JOURNAL OF CHEMICAL INFORMATION AND MODELING

Application Note

pubs.acs.org/jcim

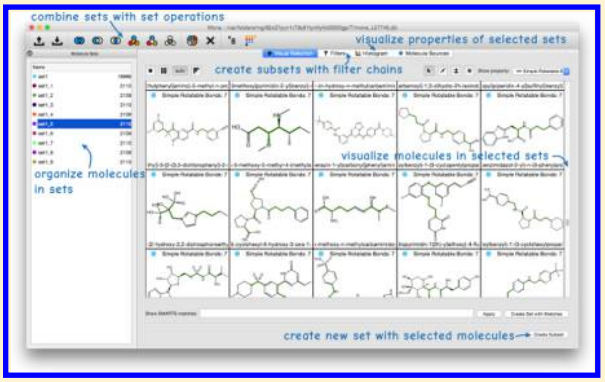
MONA 2: A Light Cheminformatics Platform for Interactive Compound Library Processing

Matthias Hilbig and Matthias Rarey*

University of Hamburg, Center for Bioinformatics, Research Group for Computational Molecular Design, Bundesstraße 43, 20146 Hamburg, Germany

Supporting Information

ABSTRACT: Because of the availability of large compound collections on the Web, elementary cheminformatics tasks such as chemical library browsing, analyzing, filtering, or unifying have become widespread in the life science community. Furthermore, the high performance of desktop hardware allows an interactive, problem-driven approach to these tasks, avoiding rigid processing scripts and workflows. Here, we present MONA 2, which is the second major release of our cheminformatics desktop application addressing this need. Using MONA requires neither complex database setups nor expert knowledge of cheminformatics. A new molecular set concept purely based on structural entities rather than individual compounds has allowed the development of an intuitive user interface. Based on a chemically precise, high-performance software library, typical tasks on chemical libraries with up to one million compounds can be performed mostly interactively. This paper describes the functionality of MONA, its fundamental concepts, and a collection of application scenarios ranging from file conversion, compound library curation, and management to the post-processing of large-scale experiments.



INTRODUCTION

For many years, the management of large compound collections via chemical databases was primarily an industrial endeavor. Complex databases that are able to handle millions of compounds accessible by many users and software systems form the core infrastructure for pharmaceutical and chemical research. Since dealing with this data often requires many processing steps, workflow systems such as BIOVIA Pipeline Pilot,[1] KNIME,[2] and AZOrange[3] have emerged. Employing graphical programming paradigms, they enable the gathering and processing of workflows in a convenient fashion. In the past decade, more and more large-scale chemical datasets became freely available via the Internet.[4] Compound vendors have made their catalogs available electronically, enabling the construction of large compound collections, such as ZINC.[5] Academic screening campaigns and the collection of bioactivity data from literature have created large data pools such as PubChem[6] and ChEMBL,[7] enriching compound data with biological knowledge. Databases that collect life science data for compounds, such as DrugBank[8] or BRENDA,[9] enable completely new ways of data mining. With all these data sources and an increasing interest of the academic community to analyze and employ them in research, the need for cheminformatics tools has increased in a wider scientific community. Compared to the industrial use of chemical databases, the way in which these data sources are initially analyzed has changed. First of all, the compound collections are less static. Depending on the research question, different data collections come into play, must be browsed, filtered, and put into relation. Second, the processes and workflows are usually more dynamic and require a high level of processing flexibility. Instead of a few data analysis scenarios, such as "compound acquisition", "ADME prefiltering", and "diverse or focused library selection" as they appear in industrial pharmaceutical research, an infinite number of possible scenarios, analyzing multiple data sources in the context of specific scientific questions, emerge. Third, the cheminformatics user community prospers from experts in the field, whereas life scientists only occasionally use these tools.

The scientific community reacted to these needs in two ways. On the one hand, the data providers on the Internet enriched their web portals with cheminformatics functionality. The ZINC portal, for example, started as a download service for compound catalogs. In the most recent version, it not only has many preprocessed database subsamples for download, it also allows one to interactively browse the catalog by many query types. On the other hand, interactive standalone tools for chemical data analysis appeared. Screening Assistant 2 (SA2)[10] and Data Warrior[11] are two recent examples for such tools. Both are very comprehensive cheminformatics tools for library processing and data analysis with a huge functionality pool.

Received: May 19, 2015
Published: September 21, 2015

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078

Consequently, they address the needs of experienced cheminformaticians. Another interactive analysis tool is Scaffold Hunter,[12] which is completely focused on the visual exploration of chemical space. However, it currently does not support the management of subcollections of small molecules. Lastly, StarDrop[13] is a software system for the analysis of experimental SAR data. Although developed with a slightly different focus, StarDrop allows molecules to be placed on on "cards" that can be clustered and piled, similar to the set concept in MONA.

In order to address the necessity for a generic, simple, and interactive compound browser and manager, we developed MONA. In contrast to workflow engines, the design of MONA allows an intuitive interactive organization and exploration of chemical small molecule datasets on current desktop computer hardware. For the sake of a wide functionality range, MONA focuses completely on this task, which makes it easy to use. The guiding design objectives behind MONA can be summarized as follows:

- *Simplicity*: MONA must be easy to use and easy to install on a wide variety of hardware, such as desktop PCs and laptops. An intuitive user interface should keep the learning barrier low and also support occasional use.
- *Chemical precision*: MONA should circumvent classical pitfalls of cheminformatics by precise and automatic processing of compounds from various file formats, including the handling of protonation, tautomerism, and stereo isomerism.
- *Completeness*: MONA should address a broad spectrum of application scenarios, from initial chemical library processing to the post-processing of large-scale experimental data.
- *Interactivity*: MONA should enable interactive work with data collections. Decisions should be able to be made and reversed based on available data and results of previous analysis steps, employing efficient visualization aids.

In the following, we present a rough technological overview showing how these design objectives have been addressed. Furthermore, we survey the functionality provided by MONA. Since most of these design objectives are difficult to validate rationally, we give a series of application scenarios and show how to solve them with MONA. A more-detailed description of the corresponding workflows is given in the Supporting Information. In order to give an impression of MONA's processing performance, we exemplary summarize computing times and memory requirements.

METHODS: MONA TECHNOLOGY

MONA is a standalone desktop application and is available for the Windows, Linux, and OS X operating systems. A first version of MONA[14] was published in 2013; therefore, only a brief summary of the basic technology is given. Several new methods and features have found their way into MONA 2 and will be described in more detail below.

MONA is based on the cheminformatics library NAOMI. NAOMI provides a consistent internal molecule structure, which is used to convert between different file formats.[15,16] NAOMI is handling the standard cases of stereo isomerism, including the explicit modeling of the unknown state. In addition, the library provides methods to generate canonical tautomers and protonation states, which were extensively validated.[17] For ring perception, the unique ring family concept[18] is applied as an extension of the standard Smallest-Set-of-Smallest-Rings approach. This results in the molecular properties "Rings", which counts the number of relevant cycles, and "Ring Families", which contains the number of unique ring families.

Internally, MONA employs an SQLite[19] database as a backend to hold all molecule data.[20] This is mainly for reasons of simplicity, because SQLite databases are completely contained in one file and do not need any setup steps. Overall, MONA does not require any additional setup steps and can be deployed directly by starting the provided binary. The fundamental concept behind MONA is to consider the molecular structure as the primary key, rather than any type of molecule ID resulting from an input file. In order to keep the link to the original data, a molecule might have multiple so-called "instances", reflecting duplicates in files. Prior to loading data, the user defines the level at which molecules are considered identical. The requested setting modifies the molecules during database import. Requesting to ignore tautomerism creates normalized tautomers[17] for all input molecules. Selecting to ignore stereoisomerism resets all stereo descriptors to "unknown". Similarly, any charge differences between molecules are ignored on request by creating the default protonations for the molecules. This identity criterion follows the basic understanding of chemists and makes operations on sets of molecules intuitively understandable. MONA's principal form of organization is a classical set concept. New sets can be created in many different ways (e.g., by importing files, by filtering, or by combining other sets with classical set operations such as intersection, union, or subtraction). Afterward, sets can be exported again into files. MONA supports the common cheminformatics file formats SDF, MOL2, and SMILES, as well as the import of small molecules from Protein Data Bank (PDB) files. Great care was taken to make all conversions between different file formats consistent.[15] For further details on how the different operations were implemented, see ref 14. An additional key feature of MONA is its fast structure depiction engine. MONA creates two-dimensional (2D) structure diagrams of molecules on the fly, exploiting multithreading. Therefore, even large compound collections can be visually browsed instantaneously.

The new version of MONA was improved in three major areas. First, arbitrary properties may now be added to any set of molecules. Second, clustering and a new cluster visualization mode further support the initial analysis of compound collections. Third, a novel, highly robust 2D alignment routine supports the visual detection of common structural patterns.

PROPERTY HANDLING

In the previous version of MONA, 29 different physicochemical properties were calculated for each molecule and saved in the database. The database scheme of the new version allows the annotation of any number of properties to each molecule. MONA calculates only a handful of the most-often-used properties such as heavy atom, donor, acceptor and stereo-center count, molecular weight, and logP estimates when loading molecules. All other properties may be calculated on demand. MONA discerns between molecule and instance properties. Molecule properties such as the molecular weight can be added directly to molecules in sets. Instance properties, such as the name of an instance, must be converted to molecule properties before they can be added. MONA provides multiple methods to convert instance properties: "First Occurrence" takes the first instance property and "All Occurrences" concatenates all instance properties. The last three methods—"Minimum", "Average", and "Maximum"—only work on

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078

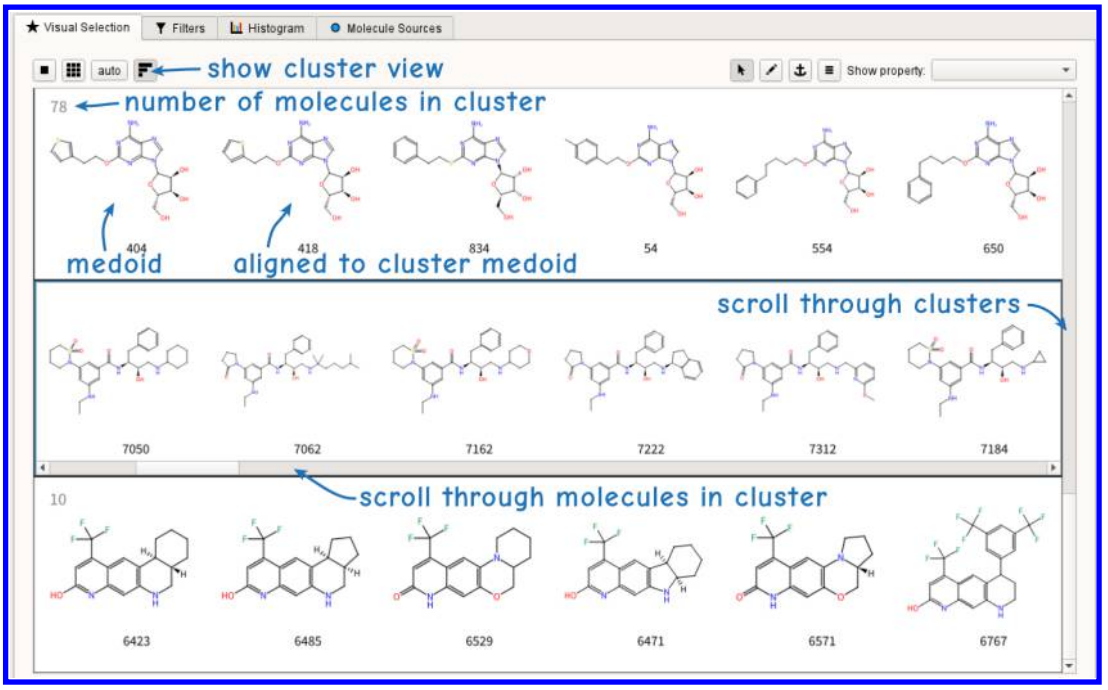


Figure 1. Cluster view in MONA shows the molecules in a cluster horizontally and the different clusters vertically. Molecules within one cluster are structurally aligned to the cluster medoid shown first.

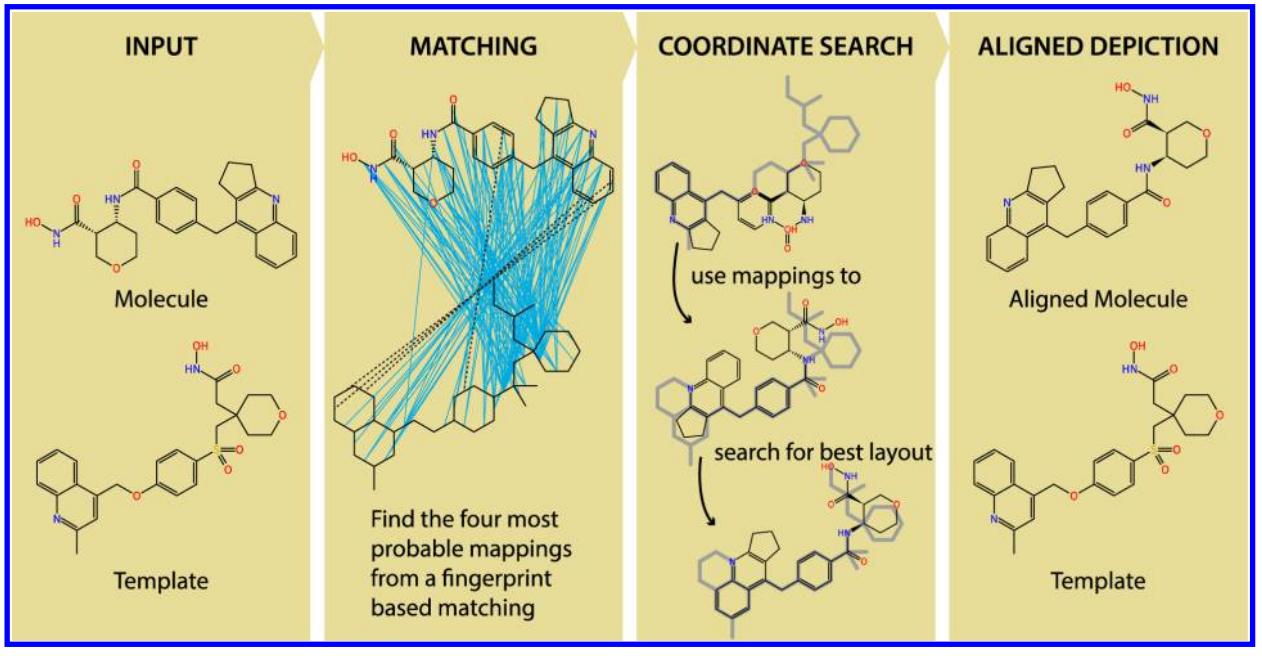


Figure 2. An aligned depiction of a molecule to its template is calculated in two phases: matching and coordinate search (also see the main text). The bond-pair mappings in the matching phase are shown by blue lines. The four most probable mappings (dashed black lines) are used in the next stage. During coordinate search, the created layouts are scored by the percentage of overlay between the molecule and its template.

numeric values and respectively return the minimum, average, and maximum value of all instance properties. The increased flexibility is used for three additional classes of properties:

(1) *External properties*, which may come from any external source: Currently, all properties annotated in SD files and information found in the remark sections of PDB files are supported.

(2) *Internal properties* that are less frequently used: Several descriptors related to molecular structure, such as atom and ring counts, as well as chemical descriptors, such as the topological polar surface area, can be added. Furthermore, identifiers representing the canonical tautomeric and protomeric form can be calculated and assigned.

(3) *Database properties* are those properties that are dependent on all currently loaded molecule sources in MONA. Namely, the duplicate properties determine the number of instances currently contained in the database for each molecule. Similarly, the source entry number and source filename contain the entry number and name of molecule sources for all instances of a molecule.

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078

Internal and external properties are handled equally and can be used for filtering, sorting and clustering of sets. The only exceptions are the tautomer and protomer identifiers, which cannot be used for filtering.

CLUSTERING AND CLUSTER VISUALIZATION

To group compounds into new sets automatically, MONA contains a new clustering functionality (see Figure 1).

Molecules may either be clustered by a property or by molecular similarity. In case of a string or a whole number property, the clusters are created by grouping molecules with equal values. In case of real valued properties, the values are binned into appropriately chosen ranges before grouping. Clustering by molecular similarity employs ECFP6 fingerprints[21] together with a *k*-medoids algorithm.[22] Different variations of fingerprints (ECFP, FCFP, Torsion fingerprint, and a purely topologically based ECFP) and similarity measures[23] (Tanimoto, Cosine, Hamming, Euclidean, and Dice) can be specified. Note that the ECFP and FCFP implementation differs in detail (hash functions, initial atom identifiers) from the original. The *k*-medoid algorithm assigns the molecules to a chosen number of clusters. The algorithm stops iterating if the clusters no longer change or after a maximum number of iterations has been reached. Note that (i) clustering, by itself, is a highly complex problem, and (ii) MONA provides only basic functionality. The *k*-medoids algorithm runs in quadratic time, making similarity clustering only feasible for small sets of up to 15 000 compounds.

A new cluster view, as alternative visualization to the tabular view of all molecules in a set, allows faster browsing. In the cluster view one cluster of molecules is depicted horizontally and the different clusters vertically. Each cluster can be scrolled individually in the horizontal direction or converted into a set for further visualization.

ALIGNED DEPICTION

The structure depiction engine was enhanced to allow the aligned layout of molecule depictions (see Figure 2). An important difference to existing alignment functionality in other toolkits[24−26] is that the alignment is done neither by post-rotating coordinates to align the largest found common subgraph nor by copying 2D coordinates from a template molecule. Our new approach calculates an optimal individual mapping of bonds from the molecule to be drawn to the template using radial fingerprints. Then, four pairs of bonds are picked based on the highest atomic similarity values of the neighboring atoms. Each of these bond pairs are, in turn, superimposed. Eventually, the structure diagram is generated with additional score values estimating the percentage of atoms of the molecule superimposed with the template. In contrast to MCS-based alignments, this method is robust against small changes in the common region of the molecule. Furthermore, the avoidance of coordinate copying leads to many more collision-free depictions. The aligned layout is used in three different places in MONA. First, anchoring an arbitrary molecule in the tabular view aligns all other molecule to the anchored molecule. Second, all members of a cluster are aligned to the cluster medoid in the cluster view; naturally, this is more visible if the cluster were created by molecular similarity. Lastly, the search for SMARTS patterns uses a depiction of this SMARTS to align the first matching subgraph in each molecule.

RESULTS: APPLICATION SCENARIOS AND PERFORMANCE

In the following, a series of application scenarios for MONA—organized in three areas—are described:

(1) The first area covers efficient handling of files with small molecules. The scenarios in this section exemplarily show basic tasks, such as file format conversion or the selection of specific molecules from files. Because of the increasing size of compound collections, quickly getting an overview of the content and creating application-specific subcollections on the fly become essential elementary tasks today.

(2) The second area involves the management and curation of large data collections. Depending on the way compound collections are created, utilizing them causes the need to fix deficiencies and remove inconsistencies. Since datasets evolve over time, keeping track of changes is very important for data mining approaches, as well as cheminformatics tool development.

(3) Showing differences and similarities between various compound collections is the third area for which MONA is especially suited. The software enables (the user to make) cross connections between datasets by combining information from multiple sources. With these capabilities, MONA allows one to perform elementary data mining steps necessary to prepare data collections for sophisticated machine learning approaches.

Here, the steps necessary to solve each of the tasks are summarized only briefly. Detailed descriptions can be found in the Supporting Information. Note that MONA is not bound to these or any specific list of workflows. All elementary steps described can be freely combined in an arbitrary order, which gives MONA the flexibility for interactive and intuition-driven work.

USING MONA AS A CONVERSION AND SELECTION TOOL

LigandExpo[27,28] is the collection of all small molecules appearing in the Protein Data Bank (PDB),[29] which currently contains 18 986 molecules. Because of the specification of the PDB format, extracting small molecules is challenging, making LigandExpo an important additional information source. Here, we will use this collection to demonstrate elementary cheminformatics tasks.

Task 1: Which Molecules Are in My File? The original LigandExpo file contains more than 500 000 entries; however, only ∼19 000 of those are unique molecules. Loading LigandExpo into MONA results in this unique collection of molecules, which can be directly written to the file. MONA is able to provide information about duplicates. The duplicate count can be added to the list of properties, enabling one to extract molecules with a certain frequency from the dataset. Furthermore, a list of up to 100 occurrences of a molecule in the database, so-called "instances", can be browsed. Histograms, which can be created for properties either contained in the file or calculated by MONA, give a statistical overview of the dataset content. The structure depiction engine allows instant browsing of collections with arbitrary size in a structure table view sorted by any of the available properties.

Task 2: How Can I Convert a Large Table of SMILES into Multiple SD Files? Many computationally intense cheminformatics tools such as virtual screening parallelize by dividing input data into evenly distributed sets. LigandExpo might be used to evaluate the performance of such a tool.

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078

Splitting data collections come in different flavors related to the order of compounds. MONA supports three different scenarios. First, the original order can just be maintained; second, the order can be randomized, which is extremely important for the creation of training and test data in machine learning. Eventually, any specific order can be specified before splitting to construct, for example, sets of compounds with increasing molecular weight. MONA allows splitting into (i) a fixed number of sets, (ii) chunks with a fixed number of molecules, or (iii) two sets with unbalanced size. Afterward, these sets can be converted to multiple SD files by exporting the created sets. Before export, the user can choose the molecule sources used for recreating instances in the output files. That means if a molecule in MONA originated from both a SMILES instance and an SD File instance, selecting only the SD File source exports the molecule with the conformation and name of the SD File instance. Similarly, if only the SMILES source is selected, the SD file entry is created without coordinates and with the name of the SMILES instance.

Task 3: How To Filter Out Compounds with Undesired Properties? There are many reasons for the removal of specific compounds from data collections. Let us assume that a user wants to inspect all small molecules from the PDB to mine for interaction patterns. He might come up with the following constraints: the relative molecular weight should be between 200 and 400, and all molecules should contain at least one halogen and a ketone group, as well as at least two hydrogen bond donors. MONA allows one to create arbitrary filter chains and store them for later reuse. However, a much better way to run through this task is by using MONA's set concept. LigandExpo contains 8557 compounds in the given MW range, 4383 halogenated compounds, 3859 molecules with a ketone group, and 13 881 molecules with two donors. Inspecting these temporary sets allows interactively adapting the chosen parameters for each filter step. The final intersection of these subsets results in 217 molecules.

Task 4: How Can I Filter by a Chemical Pattern? Let us assume experiments reveal the importance of a specific core fragment—a pteridine ring system, for example. When loading the corresponding SD file, a SMARTS pattern of the entire molecule can be created. Opening it with the included SMARTSeditor allows one to cut away all unnecessary parts and create a pattern that only matches the core fragment. Alternatively, the pteridine pattern can be drawn from scratch with the SMARTSeditor or simply by entering the corresponding SMILES string into it. Afterward, a new subset can be created from LigandExpo with all molecules matching this pattern. To easily compare the molecules in the newly created set, the first molecule is anchored and all core fragments are highlighted and depicted in the same orientation.

Task 5: How Do I Extract Small Molecules from a PDB File? Reading PDB files is an error-prone task, because the molecules must be recreated purely from three-dimensional (3D) coordinates. MONA is built on the NAOMI library, which contains a robust method to perceive small molecules from PDB files.[16]

Opening a PDB file in MONA builds all small molecules and creates a molecule set from them. Typical annotations from the PDB file, such as the ID, the resolution, or the EC numbers, are shown in the Instance View and are ready to be added as molecule properties.

USING MONA TO CURATE AND MANAGE DATABASES

Manually curating databases usually leads to increasing numbers of inconsistencies and errors as the database becomes larger. MONA allows one to work comfortably with datasets consisting up to 1 million molecules. Operations on sets of molecules, clustering, and property histogram plots enable the curation and management of such databases.

Task 6: Finding and Fixing Inconsistencies in Cheminformatics Databases. Loading the DrugBank dataset results in a set with 6613 molecules. With the Property Dialog, the Duplicates and Tautomer Identifier properties are added to this set. Sorting the set by the Duplicates Property and scrolling to the end shows all molecules that occur more than once in the file. In the DrugBank case, there are 110 molecules with duplicates. Clustering the set by the Tautomer property and viewing the result in the Cluster view shows one molecule existing in three different tautomeric forms and 10 molecules with two different forms in the dataset. Duplicates are removed automatically when writing the set to a file with default options. Tautomers can be removed by reimporting the file while ignoring tautomerism.

Vendors of small molecules may create up to 20 000 new compounds each month.[30] Keeping an overview of which molecules were added and which were removed becomes difficult without the help of tools. MONA allows one to find and inspect the changed molecules efficiently.

Task 7: Juggling Multiple Databases. A vendor sends a new catalog that must be incorporated into an in-house database. After loading the vendor catalog with MONA, all new compounds can be determined by calculating the set difference between the catalog set and the in-house database set. Substructure searches can now be used to exclude molecules with undesired properties. A famous example is the collection of 482 PAINS[31,32] patterns enriching known frequent hitters. MONA allows one to load a list of SMARTS patterns into a single filter, making the exclusion of PAINS-related compounds a very simple task. The remaining new molecules can now be clustered by similarity to see the classes of new molecules that would be added to the database. After manually removing some additional unwanted molecules, the result can finally be united with the in-house database.

USING MONA TO POST-PROCESS RESULTS OF EXPERIMENTS

After performing any type of chemical experiment, the results must be processed to generate new insights. Using set operations, cluster view and histogram plots provided by MONA can help in this endeavor. Combining information from different datasets such as PubChem, ChEMBL, or DrugBank is difficult, because they often do not have a similar structure. MONA allows one to use the common element of these datasets, namely, the molecules in the database, to connect information from different sources.

Task 8: Finding Information for Molecules. Two sets are created in MONA: the first set, with all molecules from the LigandExpo, and the second set, with all DrugBank molecules loaded from the SD file. How many molecules do the two datasets have in common? The intersection between both sets consists of 3770 molecules. To create an SD file with combined information on both databases, the DRUGBANK_ID and DRUG_GROUPS properties from the DrugBank and the

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078

Molecule Name property from the LigandExpo are added to the intersected set. MONA now allows the intersected set to be filtered and sorted by these new properties and exported as SD file combining information from both datasets.

Task 9: Preparing and Analyzing Virtual Screening Datasets. One ubiquitous task in cheminformatics is the preparation of virtual screening datasets. All virtual screening tools have the common property that they start with a huge number of molecules to create a small set of possible screening candidates. MONA allows one to work comfortably with sets of up to 1 million molecules. Much larger sets should be prefiltered by an external program before using MONA.

Starting with molecules from different ZINC vendor sources, a large set in MONA is created, consisting of the union of all vendor sets. Potential duplicates are already removed and the set is ready to be further inspected. Viewing the distribution of the molecular weight or other properties in this set can create a first impression. Any molecules not fitting into the set (typically molecules that are not druglike) are now removed by filtering operations. The resulting set is written to disk and used as input for a virtual screening pipeline. After performing the virtual screening, the top-ranked 1000 molecules must be inspected to determine which molecules should be ordered. Loading and clustering them by similarity allows one to examine the different classes of the molecules in the cluster view. The molecules with the highest virtual screening scores are chosen for each cluster and put together into the set of final candidates. Using properties can be useful in the decision where these final compounds should be bought: In MONA, the source filename property contains the filename from which each instance was originally loaded. This property can be used to discern the vendor availability for each molecule by adding all occurrences of this instance property to the molecules in the final set. This set may now be filtered for all molecules that originated from vendor A. In addition, if price information for instances are provided as additional properties in the SD files, the minimum, average, or maximum price can be added as a molecule property. Afterward, sorting the available compounds from vendor A by their average price allows one to choose the least-expensive compounds from this vendor.

PERFORMANCE OF MONA OPERATIONS

Most operations in MONA run interactively on small- and medium-sized sets or in reasonable time, even on large sets. The runtime for typical MONA operations is exemplarily shown for three different sets in Table 1: this consisted of one small set, with 18 986 molecules (LigandExpo); a medium-sized set, with 148 216 molecules (Enamine Building Blocks catalog from ZINC[33]); and a large set, with 1 172 433 molecules (all DUD-E decoys[34] combined).

All three sets were imported from multiple SMILES files into MONA and exported again into a single SD file. Four distinct types of filters were timed. The easiest possible property filter, which matches all molecules of the set with MW 200−400. Second, a more elaborate property filter, which employs tolerances (Rule-of-Five[35] removes potentially orally non-bioavailable molecules, which violate at least two of the four criteria: MW 0−500, acceptors 0−10, donors 0−5, and logP < 5). The final two SMARTS filters were described as follows: one had a single pattern matching all molecules containing phenyl rings, and the other one excluded all molecules matching any of the 482 PAINS SMARTS expressions. While property filters are completed within a matter of a few minutes, even with the largest set, SMARTS filters take much longer (in the case of PAINS up to multiple days). Note that MONA currently does not have an index to accelerate SMARTS searching.

Table 1. Durations for Typical MONA Operations Were Measured with Three Different-Sized Sets[a]

	Small Set	Medium Set	Large Set
molecules in set	18 986	148 216	1 172 433
import from multiple SMILES files	14.4 s	51.6 s	9:13 min
export to SD file	16.8 s	1:46 min	17:55 min
simple property filter	0.1 s	1.3 s	10.7 s
druglike filter	1.2 s	10.9 s	1:29 min
phenyl SMARTS filter	7.6 s	37.2 s	9:08 min
482 PAINS SMARTS filter	33:39 min	1:42 h	42:26 h
cluster by property	0.06 s	0.4 s	3.6 s
cluster by similarity	3:22 min	16:24 min	24:43 h
set union	0.07 s	0.6 s	4.8 s
set difference	0.2 s	1.3 s	10.7 s
set intersection	0.3 s	2.5 s	21.9 s
split into two sets	0.1 s	1.1 s	8.9 s

[a]Using an Intel Core i7-4770 with 3.4 GHz, 16 GB of RAM and an SSD. For all operations with <10 min, the average value of five independent runs is shown.

MONA contains two separate algorithms to cluster molecule sets. First, clustering properties employ bins to create clusters of molecules with similar property values. This takes linear time in the number of molecules and runs within <5 s for all sets. Second, clustering by similarity calculates fingerprints for all molecules, employing the *k*-medoid algorithm, which requires quadratic time. This takes ∼1 min for the small set and up to 25 h for the largest set.

Set operations are used in MONA to combine, intersect, and subtract sets from each other. Their runtime is heavily dependent on the size of the input and the output set. Therefore, the different operations were tested as follows.

Starting with two equal sets created from the input, union and intersection used both of them as input and produced a same-sized set as a result. Difference was tested between the input set and the first half of the input set, resulting in all molecules from the second half of the set.

Set operations are generally instantaneous for small sets and require <3 s for medium-sized sets and <30 s for large sets.

In conclusion, importing and exporting molecules from the database are the most time-consuming tasks. Working with already-imported sets is fast, filtering them with simple filters and applying set operations on them is instantaneous for small sets and fast enough (<2 min) for large sets. Similarity clustering and multiple SMARTS matching remain the only operations currently not supporting interactive use. As all set operations run asynchronously, MONA can be fully used while similarity clustering and SMARTS filtering tasks are running in the background.

CONCLUSIONS

MONA is a software application that supports an interactive and exploration-driven workflow for solving the most ubiquitous cheminformatics tasks. The introduction of arbitrary molecule properties leads to many new use cases, complementing the basic functionality introduced in the first version of MONA. The added similarity clustering and structure depiction

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078

alignment help in detecting similarities and differences between molecules on a small scale while set operations in MONA already provide these comparisons on a large scale. At the same time, the UI remained simple and is comprehensible for nonexperts, supporting a case-by-case workflow, in which the exact order of operations is not known beforehand. All operations in MONA run asynchronously, which means the UI is fully usable while long running tasks work in the background. Nevertheless, most operations even can be used interactively, as they were implemented efficiently with the employed SQLite database. A stable underlying cheminformatics library guarantees the consistency and correctness of all operations. Future enhancements of the software will be guided by the simplicity principle and incorporate features supporting more daily tasks of life scientists, like similarity searches or automatic creation of subsets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00292.

The Supporting Information contains the detailed documentation of the use cases including tutorial screencasts and a description of all internal MONA properties; the documentation is provided as HTML file and is viewable in any modern browser (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rarey@zbh.uni-hamburg.de.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

MONA is available for Linux, Windows, and OS X from http://www.zbh.uni-hamburg.de/mona to academics free of charge and is temporarily for free to commercial users. All feedback is greatly appreciated and directly influences the further development of MONA.

The authors declare the following competing financial interest(s): The authors declare a potential financial interest in case the MONA software will be licensed for a fee to non-academic institutions in the future.

■ ACKNOWLEDGMENTS

The authors wish to thank Therese Inhester for helping in redesigning the database layer and Marcus Gastreich and Christian Lemmen (BioSolveIT GmbH) for even more usability testing of MONA.

■ REFERENCES

(1) *BIOVIA Pipeline Pilot 9.5*; Accelrys Software, Inc., 2015; http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/ (accessed Aug. 28, 2015).

(2) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, P. D. H., Schmidt-Thieme, P. D. L., Decker, P. D. R., Eds.; Springer: Berlin, Heidelberg, 2008; pp 319–326.

(3) Stålring, J. C.; Carlsson, L. A.; Almeida, P.; Boyer, S. AZOrange - High Performance Open Source Machine Learning for QSAR Modeling in a Graphical Programming Environment. *J. Cheminf.* **2011**, *3*, 28.

(4) Barnes, M. R.; Harland, L.; Foord, S. M.; Hall, M. D.; Dix, I.; Thomas, S.; Williams-Jones, B. I.; Brouwer, C. R. Lowering Industry Firewalls: Pre-Competitive Informatics Initiatives in Drug Discovery. *Nat. Rev. Drug Discovery* **2009**, *8*, 701–708.

(5) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(6) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*, Vol. *4*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier: Boston, MA, 2008; pp 217–241.

(7) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090.

(8) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091–D1097.

(9) Schomburg, I.; Chang, A.; Placzek, S.; Söhngen, C.; Rother, M.; Lang, M.; Munaretto, C.; Ulas, S.; Stelzer, M.; Grote, A.; Scheer, M.; Schomburg, D. BRENDA in 2013: Integrated Reactions, Kinetic Data, Enzyme Function Data, Improved Disease Classification: New Options and Contents in BRENDA. *Nucleic Acids Res.* **2013**, *41*, D764–D772.

(10) Guilloux, V. L.; Arrault, A.; Colliandre, L.; Bourg, S.; Vayer, P.; Morin-Allory, L. Mining Collections of Compounds with Screening Assistant 2. *J. Cheminf.* **2012**, *4*, 20.

(11) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C.; DataWarrior. An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473.

(12) Klein, K.; Kriege, N.; Mutzel, P. Scaffold Hunter: Facilitating Drug Discovery by Visual Analysis of Chemical Space. In *Computer Vision, Imaging and Computer Graphics. Theory and Application*; Csurka, G., Kraus, M., Laramee, R. S., Richard, P., Braz, J., Eds.; Springer: Berlin, Heidelberg, 2013; pp 176–192.

(13) *StarDrop*; Optibrium, 2015; http://www.optibrium.com/stardrop/ (accessed Aug. 28, 2015).

(14) Hilbig, M.; Urbaczek, S.; Groth, I.; Heuser, S.; Rarey, M. MONA—Interactive Manipulation of Molecule Collections. *J. Cheminf.* **2013**, *5*, 38.

(15) Urbaczek, S.; Kolodzik, A.; Fischer, J. R.; Lippert, T.; Heuser, S.; Groth, I.; Schulz-Gasch, T.; Rarey, M. NAOMI: On the Almost Trivial Task of Reading Molecules from Different File Formats. *J. Chem. Inf. Model.* **2011**, *51*, 3199–3207.

(16) Urbaczek, S.; Kolodzik, A.; Groth, I.; Heuser, S.; Rarey, M. Reading PDB: Perception of Molecules from 3D Atomic Coordinates. *J. Chem. Inf. Model.* **2013**, *53*, 76–87.

(17) Urbaczek, S.; Kolodzik, A.; Rarey, M. The Valence State Combination Model: A Generic Framework for Handling Tautomers and Protonation States. *J. Chem. Inf. Model.* **2014**, *54*, 756–766.

(18) Kolodzik, A.; Urbaczek, S.; Rarey, M. Unique Ring Families: A Chemically Meaningful Description of Molecular Ring Topologies. *J. Chem. Inf. Model.* **2012**, *52*, 2013–2021.

(19) Hipp, D. R.; Kennedy, D.; Mistachkin, J. *SQLite 3.8.6*; 2014; http://www.sqlite.org.

(20) Miller, M. A. Chemical Database Techniques in Drug Discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 220–227.

(21) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(22) Kaufman, L.; Rousseeuw, P. J. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*; John Wiley & Sons, Inc.: New York, 1990; pp 68–125.

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071–2078

(23) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.
(24) Landrum, G. *RDKit: Open-Source Cheminformatics*, Release_2015.03.1; 2015; http://www.rdkit.org (accessed Aug. 28, 2015).
(25) Ihlenfeldt, W.-D. *Cactvs 3.423*; Xemistry GmbH; http://www.xemistry.com/ (accessed Apr. 29, 2015).
(26) *OEChem v2015.June*; OpenEye Scientific Software, Inc.; http://www.eyesopen.com/oechem-tk (accessed Apr. 29, 2015).
(27) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* **2004**, *20*, 2153−2155.
(28) *LigandExpo*; http://ligand-expo.rcsb.org/ (accessed May 1, 2015).
(29) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(30) Irwin, J. J. Using ZINC to Acquire a Virtual Screening Library. In *Current Protocols in Bioinformatics*; John Wiley & Sons, Inc.: New York, 2002.
(31) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719−2740.
(32) Guha, R. *PAINS Substructure Filters as SMARTS*; http://blog.rguha.net/?p=850 (accessed Apr 29, 2015).
(33) *ZINC—Enamine Building Blocks*; http://zinc.docking.org/catalogs/enaminebb (accessed May 1, 2015).
(34) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.
(35) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

DOI: 10.1021/acs.jcim.5b00292
J. Chem. Inf. Model. 2015, 55, 2071−2078