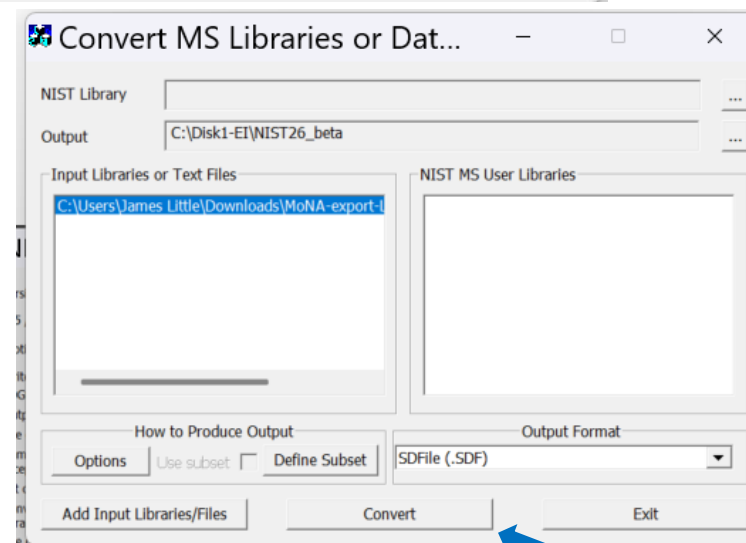
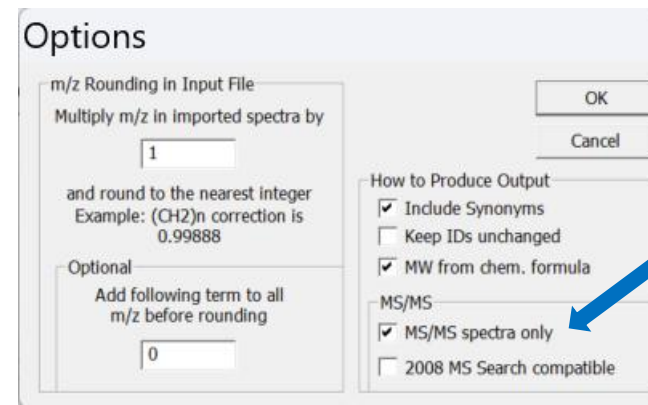
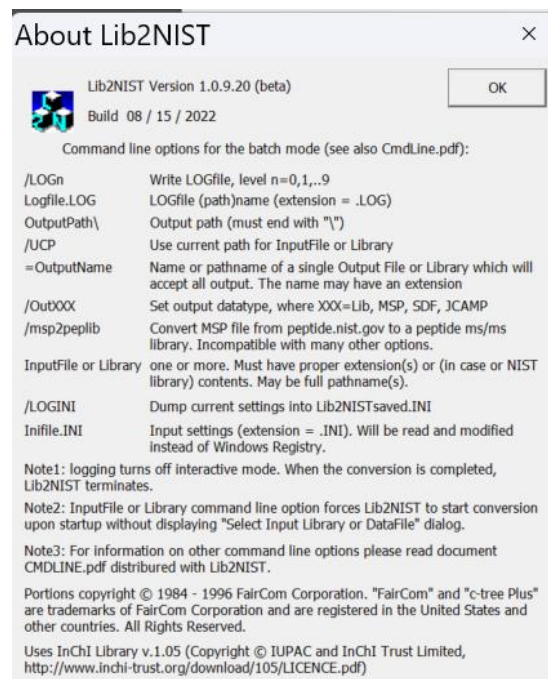
- Demonstrated with All LC-MS/MS Orbitrap (159,463 spectra)
- Download SDF File from MoNA website, SDF file includes spectra and structures in zip file
- Link to MoNA website
- MSP would ONLY be spectra
- Unzip the SDF file
- Convert to NIST format with Lib2NIST utility and output to your library search software MSSearch folder
- If not in proper MSSearch folder, e.g. C:/NIST26/Mssearch
- Then place in proper folder for accessing by NIST search program



Lib2NIST Windows: Always Use Newest Version of Lib2NIST

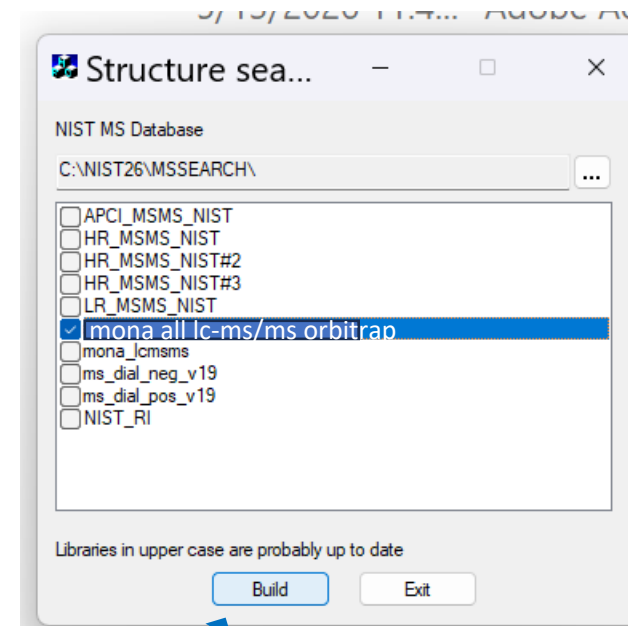
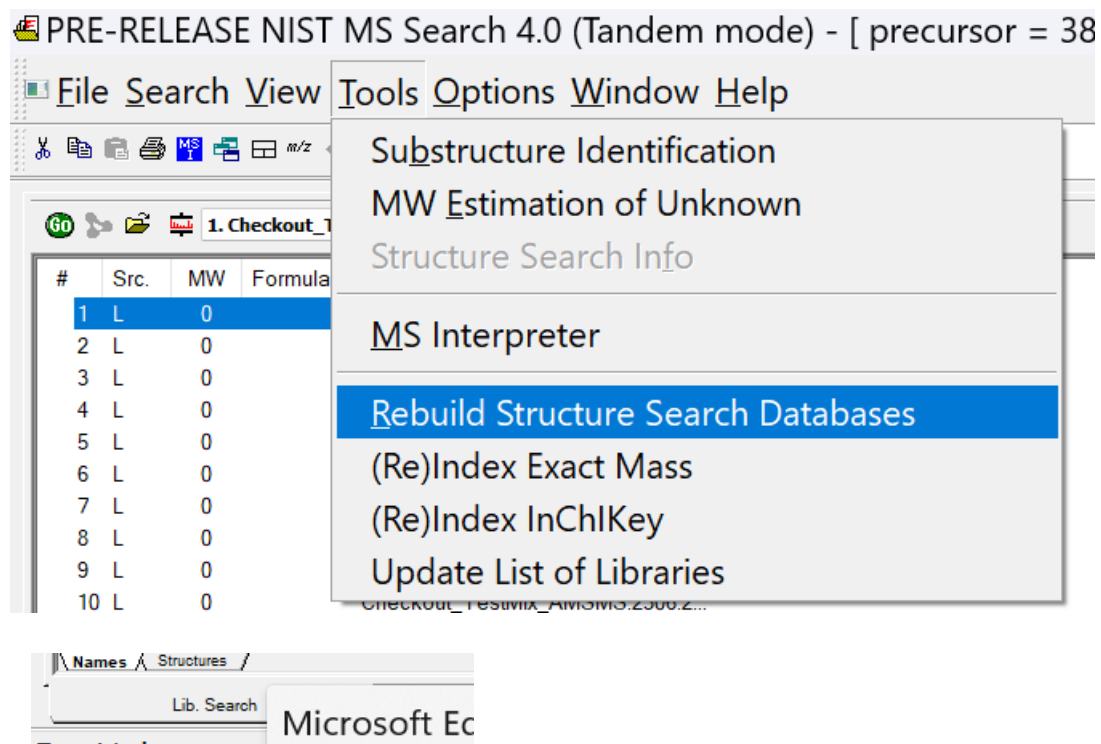


Version with NIST26



Updating MoNA LC-MS/MS Library in NIST Format

- After converting from SDF to NIST formatted library
- In Lib Search Tab, Tools/Rebuild Structure Search Database
- If not updated, structures will be displayed with search results
- **However**, search by structure WILL NOT work



“OK Problem:” Creating Update MoNA LC-MS/MS Library in NIST Format

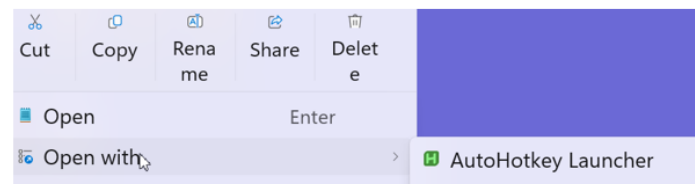
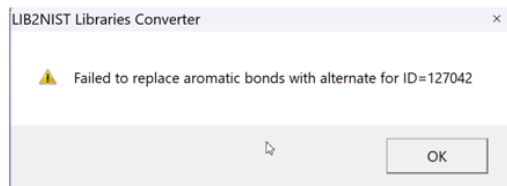
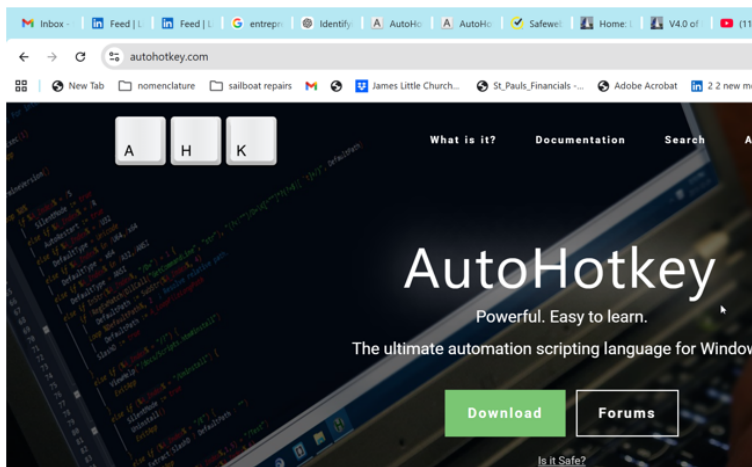
Problem with Structures, Need to select OK button hundreds of time, see automatic script below:

Auto “OK” for NIST Lib2NIST for Mono SDF File Conversions

- Converting the Mona LC-MS/MS SDF file to NIST file format with Lib2NIST
- Process requires dismissing several 100's of windows
- Typical Error Message shown below
- AutoHotkey with a Script will close these windows automatically
- ChatGPT suggested approach from description of problem
- Wrote a script and save as file type .ahk
- Right click on the script, and Open with AutoHotkey Launcher
- Automatically closes these error message windows

Stopping the Script:

1. Look in the lower-right corner of Windows near the clock.
 2. Find the green AutoHotkey icon (usually a green square with a white “H”).
 3. Right-click the icon.
 4. Select **Exit**.
- That immediately stops the script.



#Requires AutoHotkey v2.0

```
Loop
{
    Send "{Enter}"
    Sleep 700
}
```

Good Practice to Test Public Library after Building in NIST Format

Test the following:

- Identity MSMS (with Precursor)
- Identity HiRes No Precursor
- Identity EI Normal
- MS/MS Hybrid
- Structure Search

