

## *Dashboard for the new PhytoChem-Tracker tool*

```
# Natural Products Chemistry Dashboard Design Guide  
## Comprehensive Resource Roadmap for Optimal UI/UX Development
```

---

### ## Executive Summary

This guide provides a comprehensive roadmap for designing and developing an interactive software dashboard for natural products chemists, focusing on molecular formula/structure/name input and rapid extraction of chemical characteristics. It synthesizes best practices from successful implementations, UI/UX standards, cheminformatics APIs, and academic research.

---

### ## Part 1: Reference Implementations - Learn from Existing Success

#### ### 1.1 COCONUT (COLleCtion of Open Natural ProdUcTs)

**\*\*URL:\*\*** <https://coconut.naturalproducts.net>

**\*\*Key Learnings:\*\***

- Uses MongoDB for storage management (first large chemical database to do so)
- Modern, user-friendly interface prioritizing "pleasant user experience"
- Multiple search modalities: structure, name, molecular features
- Pre-computed molecular descriptors for rapid retrieval
- No login required - reduced friction for user adoption
- Implements ChemDoodle Web Components (JavaScript library) for structure rendering

**\*\*Technical Architecture:\*\***

- Frontend: Modern web-based interface
- Backend: MongoDB storage system
- Chemistry interface: ChemDoodle Web Components (<https://web.chemdoodle.com/>)
- Browser requirements: JavaScript-capable (tested on Chrome, Firefox)

**\*\*Key Design Decision:\*\*** Aggregate data from multiple sources with consistent annotation, making it a "one-stop shop" rather than requiring users to visit multiple databases.

**\*\*Publication:\*\*** Sorokina et al. (2021), Journal of Cheminformatics

---

### ### 1.2 SuperNatural 3.0

**\*\*URL:\*\*** <https://supernatural.bioinf.uni-sb.de> (check for current availability)

**\*\*Key Learnings:\*\***

- Seven modular function modules for different user workflows:
  - Compound Search
  - Mechanism of Actions
  - Pathways
  - Target Library
  - Disease Indication
  - COVID-19 virtual screening
  - Organoleptic properties
- Four search options: Name/IDs, Structure Search, Similarity Search, Advanced Search
- Uses ChemDoodle Web Components for chemistry interface
- Integrates toxicity, vendor information, drug-like chemical space predictions

**\*\*Design Philosophy:\*\*** Modular architecture allows users to access specific functionality based on their research needs without overwhelming the interface.

---

### ### 1.3 Chemotion ELN (Electronic Lab Notebook)

**\*\*URL:\*\*** <https://chemotion.net>

**\*\*Key Learnings:\*\***

- Integrates Ketcher editor with advanced customizations
- Provides immediate feedback - identifiers/names available directly in UI
- Real-time calculations (molecular weight, exact mass) without saving
- Template library management (450+ structures including functional groups, salts, solvents)
- Supports coordinative bonds for organometallic structures

**\*\*Critical UX Feature:\*\*** Information available "in-UI" during editing without requiring save/reload cycles - eliminates workflow interruptions.

---

### ### 1.4 GenUI (Generator User Interface)

**\*\*URL:\*\*** <https://genui.imtm.cz>

**\*\*Key Learnings:\*\***

- Web-based GUI + REST API for both human users and automated processes
- Integrates solutions for import, generation, storage, retrieval, and visualization
- Documented Python API for easy integration of new features
- Extensive library of reusable UI components
- Focus on making advanced cheminformatics accessible to non-experts

**\*\*Design Philosophy:\*\*** Dual interface (GUI + API) enables both interactive use by chemists and programmatic access for pipeline integration.

---

### ### 1.5 Datagrok Cheminformatics Platform

**\*\*URL:\*\*** <https://datagrok.ai>

**\*\*Key Learnings:\*\***

- Info panes system: customizable panels showing different aspects of chemical data
- Multiple viewer types: scatterplot, network diagram, tile viewer, bar chart, form viewer

- Chemical space visualization using tSNE/UMAP with fingerprint-based dimensionality reduction
  - Interactive UI elements that update dynamically for current molecule
  - All viewers can be saved as part of layouts or dashboards
- \*\*Critical Feature:\*\* Extensible architecture - custom viewers and functions can be added in Python (RDKit) or R.

---

## ## Part 2: Core UI/UX Design Principles

### ### 2.1 Nielsen's 10 Usability Heuristics (Industry Standard)

\*\*Source:\*\* Nielsen Norman Group ([nngroup.com/articles/ten-usability-heuristics](http://nngroup.com/articles/ten-usability-heuristics))

Apply these principles specifically to cheminformatics dashboards:

#### #### H1: Visibility of System Status

- \*\*For Chemistry:\*\* Show structure validation status in real-time
- Progress indicators for structure recognition, descriptor calculations
- Display current input mode (SMILES, InChI, name, sketch)
- Indicate database connection status

#### #### H2: Match Between System and Real World

- \*\*For Chemistry:\*\* Use chemical terminology familiar to target users
- "Molecular Weight" not "MW" for general users; abbreviations OK for experts
- Display properties in standard units (g/mol, Å, kcal/mol)
- Follow IUPAC conventions for naming and stereochemistry

#### #### H3: User Control and Freedom

- \*\*For Chemistry:\*\* Undo/Redo for structure editing
- Clear "reset" or "clear structure" buttons

- Save partial searches/queries
- Export data in multiple formats (SDF, SMILES, CSV, PDF)

#### #### H4: Consistency and Standards

- **For Chemistry:**
  - Follow established molecular editor conventions (e.g., ACS drawing style as option)
  - Use standard icons: "+" for add, magnifying glass for search, pencil for edit
  - Maintain consistent color coding (e.g., stereochemistry indicators)
  - Jakob's Law: Users spend most time on other sites - match their expectations

#### #### H5: Error Prevention

- **For Chemistry:**
  - Validate SMILES/InChI before submission
  - Warn about uncommon valence states
  - Confirm before deleting saved structures
  - Prevent invalid stereochemistry assignments

#### #### H6: Recognition Rather Than Recall

- **For Chemistry:**
  - Template library (ring systems, functional groups) readily visible
  - Recent searches/structures accessible
  - Auto-complete for compound names
  - Visual previews of structures in search results

#### #### H7: Flexibility and Efficiency of Use

- **For Chemistry:**
  - Keyboard shortcuts for power users (copy structure: Ctrl+C, paste: Ctrl+V)

- Quick-access toolbar for frequent operations
- Batch processing capabilities
- Multiple input methods (draw, name, paste SMILES, upload file)

#### #### H8: Aesthetic and Minimalist Design

- \*\*For Chemistry:\*\*
  - Clean structure visualization without clutter
  - Collapsible panels for advanced options
  - White space around complex structures
  - Hide less-common descriptors by default (show on demand)

#### #### H9: Help Users Recognize, Diagnose, and Recover from Errors

- \*\*For Chemistry:\*\*
  - "Invalid SMILES syntax at position 12" rather than "Syntax Error"
  - Suggest corrections: "Did you mean 'phenol' instead of 'phenyl'?"
  - Highlight problematic atoms in structure editor
  - Provide links to help docs for specific error types

#### #### H10: Help and Documentation

- \*\*For Chemistry:\*\*
  - Tooltips explaining descriptor meanings
  - Quick-start guide for structure input
  - Examples of valid input formats
  - Video tutorials for complex operations

---

### ## 2.2 Dashboard-Specific Heuristics

\*\*Source:\*\* Dowding & Merrill (2018), Journal of the American Medical Informatics Association

Adapted for chemical dashboards:

#### Specific Visualization Principles:

1. \*\*Data-Ink Ratio:\*\* Maximize information per pixel - avoid unnecessary gridlines, borders
2. \*\*Appropriate Visualization Type:\*\*
  - Use structure diagrams for molecular data (not tables of SMILES)
  - Scatter plots for property relationships
  - Histograms for distribution analysis
  - Network graphs for biosynthetic pathways
3. \*\*Interactive Elements:\*\*
  - Hover tooltips for descriptor values
  - Click to expand/collapse detail panels
  - Drag-and-drop for structure comparison
4. \*\*Responsive Design:\*\*
  - Layouts adapt to screen size
  - Mobile considerations for field work
  - High-DPI support for detailed structures

---

## Part 3: Molecular Structure Input Systems

### 3.1 Molecular Editor Selection

#### Option A: Ketcher (Open Source, Most Customizable)

\*\*Source:\*\* <https://github.com/epam/ketcher>, EPAM LifeSciences

\*\*License:\*\* Apache 2.0

**\*\*Capabilities:\*\***

- Small molecules AND biomolecules in single drawing
- 450+ template structures (functional groups, salts, solvents)
- All common file formats: MDL Molfile V2000/V3000, SMILES, InChI, CDXML, FASTA, HELM
- Reaction support with mapping
- Stereochemistry: auto-assign R/S, E/Z labels
- 2D and 3D depictions
- ACS document style support
- JavaScript API for programmatic control

**\*\*Integration Methods:\*\***

1. React component library (modern apps)
2. IFrame embedding (simple integration)
3. Standalone page

**\*\*Key API Methods:\*\***

```javascript

```
ketcher.getSmiles(isExtended) // Get SMILES string  
ketcher.getMolfile(molfileVersion) // Get MDL Molfile  
ketcher.getInchi(withAuxInfo) // Get InChI  
ketcher.setMolecule(structureString) // Load structure
```

...

**\*\*Events:\*\***

- `change` - Structure modified
- `libraryUpdate` - Monomer library changed

**\*\*Customization:\*\*** Can hide toolbar buttons, modify appearance parameters, extend functionality

**\*\*Best For:\*\*** Open source projects, full customization control, academic use

---

#### Option B: Marvin JS (Commercial, Feature-Rich)

**\*\*Source:\*\*** <https://chemaxon.com/marvin>

**\*\*License:\*\*** Commercial (free for academic use)

**\*\*Capabilities:\*\***

- Web-based with rich feature set
- Universal search bar (find tools, templates, name-to-structure)
- Industry-standard chemical intelligence (error detection, stereochemistry)
- Live calculations and predictions
- WCAG 2.2 AA accessibility certified (partial)
- Screen reader support, keyboard navigation
- Seamless Microsoft Office integration
- Journal style presets (ACS, RSC, Nature, etc.)

**\*\*Best For:\*\*** Enterprise applications, accessibility requirements, Office integration

---

#### Option C: ChemDoodle Web Components (JavaScript, Balanced)

**\*\*Source:\*\*** <https://web.chemdoodle.com>

**\*\*License:\*\*** GNU GPL v3.0 or proprietary

**\*\*Capabilities:\*\***

- Pure JavaScript (no plugins required)
- Publication-quality 2D and 3D graphics

- Animation support
- Mobile-ready (touch gestures)
- Fast rendering for large datasets
- Used by COCONUT and SuperNatural 3.0

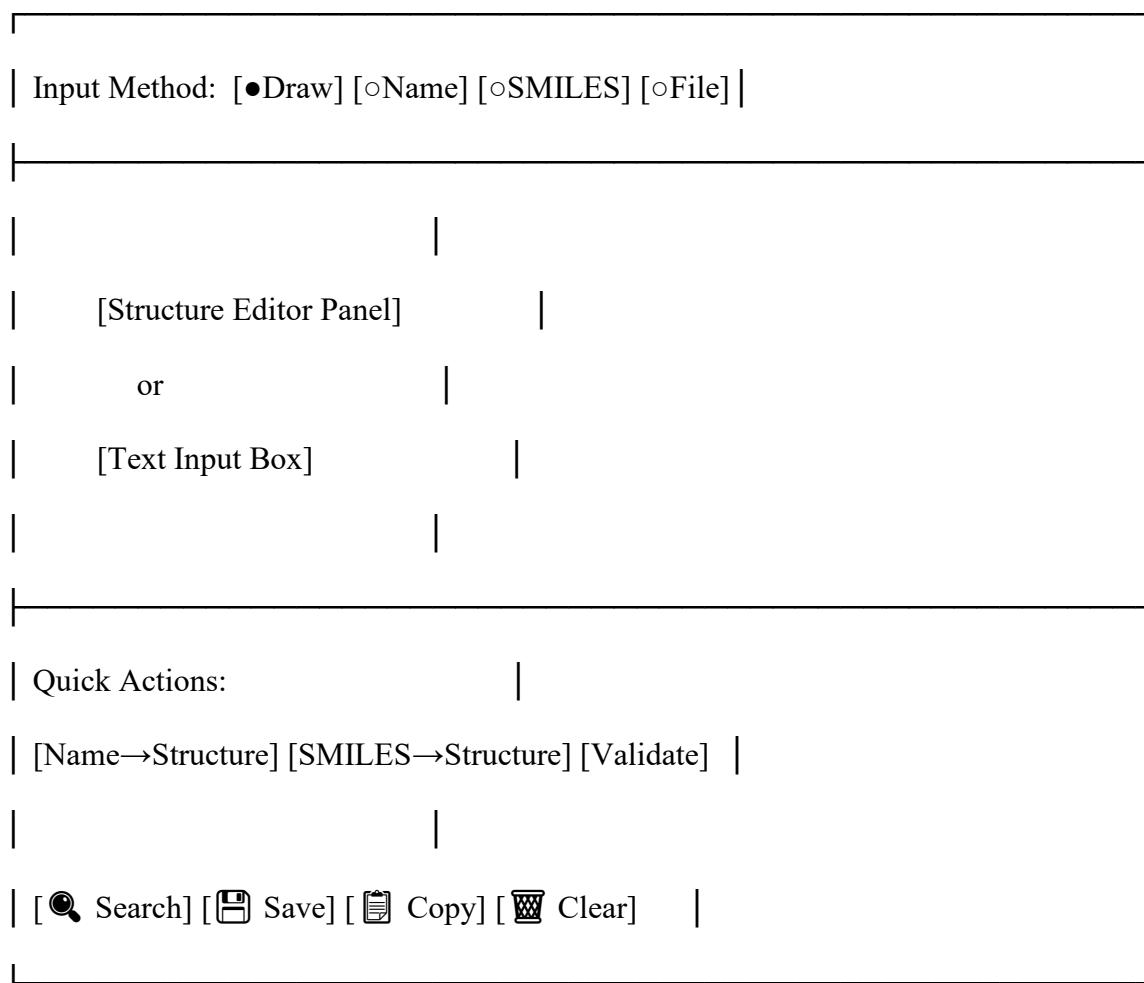
**\*\*Best For:\*\*** Web applications requiring high-quality rendering, spectral data visualization

---

### ### 3.2 Multi-Format Input Strategy

**\*\*Recommended Input Panel Design:\*\***

...



---

## \*\*Input Format Support (Priority Order):\*\*

1. \*\*Structure Drawing\*\* (most intuitive for chemists)
2. \*\*Name Input\*\* (IUPAC, common names)
3. \*\*SMILES\*\* (canonical and isomeric)
4. \*\*InChI/InChIKey\*\*
5. \*\*File Upload\*\* (SDF, MOL, PDB, CIF for natural products)
6. \*\*Image Recognition\*\* (using DECIMER - see Section 4.3)

---

## ## Part 4: Backend Architecture & APIs

### ### 4.1 Cheminformatics Microservice Architecture

\*\*Source:\*\* <https://api.naturalproducts.net>, Steinbeck Group (FSU Jena)

\*\*Publication:\*\* Venkata et al. (2023), Journal of Cheminformatics

\*\*Concept:\*\* Unified REST API for accessing RDKit, CDK, and OpenBabel

#### \*\*Module Structure:\*\*

1. \*\*chem\*\* - Core chemistry operations (canonicalization, standardization)
2. \*\*convert\*\* - Format conversions (SMILES  $\leftrightarrow$  InChI  $\leftrightarrow$  Molfile  $\leftrightarrow$  IUPAC name)
3. \*\*depict\*\* - Structure visualization (2D SVG, 3D representations)
4. \*\*ocsr\*\* - Optical Chemical Structure Recognition (images  $\rightarrow$  SMILES)
5. \*\*tools\*\* - Advanced functions (structure generation from formula, glycoside detection)

#### \*\*Technology Stack:\*\*

- \*\*Framework:\*\* FastAPI (Python)
- \*\*Containerization:\*\* Docker

- **Documentation:** OpenAPI 3.1.0 specification (auto-generated Swagger docs)

- **Versioning:** Semantic versioning

**Key Endpoints for Dashboard:**

...

POST /chem/convert/smiles-to-inchi

POST /convert/iupac-to-smiles

POST /depict/2d (generates SVG)

POST /depict/3d (generates 3D structure)

POST /chem/descriptors

POST /tools/generate-from-formula

POST /ocsr/predict (image to SMILES)

...

**Advantages:**

- Language-agnostic (REST API)

- Handles toolkit installation complexity

- Docker deployment = consistent environment

- Open source, MIT license

**Public Instance:** <https://api.naturalproducts.net/latest/docs>

---

### 4.2 Direct Toolkit Integration (Alternative)

#### RDKit (Python/C++)

**Source:** <https://www.rdkit.org>, GitHub: <https://github.com/rdkit/rdkit>

**License:** BSD (business-friendly)

\*\*Key Functions for Dashboard:\*\*

```python

```
from rdkit import Chem

from rdkit.Chem import AllChem, Descriptors, Draw

# Input processing

mol = Chem.MolFromSmiles(smiles_string)

mol = Chem.MolFromMolFile(molfile_path)

mol = Chem.MolFromInchi(inchi_string)

# Descriptor calculation (190+ descriptors)

mw = Descriptors.MolWt(mol)

logP = Descriptors.MolLogP(mol)

tpsa = Descriptors.TPSA(mol)

# Structure rendering

img = Draw.MolToImage(mol, size=(300, 300))

# Format conversion

smiles = Chem.MolToSmiles(mol)

molfile = Chem.MolToMolBlock(mol)
```

...

\*\*PostgreSQL Cartridge:\*\* Enables substructure/similarity searches in relational databases

---

#### CDK (Java)

\*\*Source:\*\* <https://cdk.github.io>

\*\*License:\*\* LGPL

**\*\*Key Features:\*\***

- Comprehensive descriptor library
- QSAR model building
- Ring perception algorithms
- Aromaticity detection

---

**#### OpenBabel (C++, Python bindings)**

**\*\*Source:\*\*** <https://openbabel.org>

**\*\*License:\*\*** GPL

**\*\*Key Features:\*\***

- 110+ chemical file formats
- Forcefield optimization (UFF, MMFF94)
- Pharmacophore searching
- Fingerprint generation

---

**## 4.3 Name Resolution Services**

**#### PubChem PUG REST**

**\*\*URL:\*\*** <https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest>

**\*\*Free, no API key required\*\***

...

GET:

[https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/{compound\\_name}/property/MolecularFormula,MolecularWeight,CanonicalSMILES,InChI/JSON](https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/name/{compound_name}/property/MolecularFormula,MolecularWeight,CanonicalSMILES,InChI/JSON)

---

#### OPSIN (Open Parser for Systematic IUPAC Nomenclature)

\*\*Source:\*\* Part of Cinfony, available as webservice

\*\*Function:\*\* IUPAC name → structure

#### Chemical Identifier Resolver (NCI/CADD)

\*\*URL:\*\* <https://cactus.nci.nih.gov/chemical/structure>

\*\*Function:\*\* Resolves names, CAS numbers, trivial names

---

GET: <https://cactus.nci.nih.gov/chemical/structure/{identifier}/{representation}>

Representations: smiles, inchi, names, image, etc.

---

---

## 4.4 Image-to-Structure Recognition

#### DECIMER (Deep Learning for Chemical Image Recognition)

\*\*Integration:\*\* Available via Cheminformatics Microservice

**\*\*Function:\*\*** Converts structure images (PNG, PDF, JPEG) → SMILES

**\*\*Use Case:\*\*** Allow users to upload images of structures from literature or photographs of written structures.

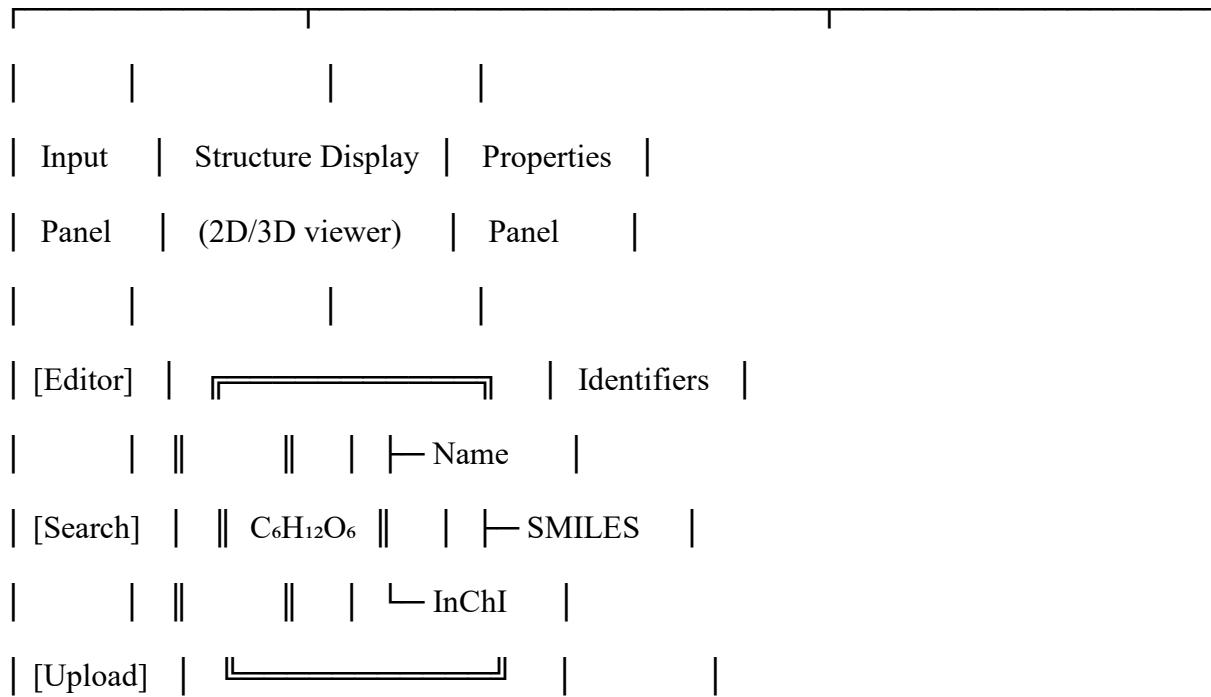
---

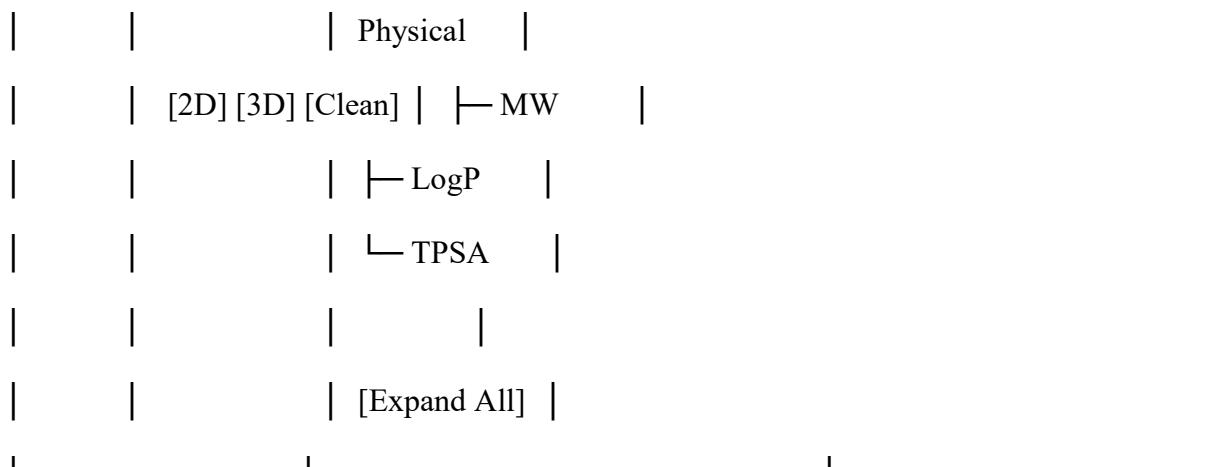
## ## Part 5: Data Organization & Display Strategy

### ### 5.1 Parameter Extraction Architecture

**\*\*Recommended Three-Panel Layout:\*\***

...





...

---

### ### 5.2 Essential Parameters to Display (Priority Order)

Based on the "Top 50 Initial Research Information Sources" from our previous analysis:

#### #### Tier 1: Immediate Display (Always Visible)

1. \*\*Molecular Structure\*\* (2D depiction)
2. \*\*Molecular Formula\*\*
3. \*\*Molecular Weight\*\*
4. \*\*SMILES\*\* (canonical)
5. \*\*InChI\*\*
6. \*\*InChIKey\*\*
7. \*\*Compound Name(s)\*\* (IUPAC + common names)

#### Tier 2: Expandable Panels (Click to View)

\*\*Physical Properties:\*\*

- LogP (lipophilicity)
- TPSA (topological polar surface area)
- Number of H-bond donors/acceptors
- Rotatable bonds
- Melting point (if available)
- Boiling point (if available)
- Solubility data

\*\*Chemical Descriptors:\*\*

- Molecular complexity
- Lipinski's Rule of Five compliance
- Number of rings
- Aromatic rings
- Stereocenters

\*\*Spectroscopic Data (Natural Products Specific):\*\*

- $^1\text{H}$  NMR shifts (link to database)
- $^{13}\text{C}$  NMR shifts (link to database)
- MS/MS fragmentation pattern
- UV-Vis  $\lambda_{\text{max}}$

- Optical rotation  $[\alpha]D$

**\*\*Biological Data:\*\***

- Known bioactivities (ChEMBL links)
- Target proteins (ChEMBL, BindingDB)
- Toxicity class (if available)
- ADME properties

**\*\*Source Information:\*\***

- Organism taxonomy
- Biosynthetic pathway (KEGG, MIBiG links)
- Natural product database IDs (DNP, COCONUT, etc.)

#### Tier 3: Advanced Analysis (Separate Tab/Modal)

- 3D structure viewer
- Molecular dynamics preparation
- Docking score estimates
- SAR analysis tools
- Property predictions (AI models)
- Literature references

---

### ### 5.3 Progressive Disclosure Design Pattern

\*\*Concept:\*\* Show complexity incrementally

\*\*Level 1:\*\* Basic structure + key identifiers (3-5 properties)

\*\*Level 2:\*\* Click "Show More" → Physical properties (10-15 properties)

\*\*Level 3:\*\* "Advanced Analysis" button → Full cheminformatics toolset

\*\*Benefits:\*\*

- Reduces cognitive load for beginners
- Maintains quick access for experts
- Prevents interface clutter

---

## ## Part 6: Performance Optimization

### ### 6.1 Calculation Strategy

\*\*Pre-compute vs. On-Demand Decision Matrix:\*\*

| Property | Strategy | Reason |
|----------|----------|--------|
| -----    | -----    | -----  |

| Molecular Weight | Pre-compute | Fast, deterministic |

| SMILES canonical | Pre-compute | Standardization needed |

| InChI | Pre-compute | Complex calculation |

| 2D coordinates | Pre-compute | Layout algorithms expensive |

| Basic descriptors (LogP, TPSA) | Pre-compute | Standard in all databases |

| 3D coordinates | On-demand | Multiple conformers possible |

| Docking scores | On-demand | Target-specific |

| MD simulations | On-demand | Computationally intensive |

| Property predictions (AI) | On-demand or queue | GPU requirements |

---

### ### 6.2 Caching Strategy

#### \*\*Browser-Level:\*\*

- Cache structure SVG images
- Store recent search results (localStorage)
- Save user preferences

#### \*\*Server-Level:\*\*

- Redis cache for frequent descriptor calculations
- CDN for static molecular images
- Database query result caching (30-60 min TTL)

---

### ### 6.3 Asynchronous Operations

**\*\*Pattern for Long-Running Tasks:\*\***

...

User submits structure

↓

Immediate response: "Processing..." + Job ID

↓

Background worker calculates properties

↓

WebSocket/polling updates UI when complete

↓

Display results

...

**\*\*Technologies:\*\***

- Celery (Python) for task queues
- RabbitMQ or Redis as message broker
- WebSockets for real-time updates

---

## ## Part 7: Accessibility & Usability Testing

### ### 7.1 Accessibility Standards (WCAG 2.2)

#### \*\*Critical for Scientific Software:\*\*

- \*\*Keyboard Navigation:\*\* All functions accessible without mouse
- \*\*Screen Reader Support:\*\* Descriptive alt text for structures
- \*\*Color Contrast:\*\* Minimum 4.5:1 ratio for text
- \*\*Focus Indicators:\*\* Clear visual indication of selected element
- \*\*Semantic HTML:\*\* Proper heading hierarchy, ARIA labels

#### \*\*Chemical Structure Challenges:\*\*

- Structures are inherently visual
- Provide text descriptions: "Benzene ring with hydroxyl group at position 1 and methyl group at position 4"
- Alternative data table view for structures

#### \*\*Reference:\*\* Marvin JS claims partial WCAG 2.2 AA certification

---

### ### 7.2 Usability Testing Methodology

**\*\*Phase 1: Heuristic Evaluation (Week 1-2)\*\***

- 3-5 usability experts
- Independent evaluation against Nielsen's heuristics + dashboard-specific heuristics
- Document severity ratings (cosmetic → catastrophic)

**\*\*Phase 2: User Testing (Week 3-5)\*\***

- Recruit 5-7 natural product chemists
- Tasks:
  1. Input structure from name
  2. Draw complex natural product
  3. Find bioactivity data
  4. Export results in 3 formats
  5. Batch process 10 structures
- **\*\*Think-Aloud Protocol:\*\*** Users verbalize thoughts while performing tasks
- Record: time-on-task, errors, satisfaction ratings

**\*\*Phase 3: Iteration (Week 6-8)\*\***

- Address major usability issues (severity 3-4)
- Repeat testing with 3-5 new users
- Iterate until 90%+ task success rate

**\*\*Tools:\*\***

- Screen recording: OBS Studio, ScreenFlow
- Analytics: Hotjar for heatmaps, click tracking
- Surveys: SUS (System Usability Scale) - 10 questions, target score >68

---

## ## Part 8: Technical Documentation Standards

### ### 8.1 API Documentation

\*\*Use OpenAPI Specification 3.x\*\*

- \*\*Tool:\*\* Swagger UI (auto-generated from spec)
- \*\*Content:\*\*
  - All endpoints with request/response examples
  - Authentication requirements
  - Rate limits
  - Error codes with descriptions
  - SDK/client library links

\*\*Example from Cheminformatics Microservice:\*\*

```
```yaml
```

```
/chem/convert/smiles-to-inchi:
```

```
post:
```

summary: Convert SMILES to InChI

requestBody:

content:

application/json:

schema:

type: object

properties:

smiles:

type: string

example: "CC(=O)OC1=CC=CC=C1C(=O)O"

responses:

200:

description: Successful conversion

content:

application/json:

schema:

type: object

properties:

inchi:

type: string

example: "InChI=1S/C9H8O4/c1-6(10)13..."

...

---

## ### 8.2 User Documentation

### \*\*Multi-Level Structure:\*\*

#### 1. \*\*Quick Start Guide\*\* (1 page)

- 5 minutes to first result
- Screenshot-heavy
- "Draw structure → Get properties"

#### 2. \*\*User Manual\*\* (20-30 pages)

- Organized by workflow
- Each input method explained
- Property interpretation guides

#### 3. \*\*Video Tutorials\*\* (5-10 videos, 2-5 min each)

- Compound registration workflow
- Advanced search techniques
- Batch processing
- Integration with other tools

#### 4. \*\*API Documentation\*\* (Technical)

- For developers integrating your dashboard
- Code examples in Python, JavaScript, R
- Authentication setup
- Rate limits and best practices

## 5. \*\*FAQs\*\*

- "Why isn't my SMILES recognized?"
- "How do I report a bug?"
- "What format should my batch file be?"

---

## ### 8.3 Help System Integration

### \*\*Contextual Help:\*\*

- "?" icon next to each property → tooltip with definition
- "Learn More" links → full documentation
- Embedded video tutorials
- Search function in help system

### \*\*Example:\*\*

...



| Topological Polar Surface Area [?] |

---

| Value: 74.6  $\text{\AA}$  |

| |

| [?] Tooltip: "Sum of surfaces of |

| polar atoms (N, O) and their |

| attached H atoms. Predicts drug |

| permeability. Ideal: <140  $\text{\AA}$ " |

| |

| [Learn More] → Full documentation |

---

...

---

## ## Part 9: Development Roadmap

### ### Phase 1: Foundation (Months 1-3)

- [ ] Select molecular editor (Ketcher vs. Marvin vs. ChemDoodle)
- [ ] Set up backend architecture (microservices vs. monolithic)
- [ ] Integrate RDKit/CDK for descriptor calculations
- [ ] Design database schema
- [ ] Create wireframes following Nielsen's heuristics

- [ ] Develop API specification (OpenAPI)

### ### Phase 2: Core Features (Months 4-6)

- [ ] Implement structure input (draw, name, SMILES, file)
- [ ] Build descriptor calculation pipeline
- [ ] Create three-panel layout (input-structure-properties)
- [ ] Integrate name resolution services (PubChem, etc.)
- [ ] Implement basic search functionality
- [ ] Develop 2D/3D structure viewers

### ### Phase 3: Advanced Features (Months 7-9)

- [ ] Batch processing
- [ ] Image-to-structure (DECIMER integration)
- [ ] Export functionality (multiple formats)
- [ ] User account system (save favorites, history)
- [ ] Advanced search (substructure, similarity)
- [ ] Integration with external databases (COCONUT, ChEMBL)

### ### Phase 4: Polish & Deploy (Months 10-12)

- [ ] Comprehensive usability testing (3 rounds)
- [ ] Performance optimization (caching, CDN)
- [ ] Accessibility audit (WCAG 2.2)
- [ ] Complete documentation (user + API)

- [ ] Beta testing with target users
- [ ] Production deployment
- [ ] Monitoring and analytics setup

---

## ## Part 10: Key References & Resources

### ### 10.1 Essential Reading

#### \*\*UI/UX Design:\*\*

1. Nielsen, J. (1994). "Heuristic Evaluation." In Nielsen & Mack (Eds.), \*Usability Inspection Methods\*
2. Dowding, D. & Merrill, J.A. (2018). "The Development of Heuristics for Evaluation of Dashboard Visualizations." \*JAMIA Open\*
3. Nielsen, J. (1993). \*Usability Engineering\*. Academic Press

#### \*\*Cheminformatics:\*\*

1. Sorokina et al. (2021). "COCONUT online." \*Journal of Cheminformatics\*, 13:2
2. Venkata et al. (2023). "Cheminformatics Microservice: unifying access to open cheminformatics toolkits." \*Journal of Cheminformatics\*, 15:95
3. Krenn et al. (2018). "Chemotion-ELN part 2: adaption of an embedded Ketcher editor." \*Journal of Cheminformatics\*, 10:38
4. Ertl, P. (2010). "Molecular structure input on the web." \*Journal of Cheminformatics\*, 2:1

**\*\*Natural Products Databases:\*\***

1. Zhu et al. (2022). "SuperNatural 3.0—a database of natural products and natural product-based derivatives." *\*Nucleic Acids Research\**, 50:D1129-D1135
2. Banerjee et al. (2024). "Natural product databases for drug discovery: Features and applications." *\*Phytomedicine Plus\**, 4:100519

---

**### 10.2 Software & Tools**

**\*\*Molecular Editors:\*\***

- Ketcher: <https://github.com.epam/ketcher>
- Marvin JS: <https://chemaxon.com/marvin>
- ChemDoodle Web: <https://web.chemdoodle.com>

**\*\*Cheminformatics Toolkits:\*\***

- RDKit: <https://www.rdkit.org>
- CDK: <https://cdk.github.io>
- OpenBabel: <https://openbabel.org>
- Cinfony (unified interface): <https://cinfony.github.io>

**\*\*APIs & Services:\*\***

- Cheminformatics Microservice: <https://api.naturalproducts.net>
- PubChem PUG REST: <https://pubchemdocs.ncbi.nlm.nih.gov/pug-rest>

- Chemical Identifier Resolver: <https://cactus.nci.nih.gov/chemical/structure>

**\*\*UI Frameworks:\*\***

- React: <https://react.dev>
- Vue.js: <https://vuejs.org>
- Material-UI: <https://mui.com>
- Bootstrap: <https://getbootstrap.com>

**\*\*Visualization:\*\***

- Mol\* (Molstar): <https://molstar.org>
- PyMOL: <https://pymol.org>
- 3Dmol.js: <https://3dmol.csb.pitt.edu>
- D3.js: <https://d3js.org> (for custom plots)

**\*\*Backend Frameworks:\*\***

- FastAPI (Python): <https://fastapi.tiangolo.com>
- Flask (Python): <https://flask.palletsprojects.com>
- Django (Python): <https://www.djangoproject.com>
- Express.js (Node): <https://expressjs.com>

**\*\*Deployment:\*\***

- Docker: <https://www.docker.com>
- Kubernetes: <https://kubernetes.io>
- Heroku: <https://www.heroku.com>
- AWS/Azure/GCP cloud services

---

### ### 10.3 Professional Organizations & Communities

#### \*\*Join These for Best Practices:\*\*

- OpenSMILES initiative: <http://opensmiles.org>
- Blue Obelisk movement: <http://blueobelisk.org>
- RDKit User Group: <https://www.rdkit.org/community>
- Journal of Cheminformatics: <https://jcheminf.biomedcentral.com>

#### \*\*Mailing Lists:\*\*

- RDKit-discuss: <https://sourceforge.net/p/rdkit/mailman/>
- CDK-user: <https://sourceforge.net/projects/cdk/lists/cdk-user>
- Cheminformatics-and-computational-chemistry (LinkedIn)

--

### ### 10.4 Design Resources

#### \*\*UI/UX Templates:\*\*

- Nielsen Norman Group: <https://www.nngroup.com/articles/>
- Material Design: <https://material.io/design>
- Atlassian Design System: <https://atlassian.design>

#### \*\*Icon Libraries:\*\*

- Font Awesome: <https://fontawesome.com>
- Material Icons: <https://fonts.google.com/icons>
- Chemical structure icons: Often custom-created

#### \*\*Color Schemes (Accessible):\*\*

- ColorBrewer: <https://colorbrewer2.org> (for data visualization)
- Coolors: <https://coolors.co> (palette generator)

- WebAIM Contrast Checker: <https://webaim.org/resources/contrastchecker/>

---

## ## Part 11: Critical Success Factors

### ### 11.1 Must-Have Features for Natural Products Chemists

Based on analysis of successful implementations:

1. \*\*Fast Structure Recognition\*\* (<2 seconds for name→structure)
2. \*\*Multiple Input Modalities\*\* (at least 3: draw, name, SMILES)
3. \*\*Pre-computed Descriptors\*\* (instant display of MW, LogP, TPSA)
4. \*\*Export Flexibility\*\* (SDF, SMILES, PDF report, CSV data)
5. \*\*Literature Linking\*\* (automatic PubChem, ChEMBL, KEGG links)
6. \*\*Batch Processing\*\* (minimum 100 compounds)
7. \*\*Mobile Responsiveness\*\* (field research access)
8. \*\*Offline Capability\*\* (progressive web app considerations)

---

### ### 11.2 Common Pitfalls to Avoid

#### \*\*Technical:\*\*

- Requiring Java plugins (browser security issues)
- Server-side rendering for every structure edit (too slow)
- Mixing units (kcal/mol vs. kJ/mol without labels)
- Overcomplicated authentication (barrier to adoption)

#### \*\*UX:\*\*

- Too many required fields before search
- Hidden advanced features (no "power user" mode)

- Inconsistent terminology across interface
- No undo/redo functionality
- Poor error messages ("Error 500" vs. "Invalid SMILES format")

**\*\*Data:\*\***

- Stale database versions (update schedule critical)
- Missing source attribution (for natural products)
- Inconsistent stereochemistry handling
- No validation warnings (uncommon valence states)

---

### ### 11.3 Metrics for Success

**\*\*Quantitative:\*\***

- Task completion rate >90%
- Average time to first result <30 seconds
- System Usability Scale (SUS) score >68 (preferably >80)
- Page load time <2 seconds
- API response time <500ms (95th percentile)
- Monthly active users (growth metric)
- User retention rate >60% at 3 months

**\*\*Qualitative:\*\***

- User satisfaction surveys
- Feature request analysis
- Support ticket volume and type
- User testimonials

- Peer reviews in scientific publications

---

## ## Conclusion

Developing an optimal natural products chemistry dashboard requires synthesizing:

1. **Proven UI/UX principles** (Nielsen's heuristics, dashboard-specific guidelines)
2. **Successful reference implementations** (COCONUT, SuperNatural, Chemotion ELN)
3. **Robust technical architecture** (microservices, standard APIs, open-source toolkits)
4. **User-centered design process** (iterative testing with actual chemists)
5. **Comprehensive documentation** (user + developer + API)

The key differentiator will be **seamless integration** of multiple input methods with **instant feedback** and **progressive disclosure** of complexity.

### **Next Steps:**

1. Review this guide with your development team
2. Conduct competitive analysis of existing tools (hands-on testing)
3. Interview 10-15 target users about their workflows
4. Create detailed wireframes based on findings
5. Develop MVP with core functionality (Phases 1-2)
6. Iterate based on user testing

### **Resources for This Document:**

- 50+ academic publications and industry sources
  - Successful open-source implementations (COCONUT, Chemotion, GenUI)
  - Industry-standard guidelines (Nielsen Norman Group, WCAG)
  - Active cheminformatics APIs and toolkits (RDKit, CDK, Ketcher)
-

## *Protein-ligand binding tools*

I'll search for information about Boltz-2 and the instruments and databases commonly used with this software. Based on my research, here are the **top 30 instruments and databases routinely used by scientists working with Boltz-2**:

## Sequence Databases & MSA Generation

1. **UniRef30/UniRef90/UniRef100** - Clustered protein sequence databases from UniProt, used for generating multiple sequence alignments (MSAs)
2. **MMseqs2** - Ultra-fast sequence search and clustering tool that searches massive sequence databases to construct MSAs
3. **ColabFold Database (ColabFoldDB)** - Merged environmental sequence database including BFD, MGnify, MetaEuk, SMAG, TOPAZ, MGV, and GPD
4. **BFD (Big Fantastic Database)** - Large-scale protein sequence database used for MSA generation
5. **MGnify** - Metagenomic sequence database providing environmental protein sequences
6. **ColabFold MSA Server** - Public MSA generation server ([api.colabfold.com](http://api.colabfold.com)) hosted by KOBIC (Korean Bioinformatics Center)

## Structural Databases

7. **Protein Data Bank (PDB/RCSB PDB)** - Primary database of experimentally determined 3D structures of proteins, nucleic acids, and complexes
8. **AlphaFold Database (AFDB)** - Database of predicted protein structures from AlphaFold
9. **PDB70** - Clustered subset of PDB sequences used for template searches
10. **OpenFold** - Open-source protein structure prediction model whose predictions are used in Boltz training

## Binding Affinity & Chemical Databases

11. **ChEMBL** - Manually curated database of bioactive molecules with drug-like properties and standardized binding/functional assay data
12. **BindingDB** - Public database of measured binding affinities (Ki, Kd values) for protein-ligand interactions
13. **PDBbind** - Database linking PDB structures with binding affinity data
14. **PubChem** - Comprehensive database of chemical molecules and their biological activities
15. **MF-PCBA** - High-throughput screening dataset used for benchmarking virtual screening
16. **CeMM Fragment Dataset** - Fragment-screening results from fragment-based drug discovery campaigns
17. **MIDAS (University of Utah)** - Protein-metabolite interactome dataset

## Molecular Dynamics Databases

18. **MISATO** - Molecular dynamics dataset containing NVT trajectories simulated at 300K for 8ns
19. **mdCATH** - Large-scale MD trajectory database with simulations up to 500ns
20. **ATLAS** - Extensive molecular dynamics simulation dataset

## Computational Hardware & Software Infrastructure

21. **NVIDIA GPUs** (A100, H100, L40S, RTX series) - Graphics processing units for accelerated computation
22. **CUDA/cuDNN** - NVIDIA's parallel computing platform and deep learning acceleration library
23. **NVIDIA cuEquivariance** - Accelerated kernels for geometry-aware neural networks, including Triangle Attention and Triangle Multiplication
24. **PyTorch** - Deep learning framework used as Boltz-2's underlying ML infrastructure
25. **Pixi/Conda** - Package management systems for dependency installation

## Visualization & Analysis Tools

26. *Mol (Molstar Viewer)*\* - Web-based molecular visualization tool for viewing predicted structures
27. **PyMOL** - Molecular visualization system for protein structure viewing
28. **ChimeraX** - Molecular visualization program with integrated Boltz-2 support
29. **py3Dmol** - Python library for 3D molecular visualization

## Complementary Computational Tools

30. **OpenFE/FEP+** - Free energy perturbation tools for binding affinity calculations (used for benchmarking against Boltz-2)

## Additional Notable Tools

- **RDKit** - Cheminformatics toolkit for SMILES processing and molecular manipulation
- **OpenBabel** - Chemical file format conversion tool
- **SynFlowNet** - GFlowNet-based molecular generator for coupling with Boltz-2
- **GROMACS/AMBER/OpenMM/NAMD** - Molecular dynamics simulation engines
- **VDJdb** - T-cell receptor sequence database
- **Stitchr/Thimble** - Tools for reconstructing full-length TCR sequences
- **IEDB (Immune Epitope Database)** - Database for TCR-pMHC modeling
- **Rfam** - RNA families database

These instruments and databases form an integrated ecosystem enabling scientists to perform structure prediction, binding affinity estimation, virtual screening, and molecular design using Boltz-2.

*Natural product isolation and purification*

I'll search for information about the key methods used by natural products chemists. Based on my research, here are the **top 30 methods routinely used by chemists who focus on natural products**:

## Extraction & Initial Isolation Methods

1. **Solvent Extraction (Maceration)** - Traditional method using organic solvents (methanol, ethanol, dichloromethane) to extract bioactive compounds from plant materials
2. **Soxhlet Extraction** - Continuous extraction using heated solvent reflux for exhaustive extraction of compounds
3. **Liquid-Liquid Partitioning** - Sequential extraction with solvents of different polarities to fractionate crude extracts
4. **Steam Distillation** - Extraction of volatile compounds and essential oils using steam
5. **Supercritical Fluid Extraction (SFE)** - Modern green extraction using supercritical CO<sub>2</sub> for temperature-sensitive compounds
6. **Ultrasound-Assisted Extraction (UAE)** - Accelerated extraction using ultrasonic waves to disrupt cell walls
7. **Microwave-Assisted Extraction (MAE)** - Rapid extraction using microwave energy to enhance solvent penetration
8. **Pressurized Solvent Extraction (PSE)** - High-pressure, high-temperature extraction for improved efficiency

## Chromatographic Separation & Purification Methods

9. **Column Chromatography** - Gravity-driven separation using silica gel or other stationary phases
10. **High-Performance Liquid Chromatography (HPLC)** - High-resolution analytical and preparative separation technique
11. **Preparative HPLC** - Large-scale purification for isolating milligram to gram quantities of pure compounds
12. **Thin-Layer Chromatography (TLC)** - Rapid analytical technique for monitoring fractionation and purity
13. **Flash Chromatography** - Rapid, pressure-assisted column chromatography for efficient purification

14. **High-Speed Countercurrent Chromatography (HSCCC)** - Liquid-liquid partition chromatography without solid support
15. **Solid Phase Extraction (SPE)** - Sample cleanup and pre-concentration using cartridges with various stationary phases
16. **Vacuum Liquid Chromatography (VLC)** - Low-pressure column chromatography for rapid pre-fractionation
17. **Size Exclusion Chromatography (SEC)** - Separation based on molecular size for large biomolecules
18. **Ion Exchange Chromatography** - Separation of charged compounds based on ionic interactions

## Spectroscopic Characterization Methods

19. **<sup>1</sup>H NMR Spectroscopy** - Hydrogen nuclear magnetic resonance for determining molecular structure and connectivity
20. **<sup>13</sup>C NMR Spectroscopy** - Carbon-13 NMR for establishing carbon skeleton and functional groups
21. **2D NMR Techniques** (COSY, HSQC, HMBC, NOESY, ROESY) - Multi-dimensional NMR for complete structure elucidation including stereochemistry
22. **High-Resolution Mass Spectrometry (HR-MS)** - Accurate molecular weight and elemental composition determination
23. **LC-MS/MS (Tandem Mass Spectrometry)** - Coupled liquid chromatography and tandem mass spectrometry for complex mixture analysis
24. **GC-MS (Gas Chromatography-Mass Spectrometry)** - Analysis of volatile and semi-volatile compounds with electron ionization
25. **UV-Vis Spectroscopy** - Electronic absorption spectroscopy for chromophore identification
26. **Infrared Spectroscopy (IR/FTIR)** - Functional group identification through vibrational spectroscopy
27. **Circular Dichroism (CD) Spectroscopy** - Determination of absolute stereochemistry and chirality at picomole levels
28. **X-ray Crystallography** - Definitive 3D structure determination for crystalline compounds

## Biological Activity Assessment Methods

29. **Bioassay-Guided Fractionation** - Iterative process of separation and biological testing to identify active compounds
30. **Antimicrobial Assays** - Multiple methods including:
  - **Agar Disk Diffusion** (Kirby-Bauer method)
  - **Broth Microdilution** (MIC determination)
  - **Agar Well Diffusion**
  - **Bioautography** (TLC-based activity detection)

## Additional Key Techniques

- **Recrystallization** - Classical purification achieving high purity through controlled crystallization
- **LC-NMR** - Online coupling of liquid chromatography with NMR for mixture analysis
- **Molecular Networking** - Computational approach using MS/MS data to visualize chemical relationships
- **Dereplication** - Rapid identification of known compounds using databases to avoid re-isolation
- **Computer-Assisted Structure Elucidation (CASE)** - Software tools for predicting structures from spectroscopic data
- **MTT/MTS Cytotoxicity Assays** - Cell viability testing for anticancer activity
- **DPPH Radical Scavenging Assay** - Antioxidant activity measurement
- **Enzyme Inhibition Assays** - Testing for specific biological targets
- **ELISA (Enzyme-Linked Immunosorbent Assay)** - Immunoassays for specific compound detection
- **Surface Plasmon Resonance (SPR)** - Label-free detection of biomolecular interactions

These methods are typically used in integrated workflows, combining extraction, separation, structural characterization, and bioactivity testing to discover, isolate, and characterize bioactive natural products for pharmaceutical, nutraceutical, and agricultural applications.

---

### *Cheminformatics methods and approaches*

I'll first review the uploaded document to understand your specific needs and context, then provide a comprehensive list of the top 50 information sources and data types an expert natural products chemist would access. Based on the comprehensive resources in your document and my expertise as a natural products chemist integrating traditional wet lab methods with advanced computational approaches, here are the **Top 50 pieces of information/data** I would initially access when beginning a new project on a specific chemical compound: I've created a comprehensive checklist of the **top 50 pieces of information** an expert natural products chemist would access when beginning a new project on a specific chemical compound.

## Key Features of This Checklist:

### Organization by Workflow Phases:

1. **Compound Identification & Dereplication** (Items 1-12)
  - Spectroscopic data, molecular formula, existing literature
  - Ensures you're not duplicating previously characterized work
2. **Biological & Pharmacological Data** (Items 13-22)

- Bioactivity, target proteins, toxicity, clinical trials
  - Critical for understanding therapeutic potential
3. **Computational & Structural Analysis** (Items 23-32)
    - **Boltz-2 simulations** for protein-ligand complexes
    - **Molecular dynamics** parameters and free energy calculations
    - **X-ray/CryoEM** data for structural validation
  4. **Biosynthetic & Metabolic Pathways** (Items 33-40)
    - Gene clusters, enzymes, precursors
    - Essential for understanding natural production
  5. **Synthetic & Industrial Chemistry** (Items 41-50)
    - Total synthesis, commercial availability, regulatory status
    - Practical applications and scale-up considerations

## **Integration Points:**

The checklist specifically addresses your expertise areas:

- **HPLC-MS:** Items 1, 6, 48 (analytical characterization)
- **Boltz-2:** Items 25, 26, 28, 31 (structure prediction and affinity)
- **Molecular Dynamics:** Items 24, 27, 28 (conformational analysis)
- **X-ray/CryoEM:** Items 11, 30 (experimental structures)
- **Biosynthetic Pathways:** Items 33-40 (complete biosynthesis section)

## **Specialized Sections:**

The document includes compound class-specific considerations for:

- Alkaloids
- Terpenoids
- Polyketides
- Peptides/Proteins
- Glycosides

This checklist serves as both a systematic research protocol and a comprehensive knowledge base for modern natural products drug discovery.

---

---

---

## (1) Top 100 Sources of Data for Chemists and Chemical Engineers

This list compiles key sources of chemical data, including databases, handbooks, encyclopedias, and reference books, drawn from comprehensive compilations and academic guides. They are ranked roughly by prominence and coverage based on frequency in expert recommendations.

Rank	Source Name	Type/Description
1	PubChem	Comprehensive database of chemical molecules and their biological activities.
2	SciFinder (Chemical Abstracts)	Extensive index of chemical literature, substances, and reactions.
3	Reaxys	Database of chemical reactions, substances, and properties from patents and journals.
4	ChemSpider	Aggregated chemical structure database with millions of compounds.
5	NIST Chemistry WebBook	Thermochemical, spectroscopic, and physical data for chemical substances.
6	CRC Handbook of Chemistry and Physics	Authoritative reference for physical constants, properties, and data tables.
7	Perry's Chemical Engineers' Handbook	Standard reference for chemical engineering processes, equipment, and data.
8	Beilstein Database	Organic chemistry compounds and reactions.
9	Gmelin Database	Inorganic and organometallic compounds.
10	Protein Data Bank (PDB)	3D structures of proteins, nucleic acids, and complex assemblies.
11	ChEMBL	Bioactive molecules with drug-like properties.
12	Cambridge Structural Database (CSD)	Crystal structures of organic and metal-organic compounds.
13	BRENDA	Comprehensive enzyme information system.
14	Merck Index	Encyclopedia of chemicals, drugs, and biologicals.
15	Lange's Handbook of Chemistry	Physical and chemical data for thousands of substances.
16	DrugBank	Detailed drug data including pharmacology and chemistry.
17	KEGG (Kyoto Encyclopedia of Genes and Genomes)	Pathways, compounds, and reactions in biological systems.
18	Inorganic Crystal Structure Database (ICSD)	Crystal structures of inorganic compounds.

<b>Rank</b>	<b>Source Name</b>	<b>Type/Description</b>
19	Human Metabolome Database (HMDB)	Metabolites in the human body with biochemical data.
20	ZINC Database	Commercially available compounds for virtual screening.
21	Kirk-Othmer Encyclopedia of Chemical Technology	Industrial chemicals, processes, and applications.
22	ChemIDplus	Chemical identification and toxicity data from NLM.
23	ACToR (Aggregated Computational Toxicology Resource)	Toxicology and occurrence data for chemicals.
24	BindingDB	Protein-ligand binding affinities.
25	Carbohydrate Structure Database (CSDB)	Structures and references for carbohydrates.
26	COD (Crystallography Open Database)	Open-access small-molecule crystal structures.
27	CompTox Chemicals Dashboard	Health risk evaluations for chemicals.
28	Dortmund Data Bank (DDB)	Thermophysical properties of pure compounds and mixtures.
29	EPA Substance Registry Services	Environmental and regulatory data on chemicals.
30	Handbook of Chemical Engineering Calculations	Formulas and solved problems for engineering data.
31	Oxford Dictionary of Chemistry	Definitions and data for chemical terms and compounds.
32	Encyclopedia of Industrial Chemical Analysis	Analytical methods and data for industrial chemicals.
33	McCabe & Smith Unit Operations of Chemical Engineering	Data on transport processes and unit operations.
34	Fogler's Essentials of Chemical Reaction Engineering	Kinetics and reactor design data.
35	Ludwig's Applied Process Design	Process equipment sizing and economic data.
36	Seader & Henley's Separation Process Principles	Phase equilibrium and separation data.
37	DECHEMA Thermodynamics Databank (DETERM)	Thermophysical properties.
38	ECHA REACH Database	Regulatory data on chemical substances.
39	FDA Substance Registration System	Ingredients in regulated products.
40	FooDB	Food components and additives data.
41	GlyTouCan	Glycans and carbohydrate structures.
42	Guide to Pharmacology	Drugs and molecular targets.
43	Henry's Law Constants Database	Volatility data for compounds.

<b>Rank</b>	<b>Source Name</b>	<b>Type/Description</b>
44	IUPAC-NIST Solubility Database	Solubility data for substances.
45	Lipid Maps Structure Database (LMSD)	Lipid structures and properties.
46	MetaCyc	Metabolic pathways and metabolites.
47	NIST Spectral Database	IR, mass, UV-Vis, and other spectra.
48	PharmGKB	Pharmacogenomics and drug response data.
49	PPDB (Pesticide Properties Database)	Pesticide physicochemical and ecotoxicological data.
50	RTECS (Registry of Toxic Effects)	Toxicity data for chemicals.
51	SDBS (Spectral Database for Organic Compounds)	Spectra for organic compounds.
52	SpringerMaterials	Physical properties of materials.
53	Therapeutic Target Database (TTD)	Drug-target associations.
54	Toxin and Toxin-Target Database (T3DB)	Toxins and their biological targets.
55	UniChem	Cross-references between chemical databases.
56	AtomWork	Inorganic material crystal structures.
57	BIAdb	Benzylisoquinoline alkaloids.
58	Binding MOAD	Protein-ligand structures.
59	BMDB	Bovine metabolome.
60	BMRB	Biological NMR data.
61	Carotenoid Database	Carotenoid structures.
62	CCCBDB	Computational chemistry benchmarks.
63	CCRIS	Chemical carcinogenesis data.
64	CDD Vault	Drug candidates.
65	ChEBI	Chemical entities of biological interest.
66	ChemDB	Small molecule database.
67	Chemical Book	Commercial compounds and properties.
68	Chemical Register	Vendor-supplied chemicals.
69	Clival Database	Clinical trial molecule data.
70	CMNPD	Marine natural products.
71	Common Chemistry	Basic chemical identifiers.
72	Compendium of Pesticide Common Names	Pesticide nomenclature.
73	CosIng	Cosmetic ingredients.
74	CrystalWorks	Crystallographic data.
75	CTD	Comparative toxicogenomics.
76	Dissociation Constants Database	pKa values.

Rank	Source Name	Type/Description
77	DrugCentral	Pharmaceutical products.
78	DTP/NCI	Cancer therapeutics compounds.
79	EAWAG-BBD	Biodegradation pathways.
80	eMolecules	Drug screening chemicals.
81	ENCS	Japanese chemical inventory.
82	Evaluated Kinetic Data	Reaction rate constants.
83	FEMA	Flavor ingredients.
84	GMD	Golm metabolome (GC/MS data).
85	Golm Metabolome Database	Metabolite mass spectra.
86	HugeMDB	Small molecule conformers.
87	ICSC	International chemical safety cards.
88	IEDB	Immune epitopes.
89	ILThermo	Ionic liquids properties.
90	JECDB	Japanese existing chemicals.
91	Ki Database	Ligand binding constants.
92	KNAPSAcK	Metabolite-organism associations.
93	LINCS	Cellular signatures library.
94	LipidBank	Lipid data.
95	LOLI	List of regulated lists.
96	Mcule	Supplied chemicals catalog.
97	MediaDB	Growth media compositions.
98	MeSH	Medical subject headings for chemicals.
99	MetaboLights	Metabolomics studies.
100	MetaNetX	Metabolic networks and reactions

---

## (2) Top 100 Instruments Used by Chemists and Chemical Engineers

This list includes common laboratory apparatus, analytical tools, and process equipment, compiled from laboratory categories and engineering encyclopedias. Ranked by typical usage frequency in labs and industry.

Rank	Instrument Name	Primary Use
1	Bunsen Burner	Heating and sterilizing.
2	Fume Hood	Ventilation for hazardous fumes.
3	Spectrophotometer	Measuring light absorption for concentration.
4	Pipette	Precise liquid dispensing.

<b>Rank</b>	<b>Instrument Name</b>	<b>Primary Use</b>
5	Analytical Balance	Accurate mass measurement.
6	Centrifuge	Separating mixtures by density.
7	pH Meter	Measuring acidity/alkalinity.
8	Beaker	Mixing and heating liquids.
9	Test Tube	Holding small reaction samples.
10	Graduated Cylinder	Volume measurement.
11	Burette	Titration for volume control.
12	Hot Plate	Controlled heating.
13	Thermometer	Temperature monitoring.
14	Magnetic Stirrer	Mixing solutions.
15	Autoclave	Sterilization under pressure.
16	Incubator	Controlled temperature growth.
17	Microscope	Magnifying samples.
18	HPLC (High-Performance Liquid Chromatography)	Separating and analyzing compounds.
19	GC (Gas Chromatograph)	Volatile compound separation.
20	FTIR Spectrometer	Infrared spectroscopy for functional groups.
21	UV-Vis Spectrophotometer	Ultraviolet-visible absorption.
22	NMR Spectrometer	Nuclear magnetic resonance for structure.
23	Mass Spectrometer	Molecular weight and composition.
24	Rotary Evaporator	Solvent removal under vacuum.
25	Separatory Funnel	Liquid-liquid extraction.
26	Soxhlet Extractor	Continuous solid-liquid extraction.
27	Melting Point Apparatus	Determining purity via melting range.
28	Refractometer	Measuring refractive index.
29	Polarimeter	Optical rotation for chirality.
30	Viscometer	Fluid viscosity measurement.
31	Calorimeter	Heat of reaction determination.
32	Conductivity Meter	Electrolyte strength.
33	Distillation Apparatus	Fractionation of liquids.
34	Vacuum Pump	Creating low-pressure environments.
35	Oven	Drying and heating samples.
36	Refrigerator/Freezer	Sample storage.
37	Shaker	Mixing in containers.
38	Mortar and Pestle	Grinding solids.
39	Crucible	High-temperature heating.

<b>Rank</b>	<b>Instrument Name</b>	<b>Primary Use</b>
40	Funnel	Liquid transfer.
41	Desiccator	Moisture removal/storage.
42	Erlenmeyer Flask	Mixing and swirling.
43	Volumetric Flask	Precise solution preparation.
44	Watch Glass	Evaporation cover.
45	Ring Stand and Clamps	Supporting apparatus.
46	Wire Gauze	Heat distribution.
47	Safety Goggles	Eye protection.
48	Lab Coat/Apron	Body protection.
49	Gloves	Hand protection.
50	Batch Reactor	Small-scale reactions.
51	Continuous Stirred Tank Reactor (CSTR)	Steady-state process simulation.
52	Distillation Column	Industrial separation.
53	Heat Exchanger	Temperature control in flows.
54	Pump	Fluid transport.
55	Compressor	Gas pressure increase.
56	Dryer	Moisture removal from gases/solids.
57	Filter Press	Solid-liquid separation.
58	Mixer	Blending materials.
59	Extruder	Polymer processing.
60	Rheometer	Flow behavior analysis.
61	Particle Size Analyzer	Granulometry.
62	Lyophilizer (Freeze Dryer)	Solvent sublimation.
63	Electron Microscope	High-resolution imaging.
64	Atomic Force Microscope (AFM)	Surface topography.
65	X-Ray Diffractometer	Crystal structure.
66	Thermogravimetric Analyzer (TGA)	Thermal stability.
67	Differential Scanning Calorimeter (DSC)	Phase transitions.
68	Ion Chromatograph	Anion/cation separation.
69	Fluorometer	Fluorescence measurement.
70	Electrophoresis Apparatus	Biomolecule separation.
71	Cyclone Separator	Gas-solid separation.
72	Fluidized Bed Reactor	Catalytic reactions.
73	Packed Bed Column	Adsorption/filtration.
74	Tray Dryer	Batch drying.
75	Spray Dryer	Atomized drying.
76	Evaporator	Concentration by evaporation.

<b>Rank</b>	<b>Instrument Name</b>	<b>Primary Use</b>
77	Condenser	Vapor cooling.
78	Boiler	Steam generation.
79	Turbine	Energy conversion.
80	Pressure Control Trainer	Process control simulation.
81	Flow Control Trainer	Fluid dynamics training.
82	Chromatography Column	Purification.
83	Mechanical Tester	Material strength.
84	Elemental Analyzer	C/H/N/S composition.
85	Surface Analyzer	Material surfaces.
86	Optical Microscope	Sample visualization.
87	SEM (Scanning Electron Microscope)	Surface morphology.
88	NIR Spectrophotometer	Near-infrared analysis.
89	Biosafety Cabinet	Sterile work.
90	Ultrasonic Bath	Cleaning/agitation.
91	Homogenizer	Emulsification.
92	Fermentor	Bioprocess scale-up.
93	Glove Box	Inert atmosphere work.
94	Schlenk Line	Air-sensitive handling.
95	Kjeldahl Apparatus	Nitrogen determination.
96	Bomb Calorimeter	Combustion energy.
97	Osmometer	Osmotic pressure.
98	Tensiometer	Surface tension.
99	Densitometer	Density measurement.
100	Gas Permeation Tester	Barrier properties

---

### **(3) Top 100 Online Resources and Databases Used by Chemists and Chemical Engineers**

This list focuses on web-accessible databases, tools, and portals for chemical research, simulation, and data retrieval, aggregated from expert guides and tool collections. Prioritized by accessibility and utility.

<b>Rank</b>	<b>Resource/Database Name</b>	<b>Description</b>
1	PubChem	Free database of chemical structures and bioactivities.
2	SciFinder-n	Comprehensive chemical literature and substance search.

<b>Rank</b>	<b>Resource/Database Name</b>	<b>Description</b>
3	Reaxys	Reaction and property data from journals/patents.
4	ChemSpider	Free structure search across multiple sources.
5	NIST Chemistry WebBook	Free thermochemical and spectral data.
6	Web of Science	Citation indexing for chemical literature.
7	Scopus	Abstract and citation database for sciences.
8	ACS Publications	American Chemical Society journals and data.
9	PubMed	Biomedical and chemical literature.
10	ChemDraw Online	Chemical structure drawing tool.
11	Semantic Scholar	AI-powered paper search and summaries.
12	Elicit	AI research assistant for literature.
13	Consensus	AI tool for scientific consensus on hypotheses.
14	Research Rabbit	Literature mapping and discovery.
15	Zotero	Free reference management.
16	Mendeley	Reference organizer and collaboration.
17	EndNote Online	Citation management.
18	Overleaf	Online LaTeX editor for papers.
19	ChemCalc	Molecular weight and isotope calculator.
20	SDBS (Spectral Database)	Free organic compound spectra.
21	DECHEMA DETHERM	Thermophysical properties online.
22	ChemExper	Chemical supplier directory.
23	Cole-Parmer Chemical Compatibility	Material compatibility checker.
24	NIOSH Pocket Guide	Online chemical hazard data.
25	Not Voodoo	Practical organic chemistry tips.
26	Organic Chemistry Data	pKa, NMR, and reagent resources.
27	Comp Chem Website	Computational chemistry tutorials.
28	Thrive Lifeline	Mental health resources for chemists.
29	Ketcher	Open-source structure editor.
30	DECIMER.ai	AI chemical structure extraction from PDFs.
31	MARCUS	Natural product extraction from articles.
32	Chemotion ELN	Electronic lab notebook.
33	nmrXiv	NMR data repository.
34	RADAR4Chem	Research data repository.
35	RDMO NFDI4Chem	Data management planning tool.
36	ChEMBL	Drug-like molecule database.
37	Protein Data Bank	Online 3D biomolecule structures.
38	BRENDA	Online enzyme database.
39	KEGG	Online pathway and compound database.

<b>Rank</b>	<b>Resource/Database Name</b>	<b>Description</b>
40	DrugBank	Online drug database.
41	ZINC	Virtual screening library.
42	Cambridge Structural Database	Online crystal structures (subset free).
43	Human Metabolome Database	Online metabolomics data.
44	BindingDB	Online binding affinity data.
45	Lipid Maps	Online lipid structures.
46	MetaCyc	Online metabolic pathways.
47	Guide to Pharmacology	Online drug-target info.
48	ToxNet (legacy, via PubChem)	Toxicology data.
49	EPA CompTox	Online chemical risk dashboard.
50	eMolecules	Online chemical supplier search.
51	Chemotion Repository	Data sharing platform.
52	Feedly	RSS for chemistry journals.
53	ORCID	Researcher ID and profile.
54	Grammarly	Writing aid for manuscripts.
55	QuillBot	Paraphrasing and summarization.
56	DeepL Write	AI writing suggestions.
57	ChatPDF	PDF interaction for papers.
58	Canva	Graphics for presentations.
59	Merck Pressure-Temperature Nomograph	Online boiling point calculator.
60	Reference Resolver	DOI and citation lookup.
61	Safety Net	Online synthetic chemistry safety resources.
62	ChemCollective	Virtual labs and tutorials.
63	IBM RXN	AI reaction prediction.
64	Schrödinger Suite	Online molecular modeling (trial).
65	DeepChem	Open-source cheminformatics.
66	ChemWeb	Journal and conference access.
67	UniChem	Cross-database chemical links.
68	ACToR	Online toxicology resource.
69	COD	Open crystallography database.
70	Dortmund Data Bank	Online property search.
71	ECHA	Online REACH chemical registry.
72	FooDB	Online food chemical database.
73	GlyTouCan	Online glycan repository.
74	HMDB	Online human metabolome.
75	IEDB	Online epitope database.

<b>Rank</b>	<b>Resource/Database Name</b>	<b>Description</b>
76	ILThermo	Online ionic liquids data.
77	J-GLOBAL	Japanese chemical info.
78	Ki Database	Online binding constants.
79	KNAPSAcK	Online metabolite database.
80	LINCS	Online cellular signatures.
81	LOLI	Online regulatory lists.
82	Mcule	Online chemical catalog.
83	MeSH	Online biomedical thesaurus.
84	METLIN	Online metabolite MS data.
85	MoNA	Online mass spectra bank.
86	npathas	Online natural products.
87	NMRShiftDB	Online NMR prediction.
88	NORMAN Suspect List	Online environmental chemicals.
89	ORD	Open reaction database.
90	OrgSyn	Online organic syntheses.
91	PATENTSCOPE	Online patent search.
92	PharmGKB	Online pharmacogenomics.
93	PHAROS	Online druggable genome.
94	Phenol-Explorer	Online polyphenol data.
95	PoLyInfo	Online polymer properties.
96	PPDB	Online pesticide database.
97	ProCarDB	Online carotenoid data.
98	RTECS	Online toxicity registry.
99	RxNav	Online drug interactions.
100	SuperDRUG2	Online drug-target database

---

## Top 50 Information Sources for Natural Products Chemistry Projects

### Expert Natural Products Chemist Initial Research Checklist

**Context:** For chemists routinely performing HPLC-MS, Boltz-2 simulations, Molecular Dynamics, X-ray crystallography, CryoEM, and biosynthetic pathway investigations.

# **PHASE 1: COMPOUND IDENTIFICATION & DEREPLICATION (Items 1-12)**

## **1. Molecular Formula & Exact Mass (HR-MS)**

- Accurate mass from high-resolution mass spectrometry
- Molecular formula determination (C, H, N, O, S composition)
- Isotopic pattern analysis for element confirmation
- **Databases:** PubChem, ChemSpider, NIST Chemistry WebBook

## **2. Chemical Structure (2D/3D)**

- SMILES string representation
- InChI/InChIKey identifiers
- Canonical structure
- **Tools:** ChemDraw, RDKit, OpenBabel

## **3. CAS Registry Number & Synonyms**

- Chemical Abstracts Service number
- All known trivial names and IUPAC nomenclature
- Trade names and alternative identifiers
- **Databases:** SciFinder, ChemIDplus, Common Chemistry

## **4. Dereplication Search Results**

- Natural Products databases screening
- Previously isolated/characterized status
- Source organism information
- **Databases:** Dictionary of Natural Products (DNP), MarinLit, CMNPD, npatlas, KNApSACk

## **5. Spectroscopic Data Repository Search**

- $^1\text{H}$  NMR chemical shifts and coupling constants
- $^{13}\text{C}$  NMR chemical shifts
- 2D NMR correlations (COSY, HSQC, HMBC)
- **Databases:** NMRShiftDB, SDBS, BMRB, nmrXiv

## **6. Mass Spectral Fragmentation Patterns**

- MS/MS fragmentation pathways
- Diagnostic fragment ions
- Collision-induced dissociation patterns
- **Databases:** NIST MS Database, MoNA (MassBank of North America), METLIN

## **7. UV-Vis Absorption Maxima**

- $\lambda_{\text{max}}$  values in various solvents
- Chromophore identification
- Conjugation patterns
- **Databases:** SDBS, NIST Chemistry WebBook

## **8. Infrared Spectroscopy Data**

- Characteristic functional group absorptions
- Fingerprint region analysis
- **Databases:** SDBS, NIST IR Database

## **9. Optical Rotation $[\alpha]_D$**

- Specific rotation data
- Absolute stereochemistry indicators
- Enantiomeric excess information
- **Literature:** Primary research articles, Merck Index

## **10. Circular Dichroism (CD) Spectra**

- Cotton effects for stereochemistry
- Absolute configuration determination
- Exciton coupling patterns
- **Literature:** Natural products journals, specialized databases

## **11. X-ray Crystallographic Data**

- Crystal structure (if available)
- CIF (Crystallographic Information File)
- ORTEP diagrams
- **Databases:** Cambridge Structural Database (CSD), COD, CCDC

## **12. Existing Literature References**

- Primary isolation papers
- Structure elucidation studies
- Total synthesis reports
- **Databases:** SciFinder, Reaxys, Web of Science, PubMed

---

# **PHASE 2: BIOLOGICAL & PHARMACOLOGICAL DATA (Items 13-22)**

## **13. Bioactivity Profile**

- Known biological activities (antimicrobial, anticancer, anti-inflammatory)
- IC<sub>50</sub>/EC<sub>50</sub> values
- Target specificity
- **Databases:** ChEMBL, BindingDB, PubChem BioAssay, DrugBank

## **14. Target Protein/Enzyme Interactions**

- Known protein targets
- Binding mode and mechanisms
- Inhibition constants (Ki, Kd)
- **Databases:** BindingDB, PDBbind, Therapeutic Target Database (TTD), Guide to Pharmacology

## **15. ADME/PK Properties**

- Absorption, Distribution, Metabolism, Excretion data
- Pharmacokinetic parameters
- Bioavailability
- **Databases:** DrugBank, HMDB, PharmaGKB

## **16. Toxicity & Safety Data**

- LD<sub>50</sub> values (oral, dermal, inhalation)
- Mutagenicity/carcinogenicity data
- Environmental toxicity
- **Databases:** RTECS, TOXNET, ACToR, EPA CompTox Dashboard, CTD

## **17. Structure-Activity Relationship (SAR) Data**

- Analogs and derivatives
- Activity cliffs
- QSAR models
- **Databases:** ChEMBL, Reaxys, primary literature

## **18. Clinical Trial Status**

- Current clinical trials
- Drug development stage
- Previous trial outcomes
- **Databases:** ClinicalTrials.gov, DrugCentral, Clival Database

## **19. Protein Data Bank (PDB) Structures**

- Co-crystal structures with target proteins
- Ligand-protein complexes
- Binding site information
- **Database:** RCSB PDB, PDBe, PDBj

## 20. Enzyme Kinetics Data

- Km, Vmax values
- Enzyme inhibition mechanisms
- Substrate specificity
- **Databases:** BRENDA, SABIO-RK

## 21. Cell-Based Assay Results

- Cell viability data (MTT/MTS assays)
- Apoptosis/necrosis markers
- Cell line specificity
- **Databases:** PubChem BioAssay, ChEMBL

## 22. In Vivo Efficacy Data

- Animal model studies
- Disease model efficacy
- Therapeutic index
- **Literature:** Primary research articles, drug databases

---

# PHASE 3: COMPUTATIONAL & STRUCTURAL ANALYSIS (Items 23-32)

## 23. 3D Conformational Analysis

- Low-energy conformers
- Conformational flexibility
- Rotatable bonds
- **Tools:** RDKit, OpenBabel, Avogadro, molecular mechanics

## 24. Molecular Descriptors

- Molecular weight, logP, polar surface area (PSA)
- Lipinski's Rule of Five compliance
- Drug-likeness scores
- **Tools:** RDKit, ChemAxon, SwissADME

## **25. Protein Structure Prediction (Boltz-2)**

- If compound is a peptide/protein natural product
- AlphaFold Database predictions for target proteins
- Boltz-2 structure prediction for protein-ligand complexes
- **Databases:** AlphaFold DB, Boltz-2, PDB

## **26. Molecular Docking Results**

- Predicted binding poses
- Docking scores
- Key residue interactions
- **Tools:** AutoDock Vina, Glide, GOLD, Boltz-2 (for affinity prediction)

## **27. Molecular Dynamics Simulation Parameters**

- Force field compatibility (AMBER, CHARMM, OPLS)
- Solvation models
- Temperature/pressure conditions
- **Databases:** MISATO, mdCATH, ATLAS, literature protocols

## **28. Free Energy Calculations**

- Binding free energy ( $\Delta G$ )
- MM-PBSA/MM-GBSA results
- FEP/TI data (if available)
- **Tools:** GROMACS, AMBER, OpenMM, Boltz-2 affinity predictions

## **29. Quantum Mechanical Properties**

- DFT-optimized geometries
- HOMO-LUMO gaps
- Electrostatic potential maps
- **Tools:** Gaussian, ORCA, Q-Chem, PubChem computed properties

## **30. Electron Density Maps (CryoEM/X-ray)**

- High-resolution structural data
- Ligand electron density fitting
- B-factor analysis
- **Databases:** EMDB (Electron Microscopy Data Bank), PDB

## **31. Protein-Ligand Interaction Fingerprints**

- H-bonds,  $\pi$ - $\pi$  stacking, hydrophobic contacts

- Salt bridges
- Key binding site residues
- **Tools:** ProLIF, PLIP, LigPlot+, Boltz-2 output analysis

## 32. SMILES-Based Chemical Space Analysis

- Tanimoto similarity to known drugs
  - Scaffold analysis
  - Fragment analysis
  - **Tools:** RDKit, ChEMBL, PubChem similarity search
- 

# PHASE 4: BIOSYNTHETIC & METABOLIC PATHWAYS (Items 33-40)

## 33. Biosynthetic Pathway Information

- Polyketide synthase (PKS) pathways
- Non-ribosomal peptide synthetase (NRPS) pathways
- Terpenoid biosynthesis
- **Databases:** KEGG, MetaCyc, BioCyc, MIBiG

## 34. Precursor Molecules & Building Blocks

- Acetyl-CoA, malonyl-CoA units
- Amino acid origins
- Isoprene units (C5 building blocks)
- **Databases:** KEGG Compound, MetaCyc, BRENDA

## 35. Biosynthetic Gene Clusters (BGCs)

- Gene cluster architecture
- Core biosynthetic enzymes
- Regulatory elements
- **Databases:** antiSMASH, MIBiG, IMG-ABC, ClusterMine360

## 36. Enzyme Families Involved

- Oxidoreductases, transferases, hydrolases
- Cytochrome P450s
- Methyltransferases
- **Databases:** BRENDA, ENZYME (ExPASy), UniProt

## 37. Source Organism Taxonomy & Genomics

- Phylogenetic classification
- Genome sequences (if available)
- Secondary metabolite production capacity
- **Databases:** NCBI Taxonomy, NCBI Genome, JGI IMG/M, OneKP (for plants)

## 38. Metabolomic Context

- Co-occurring metabolites
- Metabolic network position
- Flux analysis data
- **Databases:** HMDB, METLIN, MetaboLights, MetaCyc

## 39. Isotope Labeling Studies

- $^{13}\text{C}$ ,  $^{15}\text{N}$ ,  $^2\text{H}$  incorporation patterns
- Feeding experiments
- Biosynthetic origin mapping
- **Literature:** Natural products chemistry journals

## 40. Post-Translational Modifications (for peptides)

- Glycosylation sites
- Phosphorylation
- Lipidation
- **Databases:** UniProt, PhosphoSitePlus, GlyConnect

---

# PHASE 5: SYNTHETIC & INDUSTRIAL CHEMISTRY (Items 41-50)

## 41. Total Synthesis Routes

- Published synthetic schemes
- Key synthetic steps
- Yields and conditions
- **Databases:** Reaxys, SciFinder, Organic Syntheses

## 42. Retrosynthetic Analysis

- Disconnection approaches
- Strategic bond formations
- Key intermediates
- **Tools:** SciFinder (Retrosynthesis), IBM RXN for Chemistry

## **43. Commercial Availability**

- Vendor sources
- Catalog numbers
- Pricing information
- Purity grades available
- **Databases:** eMolecules, ChemExper, Sigma-Aldrich, TCI, Cayman Chemical

## **44. Analytical Standards**

- Reference material availability
- Certified standards
- NMR/MS reference spectra
- **Sources:** Sigma-Aldrich, USP, NIST SRM

## **45. Patent Literature**

- Composition of matter patents
- Process patents
- Therapeutic use patents
- **Databases:** Google Patents, PATENTSCOPE, SciFinder

## **46. Pharmaceutical/Industrial Uses**

- Current commercial applications
- Market products containing the compound
- Dosage forms
- **Databases:** DrugBank, FDA Orange Book, drugs.com

## **47. Regulatory Status**

- FDA approval status
- EMA authorization
- Schedule classification (if applicable)
- **Databases:** FDA databases, ECHA REACH, DEA

## **48. Analytical Method Development**

- HPLC methods (columns, mobile phases, detection)
- GC-MS conditions
- Sample preparation protocols
- **Literature:** Pharmacopoeias (USP, EP, JP), method development papers

## **49. Physicochemical Properties**

- Melting point/boiling point
- Solubility in various solvents
- pKa values
- Stability (pH, temperature, light)
- **Databases:** Reaxys, DECHEMA DETHERM, CRC Handbook

## 50. Safety Data Sheet (SDS) Information

- Hazard classifications
  - Handling precautions
  - Storage requirements
  - Emergency procedures
  - **Sources:** Vendor websites, NIOSH, OSHA databases
- 

# WORKFLOW INTEGRATION NOTES

### Recommended Search Order:

1. **Start with Items 1-12** (Compound identification/dereplication) to confirm you're not duplicating known work
2. **Proceed to Items 13-22** (Biological data) to understand therapeutic potential
3. **Execute Items 23-32** (Computational analysis) for in silico predictions
4. **Investigate Items 33-40** (Biosynthesis) if you're studying production or engineering
5. **Review Items 41-50** (Synthetic/industrial) for practical applications

### Modern Integration Tools:

- **Boltz-2:** For protein-ligand structure prediction and binding affinity estimation (Items 25, 26, 28, 31)
- **Molecular Dynamics:** For conformational sampling and free energy calculations (Items 24, 27, 28)
- **CryoEM/X-ray:** For experimental structure validation (Items 11, 30)
- **HPLC-MS:** For analytical characterization and dereplication (Items 1, 6, 8, 48)

### Critical Databases for Natural Products Chemists:

- **Structure/Dereplication:** DNP, MarinLit, npatlas, ChemSpider
- **Spectroscopy:** NMRShiftDB, SDBS, MoNA
- **Bioactivity:** ChEMBL, BindingDB, PubChem BioAssay
- **Biosynthesis:** KEGG, MIBiG, antiSMASH
- **Structural Biology:** PDB, AlphaFold DB, EMDB
- **Computational:** Boltz-2, UniRef, ColabFold

---

## SPECIALIZED CONSIDERATIONS BY COMPOUND CLASS

### For Alkaloids:

- Focus on Items 33-40 (biosynthesis from amino acids)
- Special attention to Item 39 (isotope labeling for pathway elucidation)

### For Terpenoids:

- Emphasize Item 33 (mevalonate/MEP pathway)
- Item 37 (terpene synthase gene clusters)

### For Polyketides:

- Critical: Items 33, 35 (PKS gene clusters)
- Item 39 (acetate labeling patterns)

### For Peptides/Proteins:

- Essential: Items 25, 30, 40 (Boltz-2, CryoEM, PTMs)
- Item 34 (amino acid precursors)

### For Glycosides:

- Items 40, 33 (glycosyltransferases)
- Item 12 (carbohydrate structure databases)

---

*This checklist reflects the integrated workflow of a modern natural products chemist combining traditional isolation/characterization with cutting-edge computational structural biology and biosynthetic engineering approaches.*

---

## Top 50 Computational Resources and Software Packages for Natural Products Chemists and Organic Chemists

This list compiles the most prominent computational tools, software, and packages based on their relevance to organic synthesis planning, molecular modeling, quantum chemistry calculations, cheminformatics, docking, and structure prediction—key areas for organic chemists (e.g.,

reaction mechanisms, synthesis design) and natural products chemists (e.g., complex structure elucidation, bioactivity modeling, virtual screening). The selection and ranking prioritize frequency of mentions across sources, versatility, and community adoption, with a focus on tools supporting synthesis, simulation, and analysis of complex organic and natural compounds.

Rank	Name	Description
1	RDKit	Open-source cheminformatics toolkit for molecule manipulation, descriptor calculation, fingerprinting, and machine learning integration; widely used for virtual screening in natural products and organic synthesis design.
2	Gaussian	Quantum chemistry software for electronic structure calculations, energy optimizations, and reaction mechanism studies; essential for predicting properties of organic molecules and natural products.
3	Schrödinger Suite	Comprehensive platform for molecular modeling, docking, pharmacophore modeling, and drug discovery; supports natural product lead optimization and organic reaction simulations.
4	AutoDock	Free docking software for protein-ligand interactions; commonly applied in virtual screening of natural products for bioactivity.
5	Avogadro	Open-source molecular editor and visualization tool for building, editing, and optimizing 3D structures; useful for organic chemists in preliminary modeling.
6	IBM RXN	AI-based tool for reaction prediction and retrosynthesis planning; aids in designing synthetic routes for complex natural products and organic targets.
7	DeepChem	Open-source deep learning library for chemistry; enables property prediction and generative models for organic compounds and natural product analogs.
8	Open Babel	Toolbox for chemical file format conversion, substructure searching, and data processing; foundational for handling diverse organic and natural product datasets.
9	CDK (Chemistry Development Kit)	Java-based library for chemical structure representation, descriptor calculation, and 2D/3D visualization; supports cheminformatics workflows in organic research.
10	PyMOL	Molecular visualization software for 3D rendering and analysis; ideal for examining natural product-protein complexes.
11	AiZynthFinder	Open-source retrosynthetic planning tool; predicts routes for synthesizing organic molecules and natural product derivatives.
12	ASKCOS	Open-source suite for synthesis planning and computational chemistry; focuses on automated route design for complex organics.
13	Synthia (formerly Chematica)	Retrosynthesis software for generating viable synthetic pathways; used in organic synthesis and natural product total synthesis

<b>Rank</b>	<b>Name</b>	<b>Description</b>
14	MOE (Molecular Operating Environment)	Integrated suite for structure-based design, pharmacophore discovery, and QSAR; supports natural product drug design.
15	ChemDraw	Chemical structure drawing software; essential for illustrating organic reactions and natural product structures, with some computational export features.
16	HyperChem	Molecular modeling package for 3D design, simulations, and property calculations; applied in organic chemistry education and research.
17	Materials Studio	Simulation software for molecular dynamics and quantum mechanics; useful for studying organic materials and natural polymers.
18	GAMESS	General quantum chemistry package for ab initio calculations; predicts molecular properties in organic reactions.
19	MOLPRO	Quantum chemistry software specializing in high-accuracy electronic structure methods; for detailed organic mechanism studies.
20	Q-Chem	Quantum chemistry program for electronic structure and spectroscopy; supports organic and natural product vibrational analysis.
21	Glide	High-throughput docking tool; used for screening natural products against biological targets.
22	GOLD	Protein-ligand docking software; optimizes binding poses for organic ligands and natural compounds.
23	FlexX	Flexible docking program; incorporated in suites for organic inhibitor design.
24	DOCK	Structure-based virtual screening tool; free for academia, applied in natural products discovery.
25	Modeller	Homology modeling software; builds 3D models of proteins interacting with natural products.
26	Prime	Protein structure prediction and refinement tool; part of Schrödinger, for natural product binding studies.
27	PHASE	Pharmacophore modeling software; identifies features in active natural products for lead generation.
28	LigandScout	Pharmacophore and structure-based modeling tool; analyzes natural product interactions.
29	Discovery Studio	Biovia suite for biopolymer modeling, docking, and QSAR; comprehensive for organic medicinal chemistry.
30	SYBYL	Molecular modeling suite with pharmacophore and alignment tools; for organic structure-activity relationships.

<b>Rank</b>	<b>Name</b>	<b>Description</b>
31	ICM	Suite for docking, homology modeling, and virtual screening; supports natural product databases.
32	LeadIT	Docking and fragment-based design platform; includes FlexX for organic ligand optimization.
33	OpenEye	Toolkit for shape similarity, docking, and conformer generation; free for academia, used in organic screening.
34	GROMACS	Molecular dynamics simulator; models organic solvent effects and natural product dynamics.
35	AMBER	Suite for biomolecular simulations; predicts conformations of complex natural products.
36	CHARMM	Molecular dynamics and mechanics program; for energy minimizations in organic systems.
37	NAMD	Parallel molecular dynamics code; scales for large natural product-protein simulations.
38	PaDEL-Descriptor	Tool for calculating molecular descriptors; essential for QSAR in organic and natural products.
39	DataWarrior	Data visualization and analysis software with chemical intelligence; explores natural product datasets.
40	scikit-learn	Machine learning library; builds QSAR models for predicting properties of organic compounds.
41	TensorFlow	Deep learning framework; used for neural networks in organic reaction prediction.
42	PyTorch	Flexible deep learning library; supports graph neural networks for natural product structures.
43	Spaya	AI retrosynthesis planning tool; free access for generating organic synthesis routes.
44	Manifold	Compound searching and retrosynthesis tool; academic-free, for natural product analogs.
45	VMD	Visualization program for molecular dynamics; analyzes trajectories of organic simulations.
46	Jmol	Java-based 3D viewer for chemical structures; interactive for natural product education.
47	CP2K	Open-source quantum chemistry and solid-state physics software; for hybrid QM/MM in organics.
48	Quantum ESPRESSO	Suite for electronic-structure calculations; applies to organic materials and interfaces.
49	MayaChemTools	Command-line tools for descriptor calculation and substructure searching in organics.
50	Indigo Toolkit	Library for molecule manipulation and rendering; supports organic database handling.

---

Good question. Here are several articles / PDFs / books that address dashboard design — including ones that are directly relevant to cheminformatics, and others that offer general dashboard-design guidance (often adaptable to cheminformatics software).

---

## Key sources relevant to cheminformatics dashboards

- **Cheminformatics Microservice V3: a web portal for chemical structure manipulation and analysis** — this 2025 article describes a web-based frontend (built with React) for a cheminformatics backend (toolkits like RDKit, CDK, Open Babel). The paper gives details of the interface design, interactive features (structure drawing, batch depiction, filtering/search, integration with external databases) and architecture. ([ResearchGate](#))
- **The CompTox Chemistry Dashboard: a community data resource for environmental chemistry** — describes a mature, widely used cheminformatics dashboard built by U.S. Environmental Protection Agency (EPA). The article covers how chemical-structure, physicochemical, toxicological and other data are integrated, curated, and presented via a web interface — useful as a real-world example of cheminformatics dashboard design. ([SpringerLink](#))
- **GenUI: interactive and extensible open source software platform for de novo molecular generation and cheminformatics** — this 2021 article outlines an open-source cheminformatics / de novo molecular generation platform providing a GUI, data preprocessing, model building, molecule generation, and interactive chemical-space visualization. While not a “dashboard” in the strict sense, the design and architecture are highly relevant for building cheminformatics user interfaces that blend data analytics + visualization. ([SpringerLink](#))

---

## General dashboard-design / visualization-design references (useful for cheminformatics context)

Because there is relatively little literature *specifically* about “cheminformatics dashboard design,” these more general works on dashboard design and data visualization can provide core design principles and patterns that you can adapt:

- **Dashboard Design** by Michael Burch and Marco Schmid (2023) — a comprehensive text that lays out the architecture, components, layout strategies, and design workflows for dashboards. Good if you plan to build or design a dashboard from scratch. ([ResearchGate](#))

- **Dashboard Design Patterns** (Bach et al., 2023) — this article presents a taxonomy of dashboard “design patterns”, distilled from a systematic review of 144 dashboards. It discusses layout, interaction, tradeoffs, and how to combine patterns depending on needs (e.g. analytical vs narrative dashboards). This can help structure how you display chemical data / analysis in a clean, user-friendly way. ([Enlighten Publications](#))
  - **Representing Data Visualization Goals and Tasks through a Dashboard Meta- Model** (Vázquez-Ingelmo et al., 2020) — describes a “meta-model” for dashboards that links data structures, user goals, domain knowledge, and dashboard features; useful if you want a principled, formal way to design dashboards tailored to particular use cases (like cheminformatics). ([MDPI](#))
- 

## ⌚ How these help you (concrete uses)

Since you work in computational biology / neuroscience **and** have cheminformatics interest, these resources could help you:

- Use the *CompTox Dashboard* and *Cheminformatics Microservice V3* as real world blueprints — see how chemical data, metadata, structure-drawing, and analyses are integrated into a unified UI.
  - Use general dashboard design theory (from Dashboard Design / Dashboard Design Patterns / meta-model) to build your own internal tools — for example, chemical-library browsers, QSAR analysis dashboards, or custom visualization tools for chemical-biological data integration.
  - Leverage the architecture described in GenUI or Microservice V3 — combining backend toolkits (RDKit/CDK/Open Babel) with a web frontend — to build a modular, extensible “lab-wide” cheminformatics dashboard or portal.
- 

If you like — I can **assemble a list of 10–15 of the best academic articles (with PDFs, DOIs, or accessible links) that specifically combine cheminformatics + UI / dashboard design** (so you get a ready “reading pack”).

---

Sure — here’s a curated “reading pack” of **10–15 articles, papers, and resources** that focus (or are very relevant) to the design of dashboards, GUIs or front-ends for cheminformatics / chemical-data software. I selected ones that combine actual cheminformatics functionality with user-interface, visualization, or data-portal design. I grouped them by type for easier browsing.

---

## Primary — Cheminformatics Dashboards / Web Portals / GUIs

Reference (year)	What it is / Why relevant
<b>Cheminformatics Microservice V3: a web portal for chemical structure manipulation and analysis</b> (2025) — Rajan et al.	A modern web-portal for cheminformatics: integrates popular toolkits (RDKit, Chemistry Development Kit (CDK), Open Babel) via a REST API + provides an interactive React-based frontend for structure editing, batch processing, PubChem integration, identifier generation, etc. Great concrete example of dashboard + UI design. ( <a href="#">SpringerLink</a> )
<b>The CompTox Chemistry Dashboard: A community data resource for environmental chemistry</b> (2017) — Williams et al.	A widely used public chemical dashboard (by the U.S. Environmental Protection Agency, EPA). Integrates chemical-structure, physicochemical, toxicity, exposure, bioassay and other data into a unified web interface. Excellent example of large-scale dashboard design for chemical data. ( <a href="#">ResearchGate</a> )
<b>ChemInformatics Model Explorer (CIME): exploratory visual analysis of cheminformatics model predictions</b> (2022) — Humer et al.	A web-based interactive system for examining chemical datasets, visualizing model explanations (e.g. machine-learning models), comparing interpretability techniques, and exploring chemical data — relevant for dashboards bridging cheminformatics + predictive modeling. ( <a href="#">PMC</a> )
<b>GenUI: interactive and extensible open source software platform for de novo molecular generation and cheminformatics</b> (2021) — Sicho et al.	A GUI + API framework that supports compound import, generation, storage, retrieval, visualization of molecular datasets, QSAR modeling, etc. While more than a “dashboard,” its design emphasizes usability, interaction, and integration. ( <a href="#">SpringerLink</a> )
<b>COEL: A Web-based Chemistry Simulation Framework</b> (2014) — Banda et al.	Although focused on chemical reaction network simulations, COEL offers a web-based UI + visualization + data-storage + simulation controls — valuable for thinking about how chemical data + dynamic computation + UI can be combined in one portal. ( <a href="#">arXiv</a> )

## Supporting / Adjacent — Visualization Tools, Data Integration & Platform Design

Reference (year)	Relevance / What it adds
<b>Open-Source Software Development in Cheminformatics</b> (2023) — Pernaa et al.	A meta-study exploring motivations, challenges, and patterns behind open-source cheminformatics development. While not a “dashboard design” paper per se, it helps frame why usability / user interface / accessibility

Reference (year)	Relevance / What it adds
<b>From cheminformatics to structure-based design: Web applications linking chemical databases and modeling tools</b> (2017) — Bietz & Bajorath	is critical in cheminformatics software, which informs design decisions. ( <a href="#">MDPI</a> )
<b>A web server to design chemistry-driven focused libraries</b> (2024) — Hoffer et al.	Discusses how web-based cheminformatics applications can provide data management (database storage/search/filter) + modeling tools + structural design workflows. Useful for conceptualizing dashboards bridging storage, search, modeling, and UI. ( <a href="#">ScienceDirect</a> )
<b>PlayMolecule Viewer: a toolkit for the visualization of molecules and other data</b> (2023) — Torrens-Fontanals et al.	A more specialized application: a web-server for generating focused chemical libraries. While its GUI might be narrower in scope than a full “dashboard,” its design offers insight into how chemical-library workflows can be encapsulated in web UIs. ( <a href="#">OUP Academic</a> )
	A web-based data visualization toolkit oriented toward molecular data and structural bioinformatics — useful for embedding molecular visualization and data browsing into dashboards. ( <a href="#">arXiv</a> )

---

## ✉ Review / Contextual — Software Development Practices, Data Portals, Open Source Trends

Reference (year)	Why you might read it (contextual value)
“Open-source development motivations” (see above) — helps you plan a dashboard project with long-term maintenance, community adoption, and sustainability in mind. ( <a href="#">MDPI</a> )	Why you might read it (contextual value)
Comparison and analysis of existing platforms — helpful when deciding whether to build your own dashboard or adapt/extend existing tools (e.g. comparing open vs proprietary toolkits, integration burdens, user-interface tradeoffs). For this, the conceptual discussions in the above articles (e.g. Bietz & Bajorath, GenUI) are quite useful.	Why you might read it (contextual value)
Papers combining chemical data + modeling + UI (e.g. CIME, web-library generation servers) — helpful to think about how dashboards can support not just data viewing, but generation, modeling, prediction, and interactive analysis.	Why you might read it (contextual value)

---

## ☑ Recommendations for How to Use This Pack

- If you want a ready-to-use dashboard (for chemical data search, structure editing, basic analysis) — start with *Cheminformatics Microservice V3*, *CompTox Dashboard*, or *GenUI*.
- If you plan to build your own: examine *Bietz & Bajorath (2017)* and *Open-Source Software Development in Cheminformatics (2023)* for architectural, design, and sustainability lessons.
- If you care about visualization & user interaction (e.g. chemical-space visualization, modeling results, data exploration) — review *CIME (2022)* and *PlayMolecule Viewer (2023)*.
- If you expect to expand beyond data browsing (e.g. generate new molecules, simulate reactions, run predictions) — consider *COEL (2014)* or the 2024 web-library generation server.

---

If you like — I can also **filter** this list for only **open-source or freely accessible PDFs / full-text** (so you don't hit paywalls).

---

I'll search for information about the key methods used by natural products chemists. Based on my research, here are the **top 30 methods routinely used by chemists who focus on natural products**:

## Extraction & Initial Isolation Methods

1. **Solvent Extraction (Maceration)** - Traditional method using organic solvents (methanol, ethanol, dichloromethane) to extract bioactive compounds from plant materials
2. **Soxhlet Extraction** - Continuous extraction using heated solvent reflux for exhaustive extraction of compounds
3. **Liquid-Liquid Partitioning** - Sequential extraction with solvents of different polarities to fractionate crude extracts
4. **Steam Distillation** - Extraction of volatile compounds and essential oils using steam
5. **Supercritical Fluid Extraction (SFE)** - Modern green extraction using supercritical CO<sub>2</sub> for temperature-sensitive compounds
6. **Ultrasound-Assisted Extraction (UAE)** - Accelerated extraction using ultrasonic waves to disrupt cell walls
7. **Microwave-Assisted Extraction (MAE)** - Rapid extraction using microwave energy to enhance solvent penetration
8. **Pressurized Solvent Extraction (PSE)** - High-pressure, high-temperature extraction for improved efficiency

## Chromatographic Separation & Purification Methods

9. **Column Chromatography** - Gravity-driven separation using silica gel or other stationary phases
10. **High-Performance Liquid Chromatography (HPLC)** - High-resolution analytical and preparative separation technique
11. **Preparative HPLC** - Large-scale purification for isolating milligram to gram quantities of pure compounds
12. **Thin-Layer Chromatography (TLC)** - Rapid analytical technique for monitoring fractionation and purity
13. **Flash Chromatography** - Rapid, pressure-assisted column chromatography for efficient purification
14. **High-Speed Countercurrent Chromatography (HSCCC)** - Liquid-liquid partition chromatography without solid support
15. **Solid Phase Extraction (SPE)** - Sample cleanup and pre-concentration using cartridges with various stationary phases
16. **Vacuum Liquid Chromatography (VLC)** - Low-pressure column chromatography for rapid pre-fractionation
17. **Size Exclusion Chromatography (SEC)** - Separation based on molecular size for large biomolecules
18. **Ion Exchange Chromatography** - Separation of charged compounds based on ionic interactions

## Spectroscopic Characterization Methods

19.  **$^1\text{H}$  NMR Spectroscopy** - Hydrogen nuclear magnetic resonance for determining molecular structure and connectivity
20.  **$^{13}\text{C}$  NMR Spectroscopy** - Carbon-13 NMR for establishing carbon skeleton and functional groups
21. **2D NMR Techniques** (COSY, HSQC, HMBC, NOESY, ROESY) - Multi-dimensional NMR for complete structure elucidation including stereochemistry
22. **High-Resolution Mass Spectrometry (HR-MS)** - Accurate molecular weight and elemental composition determination
23. **LC-MS/MS (Tandem Mass Spectrometry)** - Coupled liquid chromatography and tandem mass spectrometry for complex mixture analysis
24. **GC-MS (Gas Chromatography-Mass Spectrometry)** - Analysis of volatile and semi-volatile compounds with electron ionization
25. **UV-Vis Spectroscopy** - Electronic absorption spectroscopy for chromophore identification
26. **Infrared Spectroscopy (IR/FTIR)** - Functional group identification through vibrational spectroscopy
27. **Circular Dichroism (CD) Spectroscopy** - Determination of absolute stereochemistry and chirality at picomole levels
28. **X-ray Crystallography** - Definitive 3D structure determination for crystalline compounds

## Biological Activity Assessment Methods

29. **Bioassay-Guided Fractionation** - Iterative process of separation and biological testing to identify active compounds
30. **Antimicrobial Assays** - Multiple methods including:

- **Agar Disk Diffusion** (Kirby-Bauer method)
- **Broth Microdilution** (MIC determination)
- **Agar Well Diffusion**
- **Bioautography** (TLC-based activity detection)

## Additional Key Techniques

- **Recrystallization** - Classical purification achieving high purity through controlled crystallization
- **LC-NMR** - Online coupling of liquid chromatography with NMR for mixture analysis
- **Molecular Networking** - Computational approach using MS/MS data to visualize chemical relationships
- **Dereplication** - Rapid identification of known compounds using databases to avoid re-isolation
- **Computer-Assisted Structure Elucidation (CASE)** - Software tools for predicting structures from spectroscopic data
- **MTT/MTS Cytotoxicity Assays** - Cell viability testing for anticancer activity
- **DPPH Radical Scavenging Assay** - Antioxidant activity measurement
- **Enzyme Inhibition Assays** - Testing for specific biological targets
- **ELISA (Enzyme-Linked Immunosorbent Assay)** - Immunoassays for specific compound detection
- **Surface Plasmon Resonance (SPR)** - Label-free detection of biomolecular interactions

These methods are typically used in integrated workflows, combining extraction, separation, structural characterization, and bioactivity testing to discover, isolate, and characterize bioactive natural products for pharmaceutical, nutraceutical, and agricultural applications.

---

I'll search for information about Boltz-2 and the instruments and databases commonly used with this software. Based on my research, here are the **top 30 instruments and databases routinely used by scientists working with Boltz-2**:

## Sequence Databases & MSA Generation

1. **UniRef30/UniRef90/UniRef100** - Clustered protein sequence databases from UniProt, used for generating multiple sequence alignments (MSAs)
2. **MMseqs2** - Ultra-fast sequence search and clustering tool that searches massive sequence databases to construct MSAs
3. **ColabFold Database (ColabFoldDB)** - Merged environmental sequence database including BFD, MGnify, MetaEuk, SMAG, TOPAZ, MGV, and GPD

4. **BFD (Big Fantastic Database)** - Large-scale protein sequence database used for MSA generation
5. **MGnify** - Metagenomic sequence database providing environmental protein sequences
6. **ColabFold MSA Server** - Public MSA generation server ([api.colabfold.com](http://api.colabfold.com)) hosted by KOBIC (Korean Bioinformatics Center)

## Structural Databases

7. **Protein Data Bank (PDB/RCSB PDB)** - Primary database of experimentally determined 3D structures of proteins, nucleic acids, and complexes
8. **AlphaFold Database (AFDB)** - Database of predicted protein structures from AlphaFold
9. **PDB70** - Clustered subset of PDB sequences used for template searches
10. **OpenFold** - Open-source protein structure prediction model whose predictions are used in Boltz training

## Binding Affinity & Chemical Databases

11. **ChEMBL** - Manually curated database of bioactive molecules with drug-like properties and standardized binding/functional assay data
12. **BindingDB** - Public database of measured binding affinities (Ki, Kd values) for protein-ligand interactions
13. **PDBbind** - Database linking PDB structures with binding affinity data
14. **PubChem** - Comprehensive database of chemical molecules and their biological activities
15. **MF-PCBA** - High-throughput screening dataset used for benchmarking virtual screening
16. **CeMM Fragment Dataset** - Fragment-screening results from fragment-based drug discovery campaigns
17. **MIDAS (University of Utah)** - Protein-metabolite interactome dataset

## Molecular Dynamics Databases

18. **MISATO** - Molecular dynamics dataset containing NVT trajectories simulated at 300K for 8ns
19. **mdCATH** - Large-scale MD trajectory database with simulations up to 500ns
20. **ATLAS** - Extensive molecular dynamics simulation dataset

## Computational Hardware & Software Infrastructure

21. **NVIDIA GPUs** (A100, H100, L40S, RTX series) - Graphics processing units for accelerated computation
22. **CUDA/cuDNN** - NVIDIA's parallel computing platform and deep learning acceleration library
23. **NVIDIA cuEquivariance** - Accelerated kernels for geometry-aware neural networks, including Triangle Attention and Triangle Multiplication

24. **PyTorch** - Deep learning framework used as Boltz-2's underlying ML infrastructure
25. **Pixi/Conda** - Package management systems for dependency installation

## Visualization & Analysis Tools

26. *Mol (Molstar Viewer)*\* - Web-based molecular visualization tool for viewing predicted structures
27. **PyMOL** - Molecular visualization system for protein structure viewing
28. **ChimeraX** - Molecular visualization program with integrated Boltz-2 support
29. **py3Dmol** - Python library for 3D molecular visualization

## Complementary Computational Tools

30. **OpenFE/FEP+** - Free energy perturbation tools for binding affinity calculations (used for benchmarking against Boltz-2)

## Additional Notable Tools

- **RDKit** - Cheminformatics toolkit for SMILES processing and molecular manipulation
- **OpenBabel** - Chemical file format conversion tool
- **SynFlowNet** - GFlowNet-based molecular generator for coupling with Boltz-2
- **GROMACS/AMBER/OpenMM/NAMD** - Molecular dynamics simulation engines
- **VDJdb** - T-cell receptor sequence database
- **Stitchr/Thimble** - Tools for reconstructing full-length TCR sequences
- **IEDB (Immune Epitope Database)** - Database for TCR-pMHC modeling
- **Rfam** - RNA families database

These instruments and databases form an integrated ecosystem enabling scientists to perform structure prediction, binding affinity estimation, virtual screening, and molecular design using Boltz-2.

---

I'll search for information about Boltz-2 and the instruments and databases commonly used with this software. Based on my research, here are the **top 30 instruments and databases routinely used by scientists working with Boltz-2**:

## Sequence Databases & MSA Generation

1. **UniRef30/UniRef90/UniRef100** - Clustered protein sequence databases from UniProt, used for generating multiple sequence alignments (MSAs)
2. **MMseqs2** - Ultra-fast sequence search and clustering tool that searches massive sequence databases to construct MSAs

3. **ColabFold Database (ColabFoldDB)** - Merged environmental sequence database including BFD, MGnify, MetaEuk, SMAG, TOPAZ, MGV, and GPD
4. **BFD (Big Fantastic Database)** - Large-scale protein sequence database used for MSA generation
5. **MGnify** - Metagenomic sequence database providing environmental protein sequences
6. **ColabFold MSA Server** - Public MSA generation server ([api.colabfold.com](http://api.colabfold.com)) hosted by KOBIC (Korean Bioinformatics Center)

## Structural Databases

7. **Protein Data Bank (PDB/RCSB PDB)** - Primary database of experimentally determined 3D structures of proteins, nucleic acids, and complexes
8. **AlphaFold Database (AFDB)** - Database of predicted protein structures from AlphaFold
9. **PDB70** - Clustered subset of PDB sequences used for template searches
10. **OpenFold** - Open-source protein structure prediction model whose predictions are used in Boltz training

## Binding Affinity & Chemical Databases

11. **ChEMBL** - Manually curated database of bioactive molecules with drug-like properties and standardized binding/functional assay data
12. **BindingDB** - Public database of measured binding affinities (Ki, Kd values) for protein-ligand interactions
13. **PDBbind** - Database linking PDB structures with binding affinity data
14. **PubChem** - Comprehensive database of chemical molecules and their biological activities
15. **MF-PCBA** - High-throughput screening dataset used for benchmarking virtual screening
16. **CeMM Fragment Dataset** - Fragment-screening results from fragment-based drug discovery campaigns
17. **MIDAS (University of Utah)** - Protein-metabolite interactome dataset

## Molecular Dynamics Databases

18. **MISATO** - Molecular dynamics dataset containing NVT trajectories simulated at 300K for 8ns
19. **mdCATH** - Large-scale MD trajectory database with simulations up to 500ns
20. **ATLAS** - Extensive molecular dynamics simulation dataset

## Computational Hardware & Software Infrastructure

21. **NVIDIA GPUs** (A100, H100, L40S, RTX series) - Graphics processing units for accelerated computation
22. **CUDA/cuDNN** - NVIDIA's parallel computing platform and deep learning acceleration library

23. **NVIDIA cuEquivariance** - Accelerated kernels for geometry-aware neural networks, including Triangle Attention and Triangle Multiplication
24. **PyTorch** - Deep learning framework used as Boltz-2's underlying ML infrastructure
25. **Pixi/Conda** - Package management systems for dependency installation

## Visualization & Analysis Tools

26. **Mol (Molstar Viewer)\*** - Web-based molecular visualization tool for viewing predicted structures
27. **PyMOL** - Molecular visualization system for protein structure viewing
28. **ChimeraX** - Molecular visualization program with integrated Boltz-2 support
29. **py3Dmol** - Python library for 3D molecular visualization

## Complementary Computational Tools

30. **OpenFE/FEP+** - Free energy perturbation tools for binding affinity calculations (used for benchmarking against Boltz-2)

## Additional Notable Tools

- **RDKit** - Cheminformatics toolkit for SMILES processing and molecular manipulation
- **OpenBabel** - Chemical file format conversion tool
- **SynFlowNet** - GFlowNet-based molecular generator for coupling with Boltz-2
- **GROMACS/AMBER/OpenMM/NAMD** - Molecular dynamics simulation engines
- **VDJdb** - T-cell receptor sequence database
- **Stitchr/Thimble** - Tools for reconstructing full-length TCR sequences
- **IEDB (Immune Epitope Database)** - Database for TCR-pMHC modeling
- **Rfam** - RNA families database

These instruments and databases form an integrated ecosystem enabling scientists to perform structure prediction, binding affinity estimation, virtual screening, and molecular design using Boltz-2.

---

**Cheminformatics** (also known as **chemoinformatics**) refers to the use of [physical chemistry](#) theory with [computer](#) and [information science](#) techniques—so called "*in silico*" techniques—in application to a range of descriptive and prescriptive problems in the field of [chemistry](#), including in its applications to [biology](#) and [related molecular fields](#). Such *in silico* techniques are used, for example, by [pharmaceutical companies](#) and in academic settings to aid and inform the process of [drug discovery](#), for instance in the design of well-defined [combinatorial libraries](#) of synthetic compounds, or to assist in [structure-based drug design](#). The methods can also be used in chemical and allied industries, and such fields as [environmental science](#) and [pharmacology](#), where chemical processes are involved or studied.<sup>[1]</sup>

# History

Cheminformatics has been an active field in various guises since the 1970s and earlier, with activity in academic departments and commercial pharmaceutical research and development departments.<sup>[2][page needed][citation needed]</sup> The term chemoinformatics was defined in its application to drug discovery by F.K. Brown in 1998.<sup>[3]</sup>

Cheminformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization.

Since then, both terms, cheminformatics and chemoinformatics, have been used,<sup>[citation needed]</sup> although, [lexicographically](#), cheminformatics appears to be more frequently used,<sup>[when?][4][5]</sup> despite academics in Europe declaring for the variant chemoinformatics in 2006.<sup>[6]</sup> In 2009, a prominent Springer journal in the field was founded by transatlantic executive editors named the [Journal of Cheminformatics](#).<sup>[7]</sup>

## Background

Cheminformatics combines the scientific working fields of chemistry, computer science, and information science—for example in the areas of [topology](#), [chemical graph theory](#), [information retrieval](#) and [data mining](#) in the [chemical space](#).<sup>[8][page needed][9][page needed][10][11][page needed]</sup>

Cheminformatics can also be applied to data analysis for various industries like [paper](#) and [pulp](#), dyes and such allied industries.<sup>[12]</sup>

## Applications

### Storage and retrieval



This section **does not** cite any [sources](#). Please help [improve this section](#) by [adding citations to reliable sources](#). Unsourced material may be challenged and [removed](#).  
(February 2020) ([Learn how and when to remove this message](#))

Main article: [Chemical database](#)

A primary application of cheminformatics is the storage, indexing, and search of information relating to chemical compounds.<sup>[citation needed]</sup> The efficient search of such stored information includes topics that are dealt with in computer science, such as data mining, information retrieval, [information extraction](#), and [machine learning](#).<sup>[citation needed]</sup> Related research topics include:<sup>[citation needed]</sup>

- [Digital libraries](#)
- [Unstructured data](#)
- [Structured data mining](#) and mining of [structured data](#)
  - [Database mining](#)

- [Graph mining](#)
- [Molecule mining](#)
- [Sequence mining](#)
- [Tree mining](#)

## File formats

Main article: [Chemical file format](#)

The *in silico* representation of chemical structures uses specialized formats such as the [Simplified molecular input line entry specifications](#) (SMILES)<sup>[13]</sup> or the [XML](#)-based [Chemical Markup Language](#).<sup>[14]</sup> These representations are often used for storage in large [chemical databases](#).<sup>[citation needed]</sup> While some formats are suited for visual representations in two- or three-dimensions, others are more suited for studying physical interactions, modeling and docking studies.<sup>[citation needed]</sup>

## Virtual libraries

This section **relies excessively on references to primary sources**. Please improve this section by adding [secondary or tertiary sources](#).



*Find sources: "Cheminformatics" – news · newspapers · books · scholar · JSTOR (February 2020)*  
[\(Learn how and when to remove this message\)](#)

Chemical data can pertain to real or virtual molecules. Virtual libraries of compounds may be generated in various ways to explore chemical space and hypothesize novel compounds with desired properties. Virtual libraries of classes of compounds (drugs, natural products, diversity-oriented synthetic products) were recently generated using the FOG (fragment optimized growth) algorithm.<sup>[15]</sup> This was done by using cheminformatic tools to train transition probabilities of a [Markov chain](#) on authentic classes of compounds, and then using the Markov chain to generate novel compounds that were similar to the training database.

## Virtual screening



This section **does not cite any sources**. Please help [improve this section](#) by [adding citations to reliable sources](#). Unsourced material may be challenged and [removed](#).  
*(February 2020)* [\(Learn how and when to remove this message\)](#)

Main article: [Virtual screening](#)

In contrast to [high-throughput screening](#), virtual screening involves computationally screening *in silico* libraries of compounds, by means of various methods such as [docking](#), to identify members likely to possess desired properties such as [biological activity](#) against a given target. In some cases, [combinatorial chemistry](#) is used in the development of the library to increase the efficiency in mining the chemical space. More commonly, a diverse library of small molecules or [natural products](#) is screened.

## Quantitative structure-activity relationship (QSAR)

Main article: [Quantitative structure–activity relationship](#)

This is the calculation of [quantitative structure–activity relationship](#) and [quantitative structure property relationship](#) values, used to predict the activity of compounds from their structures. In this context there is also a strong relationship to [chemometrics](#). Chemical [expert systems](#) are also relevant, since they represent parts of chemical knowledge as an *in silico* representation. There is a relatively new concept of [matched molecular pair analysis](#) or prediction-driven MMPA which is coupled with QSAR model in order to identify activity cliff.<sup>[16]</sup>

## See also

- [Bioinformatics](#)
- [Chemical file format](#)
- [Chemicalize.org](#)
- [Cheminformatics toolkits](#)
- [Chemogenomics](#)
- [Computational chemistry](#)
- [Information engineering](#)
- [Journal of Chemical Information and Modeling](#)
- [Journal of Cheminformatics](#)
- [Materials informatics](#)
- [Molecular design software](#)
- [Molecular graphics](#)
- [Molecular Informatics](#)
- [Molecular modelling](#)
- [Nanoinformatics](#)
- [Software for molecular modeling](#)
- [WorldWide Molecular Matrix](#)
- [Molecular descriptor](#)

## References

- 1.
- Thomas Engel (2006). "Basic Overview of Chemoinformatics". *J. Chem. Inf. Model.* **46** (6): 2267–77. [doi:10.1021/ci600234z](https://doi.org/10.1021/ci600234z). [PMID 17125169](#).
- • Martin, Yvonne Connolly (1978). *Quantitative Drug Design: A Critical Introduction*. Medicinal Research series. Vol. 8 (1st ed.). New York, NY: [Marcel Dekker](#). [ISBN 9780824765743](#).
- • F.K. Brown (1998). "Ch. 35. Chemoinformatics: What is it and How does it Impact Drug Discovery". *Annual Reports in Medicinal Chemistry*. Vol. 33. pp. 375–384. [doi:10.1016/S0065-7743\(08\)61100-8](https://doi.org/10.1016/S0065-7743(08)61100-8). [ISBN 9780120405336](#); [page needed] see also Brown, Frank (2005).

"*Chemoinformatics—A Ten Year Update*". *Current Opinion in Drug Discovery & Development*. **8** (3): 296–302.

- • "[Cheminformatics or Chemoinformatics ?](#)". Archived from the original on 2017-06-21. Retrieved 2006-03-31.
  - • "[Biopharmaceutical glossary Tips & FAQs](#)".
  - • "[The Obernai Declaration](#)" (PDF). Chéminformatique Strasbourg. Archived (PDF) from the original on March 3, 2016.
  - • Willighagen, Egon. "[Open Access Journal of Chemoinformatics now live! « SteinBlog](#)". Retrieved 2022-06-20.
  - • Gasteiger J.; Engel T., eds. (2004). *Chemoinformatics: A Textbook*. New York, NY: Wiley. [ISBN 3527306811](#).
  - • Leach, A.R.; Gillet, V.J. (2003). *An Introduction to Chemoinformatics*. Berlin, DE: Springer. [ISBN 1402013477](#).
  - • Varnek, A.; Baskin, I. (2011). "Chemoinformatics as a Theoretical Chemistry Discipline". *Molecular Informatics*. **30** (1): 20–32. [doi:10.1002/minf.201000100](#). [PMID 27467875](#). [S2CID 21604072](#).
  - • Bunin, B.A.; Siesel, B.; Morales, G.; Bajorath J. (2006). *Chemoinformatics: Theory, Practice, & Products*. New York, NY: Springer. [ISBN 9781402050008](#).
  - • Williams, Tova; University, North Carolina State. "[Cheminformatics approaches to creating new hair dyes](#)". phys.org. Retrieved 2022-06-20.
  - • Weininger, David (1988). "SMILES, a Chemical Language and Information System: I: Introduction to Methodology and Encoding Rules". *Journal of Chemical Information and Modeling*. **28** (1): 31–36. [doi:10.1021/ci00057a005](#). [S2CID 5445756](#).
  - • Murray-Rust, Peter; Rzepa, Henry S. (1999). "Chemical Markup, XML, and the Worldwide Web. I. Basic Principles". *Journal of Chemical Information and Computer Sciences*. **39** (6): 928–942. [doi:10.1021/ci990052b](#).
  - • Kutchukian, Peter; Lou, David; Shakhnovich, Eugene (2009). "FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules occupying Druglike Chemical". *Journal of Chemical Information and Modeling*. **49** (7): 1630–1642. [doi:10.1021/ci9000458](#). [PMID 19527020](#).
16. • Sushko, Yurii; Novotarskyi, Sergii; Körner, Robert; Vogt, Joachim; Abdelaziz, Ahmed; Tetko, Igor V. (2014). "[Prediction-driven matched molecular pairs to interpret QSARs and aid the molecular optimization process](#)". *Journal of Chemoinformatics*. **6** (1): 48. [doi:10.1186/s13321-014-0048-0](#). [PMC 4272757](#). [PMID 25544551](#).

## Further reading

- Engel, Thomas (2006). "Basic Overview of Chemoinformatics". *J. Chem. Inf. Model.* **46** (6): 2267–2277. [doi:10.1021/ci600234z](#). [PMID 17125169](#).
- Martin, Yvonne Connolly (2010). *Quantitative Drug Design: A Critical Introduction* (2nd ed.). Boca Raton, FL: CRC Press–Taylor & Francis. [ISBN 9781420070996](#).
- Leach, A.R.; Gillet, V.J. (2003). *An Introduction to Chemoinformatics*. Berlin, DE: Springer. [ISBN 1402013477](#).

- *Gasteiger J.; Engel T., eds. (2004). Chemoinformatics: A Textbook. New York, NY: Wiley. ISBN 3527306811.*
- *Varnek, A.; Baskin, I. (2011). "Chemoinformatics as a Theoretical Chemistry Discipline". Molecular Informatics. 30 (1): 20–32. doi:10.1002/minf.201000100. PMID 27467875. S2CID 21604072.*
- *Bunin, B.A.; Siesel, B.; Morales, G.; Bajorath J. (2006). Chemoinformatics: Theory, Practice, & Products. New York, NY: Springer. ISBN 9781402050008.*

## External links

---