NAOMI ChemBio Suite 22

an academic software collection

Prof. Dr. Matthias Rarey
Universität Hamburg
ZBH – Center for Bioinformatics
Research Group for Computational Molecular Design

17 – 11 – 2022
Overview

**Cheminformatics**

Mona: Managing compound collections ........................................ 5
UNICON: The universal file format converter .................................. 6
Conformator: Generating conformer ensembles .............................. 7
SMARTSeditor: Interactive and semi-automated graphical creation of SMARTS expressions .............................. 8
SMARTScompare: Analyze and compare molecular patterns .......... 9
REMUS: Shape-based alignments – interactive exploration and screening .................................................. 10
CSFPy: The all-connected-subgraph topological fingerprint .......... 11
TorsionPatternMiner/TorsionAnalyzer: Create knowledge-based torsion profiles and analyze conformations .............................. 12
Phariety: Pharmacophore matching made easy ............................. 13
ReactionViewer: Visualizing chemical reactions with high precision .................................................. 14

**Chemical Space**

SpaceLight: Topological similarity searching in large chemical spaces .................................................. 16
SpaceMACS: Maximum common substructure searching in large chemical spaces .......................................... 17
SpaceCompare: Analyze, compare, and optimize ultra-large combinatorial spaces .................................................. 18
Galileo: Genetic algorithm for general purpose searching in chemical space .................................................. 19

**Structure-Based Design**

ASCONA/SIENA: Active site search and alignment .......................... 21
EDIAscorer: Electron density score for individual atoms .................. 22
StructureProfiler: Filtering PDB files by properties and experimental criteria .................................................. 23
JAMDAscorer: Gradient-based optimization of protein-ligand poses .................................................. 24
AltLocEnumerator: Dealing with experimental alternate locations in protein structures .................................................. 25
GeoMineDB: Searching geometric patterns in protein-ligand interfaces .................................................. 26
MicroMiner: Searching similar 3D amino acid micro environments .................................................. 27
## Overview: Deprecated Tools

(still available, no updates, risk of OS incompatibilities)

<table>
<thead>
<tr>
<th>Cheminformatics</th>
<th>Structure-Based Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRAISE: Ligand-based virtual</td>
<td>iRAISE: Inverse virtual screening and structure-based</td>
</tr>
<tr>
<td>screening with user-defined partial</td>
<td>target prediction</td>
</tr>
<tr>
<td>shape constraints</td>
<td></td>
</tr>
<tr>
<td>FSEEs: Fragment space exhaustive</td>
<td>NAOMInext: Synthetically feasible fragment growing</td>
</tr>
<tr>
<td>enumeration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pelikan: Searching interaction patterns</td>
</tr>
<tr>
<td></td>
<td>NAOMInova: Intuitive analysis of interaction geometries</td>
</tr>
</tbody>
</table>
Cheminformatics
Mona: Managing compound collections

Cheminformatics in an interactive fashion: Visualize, analyze, filter, and cluster compound collections; Easily spot the differences between molecule sets; Create your own sub-collections by properties, by inclusion and exclusion rules, or just manually. No database installation, no scripting, no pipelining, just an easy-to-use graphical interface.

Main developer: Matthias Hilbig

UNICON: The universal file format converter

Convert between various file formats
including coordinate generation: Convert compound files between SDF, mol2, InChI, SMILES; Import from PDB and mmCIF; Export to InChI Key; Avoid conversion errors with a consistent chemical model; Generate 2D coordinates with high layout quality; Generate low-energy 3D coordinates; Enumerate tautomeric and protomeric forms; Generate low-energy conformational ensembles; Do it all with high speed in just one step with an easy-to-use commandline tool.

Main developers: Kai Sommer, Sascha Urbaczek, and Adrian Kolodzik

Conformator: Generation of conformer ensembles

Generate high-quality conformer ensembles for small molecules:
Read molecular structures from SD and MOL2 files as well as from SMILES and InChI notations;
Sample the conformational space of small molecules (including macrocycles) with a fast, accurate and effective knowledge-based algorithm without PDB bias; Create diverse conformer ensembles for any given ensemble size by automated RMSD thresholding.

Main developers: Nils-Ole Friedrich, Florian Flachsenberg, and Kai Sommer

SMARTSeditor: Interactive and semi-automated graphical creation of SMARTS expressions

**Simply draw your molecular pattern:**
- Visualize and understand SMARTS expressions;
- Draw scaffolds and interactively modify atomic properties;
- Navigate and specify atomic environments graphically;
- Match and validate SMARTS expressions on the fly;
- Calculate contrast patterns separating two molecule sets fully automatically.

**Main developers:** Karen Schomburg, Stefan Bietz, and Lars Wetzer

![ SMARTSeditor Interface ](image)

ACS COMP Emerging Technology Award

SMARTScompareViewer: Analyze and compare molecular patterns

Manage and analyze large filter pattern collections: Test whether one pattern A is included in another pattern B, i.e. A matches whenever B match; Construct and verify a pattern hierarchy; Calculate the similarity between two patterns; Search for similar, more specific or more generic patterns in large pattern collections like Pfizer LINT, SMARTSCyp, SureChEMBL or PAINS; Test for incompatible atom-/bondtype combinations; Visualize the node mapping between similar SMARTS expressions; supports logic and recursive SMARTS expressions, does not support stereochemistry so far.

Main developers: Robert Schmidt, Emanuel Ehmki, Farina Ohm, Andriy Mashychev, and Hans-Christian Ehrlich

REMUS: Shape-based alignments – interactive exploration and screening

Search for 3D similarity within small and medium-sized compound collections: Calculate molecular similarity based on classical Gaussian (colored) shape description; Handle conformational flexibility by a building conformer generator (Conformator) and conformer fine tuning during alignment; Employ a brand-new step-limited BFGS numerical optimizer; Visually explore molecular alignments and perform them semi-automated by RMSD-fit of user-specified atom pairs.

Main developer: Joel Graef
Improved similarity searching and machine learning: Create a topological fingerprint representing all connected subgraphs within specified size ranges (ECFPs consists of only circular substructures); precisely control the level of chemical detail encoded; convert large collections of molecules and derive similarity measurements; integrate easily due to an easy-to-use Python interface.

Main developer: Louis Bellmann
TorsionPatternMiner/TorsionAnalyzer: Create knowledge-based torsion profiles and analyze conformations

TorsionPatternMiner – Creating torsion libraries: All in one tool to populate torsion angle statistics for SMARTS torsion patterns; Cutting edge SMARTS pattern analysis with SMARTScompare; Torsion Libraries created can be used with many conformer generators like Conformator, Omega

TorsionAnalyzer – Analyzing torsions in molecule conformations: Easy to use torsion analysis tool; enables high-throughput torsion analysis (command line), GUI mode to visualize torsions, torsion patterns and their statistics.

Main developer: Patrick Penner, Robert Schmidt, Agnes Meyder
**Phariety: Pharmacophore matching made easy**

**Pharmacophore matching:** Implements a well-established, precise and efficient backtracking algorithm; enables fast and reliable pharmacophore screening with a command-line tool; Handles all standard pharmacophore features including interaction directions and exclusion volumes; Parses pharmacophore models in .json and .ph4 format (MOE), molecules in SMILES and sdf with fully integrated structure and conformer generator.

Main developer: Uschi Dolfus

![Diagram of pharmacophore matching](image)

Check for distance compatibility:

\[ |d_{1,2} - d_{0,1,2}| \leq r_1 + r_2 + \varepsilon \]

\[ + \]

\[ |d_{2,3} - d_{0,2,3}| \leq r_2 + r_3 + \varepsilon \]

Dolfus et al, manuscript in preparation
ReactionViewer: Visualizing chemical reactions with high precision

Visualizing Generic Reaction Patterns: Create easy-to-understand images for generic reaction patterns using SMARTSview technology; Supports Reaction SMILES, Reaction SMARTS, and SMIRKS; Inspired by the IUPAC’s Compendium of Chemical Terminology for reaction equations; Features direct image output in png, svg, and pdf format; Enables single-command conversion of large reaction pattern collections to single pdf documents.

Main developer: Uschi Dolfus, Robert Schmidt

Chemical Space
SpaceLight: Topological similarity searching in large chemical spaces

Explore chemical spaces containing $10^{15+}$ compounds in seconds: Search for structurally similar compounds in chemical spaces far too large for state-of-the-art substructure-driven search methods; Retrieve results similar to a classic fingerprint (e.g. ECFP) search; Detect scaffolds from large building block collections; Create your own large reaction-driven chemical space; use your own laptop and finish search runs within seconds.

Note: The ChemBio Suite SpaceLight version works with public spaces only. SpaceLight for commercial spaces can be licensed exclusively from BioSolveIT GmbH.

Main developer: Louis Bellmann

SpaceMACS: Maximum common substructure searching in large chemical spaces

**Search chemical spaces by sub-structures in seconds:** Search for structurally similar compounds in chemical spaces by maximum common substructure similarity; Retrieve hundred thousands of compounds by maximum common substructures within 10^th's of seconds on a single, standard server; apply a chemically motivated similarity measure finding the closest analog available in chemical spaces.

Note: The ChemBio Suite SpaceMACS version works with public spaces only. SpaceMACS for commercial spaces can be licensed exclusively from BioSolveIT GmbH.

**Main developer:** Robert Schmidt

**SpaceCompare: Analyze, compare, and optimize ultra-large chemical spaces**

**Analyzing chemical spaces:** Calculate the exact overlap of products shared by very large chemical fragment spaces (provided that the overlap is enumerable); Calculate exact physicochemical property distributions for ultra-large chemical spaces (heavy atom count, MW, clogP, H-donor/acceptor counts); generate chemical subspaces with optimized product properties.

Note: SpaceCompare works with public and self-created, and commercial spaces available from BioSolveIT (for customers on request due to special file format).

**Main developer:** Louis Bellmann, Patrick Penner
**Galileo: Genetic algorithm for general purpose searching in chemical space**

**Chemical Space Search:** Searching for molecules in chemical fragment spaces by arbitrary scoring functions; Genetic algorithm directly operating in fragment spaces (so each molecule created is from the space); Combinable with arbitrary external scoring functions via system call; Pharmacophore search based on Phariety already integrated; First search engine for 3D in fragment spaces, however quite resource demanding.

Note: Galileo works with public, self-created and commercial spaces available from BioSolveIT’s download area.

Main developer: Christian Meyenburg, Uschi Dolfus

Structure-Based Design
ASCONA/SIENA: Active site search and alignment

Create aligned active site ensembles: Search for sequence-similar active sites in the whole PDB; control sequence identity in the active site; find active sites in homodimeric and multimeric structures; consider structural variation upon search and alignment; create a multiple structural alignment of active sites; reduce active site collections to small ensembles with high structural variance.

Main developer: Stefan Bietz

EDIAscorer: Electron density score for individual atoms

**Estimate the experimental uncertainty of individual atoms in protein structures:** Create a map showing the support of each individual atom in a crystallographic structure; Automatically compare a structure model to an electron density; calculate an EDIAm value for each molecule emphasizing individual atoms with low EDIA values; get error analysis information for atoms with low electron density support values; Fully-automated reliability prefiltering of large structure collections.

Main developers: Agnes Meyder, Eva Nittinger, and Florian Flachsenberg

StructureProfiler: Filtering PDB files by properties and experimental criteria

Automatically profile X-ray protein structures with the most frequently applied quality selection criteria in use:
All-in-one tool for PDB file selection; Fully configurable via INI parameter files;
Preconfigured filters highly similar to Astex, Iridium, Pratinium and more;
Supports PDB and mmCIF format with optional electron density files; Includes EDIA and other electron density quality criteria; Includes torsion angle and clash checks; Includes lots of ligand descriptor filters like Lipinski Ro5; Enables fast and efficient updating of benchmark datasets with new PDB files.

Main developer: Agnes Meyder and Stefanie Kampen

JAMDAscorer: Gradient-based optimization of protein-ligand poses

Post-optimize protein-ligand complexes into precise local minima with the new JAMDA scoring function: JAMDAscorer is a new empirical scoring function made for use with gradient-based optimizers. Combined with a newly developed optimizer, LSL-BFGS, fast convergence, locality and precise detection of local minima can be guaranteed. JAMDAscorer considers hydrogen bond geometries, hydrophobic contacts, clashes, and torsion angles therefore balancing the most important terms in protein-ligand scoring functions.

Main developer: Florian Flachsenberg, Agnes Meyder, Kai Sommer, Patrick Penner

AltLocEnumerator: Dealing with experimental alternate locations in protein structures

**Modeling structural uncertainties:**
Automatic handling of alternate locations (AltLocs) in X-ray protein structures; Exploring protein flexibility through experimentally validated AltLocs; Graph-based algorithm to enumerate all valid structure conformations that do not clash or introduce chain breaks; Generating all AltLoc-induced structural variants and writing them to separate PDB files.

Main developer: Jochen Sieg, Torben Gutermuth, Tim Stohn
GeoMineDB: Searching geometric patterns in protein-ligand interfaces

**Searching protein structures:** rapid searching for spatial interaction patterns in large collections of protein-ligand complexes and binding pockets; Fully-automated database build-up from PDB and CIF; Built on free database systems SQLite and PostgreSQL; Supports radius-based pockets (around ligands) and predicted pockets (based on DoGSite); Flexible query management based on XML; Outputs statistics about matching hit structures superimposed on query in JSON format (but no graphical frontend).

Main developer: Joel Graef, Konrad Diedrich, Therese Inhester

MicroMiner: Searching similar 3D amino acid micro environments

Analyzing protein structures: Searches amino acid environments with local sequence and structure similarity; Structural mutation search in the entire PDB, your inhouse protein collection or AF2 database; Explore the mutation landscape of proteins with experimental or predicted structures; Applicable in single domains or even on protein-protein or protein-ligand interfaces; Several filter options to simplify downstream analysis.

Main developer: Jochen Sieg, Stefan Biets

Schöning-Stierand, K. et al., Nucleic Acids Research (2022), 50, W611-615
Sieg et al, manuscript in preparation
Deprecated Tools

These software tools remain available for download, however they will not be updated anymore. The reasons could be manyfold, in most cases: the dependence on deprecated external libraries, the replacement by newer software tools already available or on their way, low usage, extremely high maintenance effort.
**mRAISE: Ligand-based virtual screening**

**with user-defined partial shape constraints**

**Define your own constraints**: Fast ligand-based virtual screening of large compound collections using the RAISE index technology; Screen compound libraries either with certain percentages of shape similarity, with complex derived constraints or with your own manual selection of partial shape constraints; Find similar structures and generate accurate alignments.

Main developer: *Matthias von Behren*

FSees: Fragment space exhaustive enumeration system

Generating large compound libraries:
Enumerates a so-called fragment space under physico-chemical constraints; Ensures completeness and uniqueness among the created compounds; Allows to limit more than ten physico-chemical properties from molecular weight via structural features to logP; Allows to combine properties to important filters (e.g. lead-like or drug-like).

The basis for this method is a fragment space, a fragment-oriented model of combinatorial chemistry space. It can be constructed from retrosynthetic rules or synthetic chemical reactions thus yielding molecules that have a high likelihood of synthetic accessibility.

Main developer: Florian Lauck

iRAISE: Inverse virtual screening and structure-based target prediction

Search through thousands of proteins for potential binding events: Find potential binding sites for a small molecule to predict off-target effects or polypharmacology; Prepare and use a relational protein structure database for fast access to structures; Prepare proteins in a fully-automated fashion; Score potential binding sites with an innovative multi-step scoring procedure.

Main developer: Karen Schomburg

NAOMInext: Synthetically feasible fragment growing

**Forward synthesis planning and high throughput growing:** Grow ligands into active sites based on crystallized or docked fragments; Create a target focused library of your in house or vendor catalogue of fragments; Use 58 published incorporated robust reaction rules in SMIRKS format; Provide your own reactions which are checked for validity and consistency prior usage; Visually inspect reaction vectors; Add residue constraints to guide the growing process into a specific subpocket.

Main developer: Kai Sommer

Pelikan: Searching interaction patterns

Find all occurences of your interaction patterns: Create a database of your collection of protein-ligand complexes or search in the whole PDB; Define your spatial constraints in 3D starting from a binding site of interest or from scratch; Combine 3D search with textual and numerical constraints for a large number of ligand, protein and complex properties; Rapidely find all occurences of interaction patterns; browse and refine your results in a 3D viewer.

Precompiled Pelikan Databases available for PDB, scPDB, and several target classes.

Main developer: Therese Inhester

NAOMInova: Intuitive analysis of interaction geometries

Determine optimal interaction geometries for non-covalent interactions: Create a database of your collection of protein-ligand complexes; Query the database for atoms found in the vicinity of functional groups; Visualize all detected atoms or density clouds; Filter and inspect all detected atoms around the functional group (incl. electron density support values, EDIA); Click on any data point to inspect the original protein-ligand complex and measure arbitrary distances and angle distributions; add your own functional group descriptions.

Main developers: Therese Inhester and Eva Nittinger

NAOMI ChemBio Suite – FAQ

**Will older software become part of the collection?** We are currently in the process of integrating successful, older software tools into our collection e.g. the TorsionAnalyzer and TorsionChecker jointly developed with Roche and DoGSite and DoGSiteScorer developed with Merck and BioSolveIT. Software based on FlexX and FTrees cannot be integrated due to legal reasons.

**Will you charge a substantially higher fee once the collection is distributed?** This is not our plan. We do this as an academic group and calculate prices in order to build and maintain a fund helping us to keep the software alive. This is not a for-profit business.

**Will software tools disappear from the collection?** This might happen in case we are not able to keep it running. Reasons might be that 3rd-party libraries are not available anymore, substantial problems occur with OS updates or we find a substantial scientific flaw we are not able to fix (which never happened so far).

**What level of support can you offer?** We help wherever we can, however we are not a software company. Therefore, we cannot offer guaranteed response times and we cannot promise that we fix all problems and bugs which might come up in the future. We offer the software on an as-is level, this is the reason why you can test everything before you make a decision.

**Will the collection grow over time?** Yes, this is the plan. We want to add all software resulting from our research without exclusivity constraints (which sometimes happens in case of 3rd party funding). We have a lot of cool stuff in the pipeline in all three research areas, namely cheminformatics, structure-based molecular design, and visual analytics in chemistry.
Why is our software not available for free?

• We are strong believers in the policy that academic software should be free for academic research and evaluation purposes, but should not be free for commercial applications. You might not agree, but here are our arguments:

• PhD students are not paid for creating software. In fact, they are most frequently not even paid for doing research. The salary even in a rich country like Germany is much below that of a software developer. License fees received by us are exclusively used to support academic research in the AMD group, i.e.:
  • To improve the salary of PhD students
  • To give travel grants to PhD students for participation on international conferences

• Creating and licensing software is a valid business. Thousands of computational chemists and computer scientists around the world work for enterprises offering software and related services. Free software destroys, or at least harms, this business.

• Academic software should be designed and developed in a sustainable fashion for several reasons. First of all, a scientific result should be reproducible. Second, the software production should be economic. Third, and most important, software development during a PhD project is an important training on the job. Producing sustainable software is, however, not for free. Substantial efforts are necessary to build up the required infrastructure for reviewing, testing and cross-compiling, and achieve the necessary level of code quality and documentation. License fees help to support this endeavor.
Licence Model

- Tools are distributed as installable software packages for various Linux platforms, Windows and Mac OS.
- Single registration, access to all tools via a unified web service.
- Continuously extended by new software tools
- Regularly updated
- All tools available for free for academic use and evaluation purposes.

License model for non-academic use:

- Single license for all tools
- Unlimited site license (unlimited number of users and number of CPUs)
- Annual license fee depending on company size:

<table>
<thead>
<tr>
<th>License Type</th>
<th># of employees</th>
<th>Amount (in k€)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very small</td>
<td>1-9</td>
<td>2</td>
</tr>
<tr>
<td>Small</td>
<td>10-99</td>
<td>4</td>
</tr>
<tr>
<td>Medium</td>
<td>100-999</td>
<td>8</td>
</tr>
<tr>
<td>Large (single site)</td>
<td>≥ 1000</td>
<td>16</td>
</tr>
<tr>
<td>Large (all sites)</td>
<td>≥ 1000</td>
<td>25</td>
</tr>
</tbody>
</table>

- Discounts for project partners

All incoming fees are exclusively used to support PhD students and Postdocs participating in the development of the NAOMI ChemBio Suite.
Contact

Prof. Dr. Matthias Rarey [matthias.rarey@uni-hamburg.de]

Universität Hamburg
ZBH – Center for Bioinformatics [https://uhh.de/zbh]
Research Group for Computational Molecular Design [https://uhh.de/amd]

Software Server: https://uhh.de/naomi
Web Server: https://proteins.plus
https://smarts.plus

Source code libs: https://github.com/rareylab