

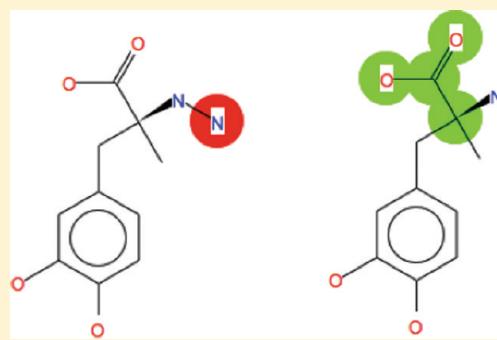
Integrated Decision Support for Assessing Chemical Liabilities

Ola Spjuth,^{*†} Martin Eklund,[†] Ernst Ahlberg Helgee,[‡] Scott Boyer,[‡] and Lars Carlsson[‡]

[†]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

[‡]Computational Toxicology, Global Safety Assessment, AstraZeneca R&D, Mölndal, Sweden

ABSTRACT: Chemical liabilities, such as adverse effects and toxicity, have a major impact on today's drug discovery process. *In silico* prediction of chemical liabilities is an important approach which can reduce costs and animal testing by complementing or replacing *in vitro* and *in vivo* liability models. There is a lack of integrated, extensible decision support systems for chemical liability assessment which run quickly and have easily interpretable results. Here we present a method which integrates similarity searches, structural alerts, and QSAR models which all are available from the Bioclipse workbench. Emphasis has been placed on interpretation of results, and substructures which are important for predictions are highlighted in the original chemical structures. This allows for interactively changing chemical structures with instant visual feedback and can be used for hypothesis testing of single chemical structures as well as compound collections. The system has a clear separation between methods and data, and the extensible architecture enables straightforward extension via addition of more plugins (such as new data sets and computational models). We demonstrate our method on three important safety end points: mutagenicity, carcinogenicity, and aryl hydrocarbon receptor (AhR) activation. Bioclipse and the decision support implementation are free, open source, and available from <http://www.bioclipse.net/decision-support>.



INTRODUCTION

Drug discovery is a complex and expensive undertaking, and the number of new approved drugs is declining.¹ The major part contributing to the high costs is attributed to late failures,² and it is hence important to select the most promising compounds as early as possible and also to identify chemical liabilities early to reduce the risk of late-stage attrition.³ Much effort is spent on improving and developing new safety models, and *in silico* modeling has been given increased attention as it can complement, and in some cases even replace, traditional *in vitro* and *in vivo* testing. Recent technological advances, such as high-throughput screening, have facilitated collection of large amounts of drug safety data and paved the way for more accurate computational models.

Existing methods for chemical liability assessment have mostly been based on quantitative structure–activity relationship (QSAR),⁴ querying chemical libraries for similar compounds,⁵ and text mining to some extent.⁶ Existing software applications that predict chemical liabilities contain predictive models based on experimental data and offer a similar range of end points including carcinogenicity, mutagenicity, teratogenicity, irritation, skin sensitization, acute toxicity, neurotoxicity, and skin permeability.

Many aspects must be taken into consideration when evaluating chemical liabilities, and the multitude of parameters that affect a drug's safety profile cannot be treated independently. This calls for integrative approaches where individual components run quickly and reliably in order to provide decision support to scientists within reasonable time. It is also important to be able to interpret results from *in silico* safety models.⁷ If

given the opportunity to see how structural changes to compounds would affect the pharmacological profile, scientists can proactively try to avoid chemical liabilities. Also, people with chemical expertise are not always as experienced with computers. There is hence a need for decision support systems with user-friendly interfaces which are accessible for people with little or no computational modeling expertise and which present results in easily interpretable formats.

We have created a safety assessment platform based on Bioclipse^{8,9} which integrates various forms of computational models and provides an end user workbench where scientists get instant decision support based on the chemical structure of a compound. We demonstrate the flexibility with various computational models, including similarity searches, structural alerts, and QSAR models and apply this on three important safety end points: mutagenicity, carcinogenicity, and aryl hydrocarbon receptor activation.

DECISION SUPPORT MODELS

Similarity Searching. A common approach when evaluating new chemical entities is to compare them with similar compounds in chemical libraries for which interesting properties are known.¹⁰ This could be for example publicly available toxicity data, results from *in vitro* assays or *in vivo* models, and metabolic pathways. Such comparisons can be implemented by querying

Received: June 1, 2011

Published: July 20, 2011

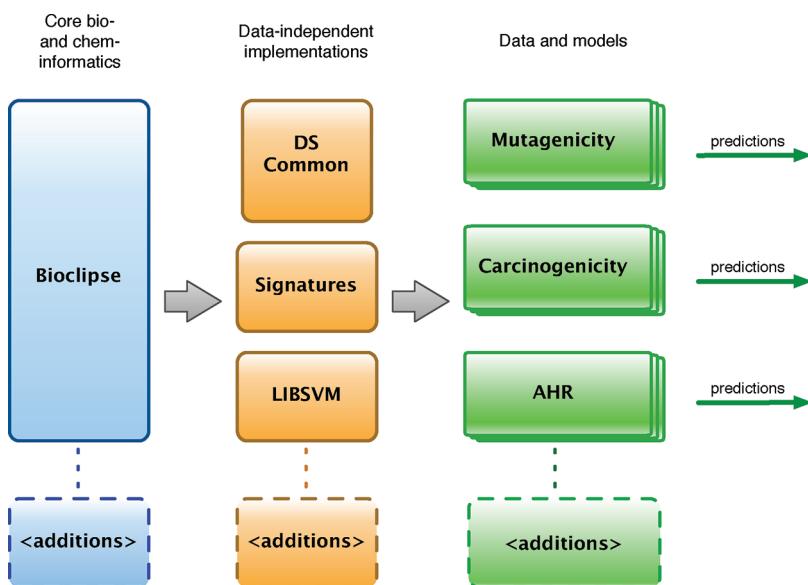


Figure 1. The architecture of the Bioclipse Decision Support system. Core bio- and cheminformatics functionality is provided by Bioclipse, upon which components for safety assessment are based. The Bioclipse-DS system consists of several data-independent implementations, including structural alerts and similarity searches (DS Common), and other resources that can be used for predictions (i.e., Signatures and LIBSVM). Separated from implementations are the bundles containing data and models, providing e.g. chemical libraries or models for a certain end point. Both data and analysis functionality are bundled as reusable components (plugins), which are represented as boxes in the figure. This implies that the system can easily be extended on all levels by addition of more plugins, which for example could be a new type of query (such as a new fingerprint or similarity measure), a new machine-learning algorithm, a new end point, or a new data set.

existing databases of chemical structures, either for exact matches or for molecules with similar structure.¹¹ The approach is straightforward and used extensively in drug discovery.⁵

We use two different, but conceptually similar, methods to assess if compounds are identical: InChI¹² and the signature molecular descriptor.¹³ To obtain similar compounds (near neighbors) we use the Chemistry Development Kit (CDK)^{14,15} version 1.3.8 to calculate CDK fingerprints, which then are compared using the Tanimoto distance.

Structural Alerts. Structural alerts are chemical substructures which have been associated, either manually or via computational modeling, to a chemical liability.¹⁶ We use two different technologies to represent structural alerts: SMARTS (<http://www.daylight.com>) and atom signatures.¹⁷

Interpretable QSAR. QSAR is a ligand based approach which aims to correlate molecular structural features to an observed biological activity.⁴ In QSAR, chemical structures are described with mathematical representations, so-called descriptors, and regression or machine learning methods are used to predict the response of new compounds. QSAR is widely used in chemical liability assessment for various end points.^{3,7,18,19} In order to obtain interpretable models, the statistical model should identify the most important descriptors, which also must be chosen so that they may give hints on how the structure could be changed in order to obtain the desired change in the predicted property.

We describe chemical structures with the signature descriptor,²⁰ which for an atom is a canonical representation of the atom's environment up to a predefined height (the number of neighboring atoms which the signature spans), and for a molecule a vector of occurrence numbers of atomic signatures. The signatures are then used to train a statistical model. Based on the result, we retrieve the variable corresponding to the largest component of the decision-function gradient at any point in the model and trace this back to the corresponding signature.²¹

This signature corresponds to an atom and an associated height, which can be highlighted in the original compound's chemical structure as the substructure which contributed the most to the prediction (regardless of the prediction outcome).

IMPLEMENTATION

Bioclipse^{8,9} is a Rich Client for the life sciences that provides means to run and integrate algorithms and tools without a network connection, while taking advantage of remote services if a network connection is available. Bioclipse is equipped with advanced chemical functionality such as 2D editing, interactive 3D visualization, I/O for the most common molecular file formats, and custom visualization of results from predictive modeling. Bioclipse is based on an advanced component-based architecture and can be extended in virtually any direction by addition of plugins, making it an ideal platform for a user-oriented workbench offering decision support for chemical liability assessment. Bioclipse and all extensions produced for this study are free, open source, and available from <http://www.bioclipse.net/decision-support>.

The decision support implementation in Bioclipse is made up of several plugins (see Figure 1 for an overview of the architecture). The main plugin, *net.bioclipse.ds*, contains the decision support integration framework and allows for execution of computational models in parallel. There is no restriction on how computational models are implemented, other than that they should accept a chemical structure as input and deliver a result which adheres to a set of rules, such as that the model should have a name and a classification into one of the following classes: POSITIVE - the model indicated a chemical liability, NEGATIVE - no chemical liability, INCONCLUSIVE - the model was unable to produce a positive or negative result, and ERROR - the model failed to produce any results.

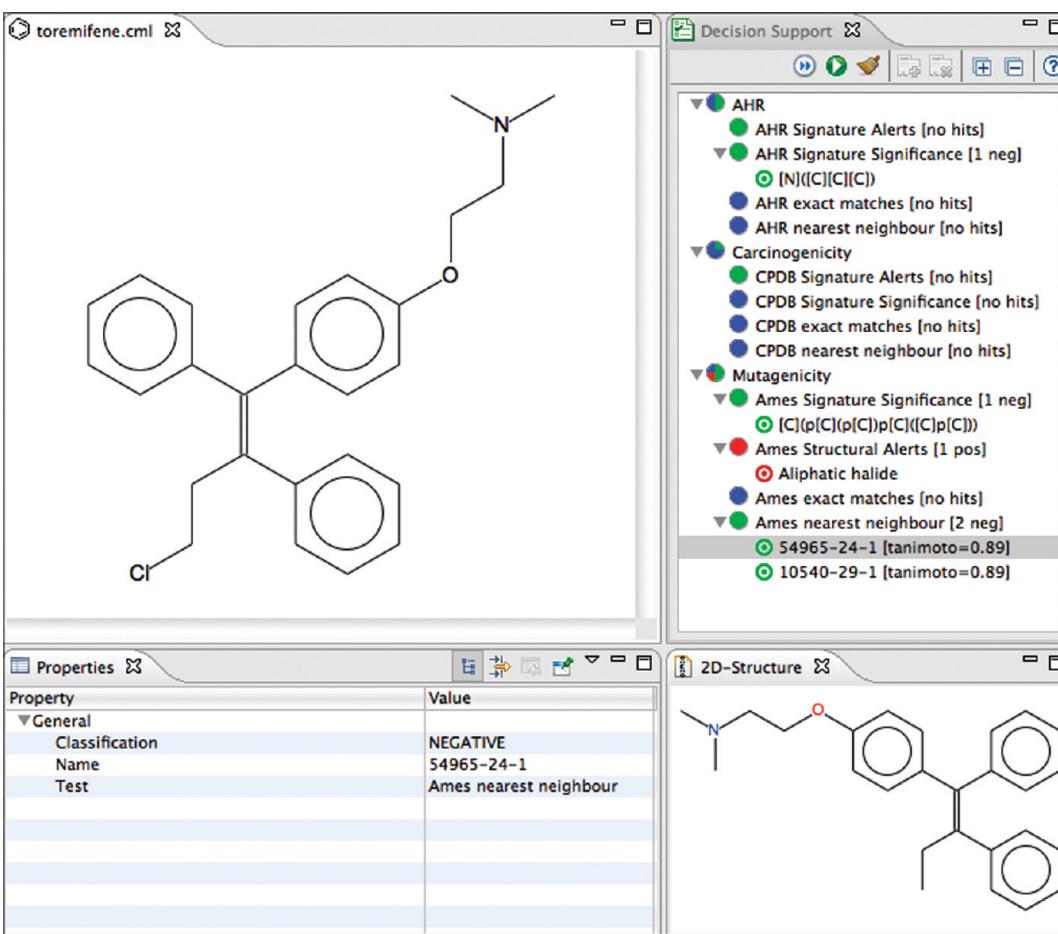


Figure 2. The compound *Toremifene*, an approved selective estrogen receptor modulator, predicted in the Bioclipse Decision Support system. In the Decision Support View (right panel) are shown individual models and their result. A near neighbor is selected, and its structure is shown in the lower right panel. In this example, a structural alert has identified an Aliphatic halide, which is an alert for mutagenicity.²⁵ However, we note near neighbors and the QSAR model (Ames Signature Significance) having negative results, and this example shows the benefit of having several models available simultaneously.

A plugin, *net.bioclipse.ds.ui*, was developed to hold user interface components, which extend the Bioclipse workbench with new graphical features. A *Decision Support View* (see Figure 2) shows the available end points and all available liability models. Users can open files containing single molecules and use the *Decision Support View* to invoke models on the chemical structure. In the case of a model producing substructure results, such hits can be highlighted in the original structure. In the case of a match in a database lookup, the returned molecules can be visualized in the separate 2D-structure view. Also added was the functionality to visualize collections of molecules in a spreadsheet with coloring of the individual decision support models (see Figure 3).

Based on the existing cheminformatics functionality in Bioclipse, several algorithm implementations were bundled in a plugin (*net.bioclipse.ds.common*), which can be used in various end points regardless of the underlying data. The implementations include querying structure-data files for similar structures using CDK fingerprints and exact matches using molecular signatures and InChI. An implementation for SMARTS querying using CDK was also added, taking a file with one SMARTS per line as input.

A simple consensus model based on majority voting was also added, together with a set of business rules to achieve a simple

weighting of results: 1) If an exact match model returns a result other than INCONCLUSIVE, then the result of this end point is set to this model's result, 2) If no near neighbors are found or if there are no matching signatures produced for a query compound, then the model returns an INCONCLUSIVE result, and 3) In all other cases, no matches are considered a NEGATIVE result.

A Bioclipse plugin, *net.bioclipse.ds.libsvm*, was constructed, integrating the libsvm library²² into the decision support system with support for classification and regression models using a Support Vector Machine (SVM) algorithm. A plugin *net.bioclipse.ds.signatures* was developed, incorporating a java implementation of the signature descriptor.

A Bioclipse plugin, *net.bioclipse.ds.report*, was assembled for creating reports of chemical decision support analyses, which can be printed and exported in various formats including PDF, RTF, XLS, CSV, XML, and HTML. A default report for chemical liability assessment was created, and it is straightforward to customize this or add new report templates using a graphical report designer.

■ END POINTS

The developed decision support system facilitates addition and integration of safety models and related informative

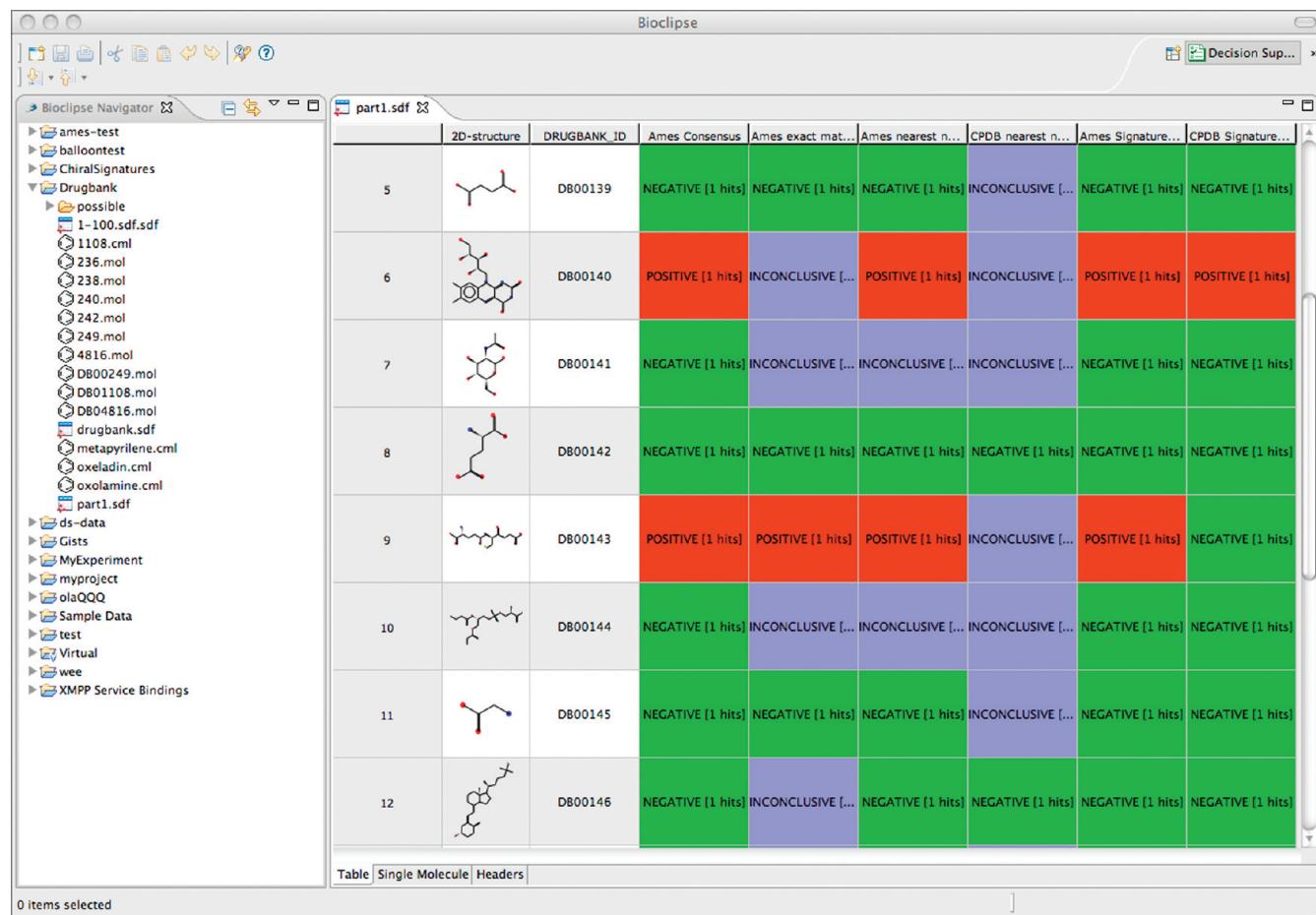


Figure 3. Screenshot from Bioclipse showing mutagenicity predictions on a collection of compounds. Negative predictions are colored in green, inconclusive in blue, and positive predictions are colored in red.

resources that can be used as decision support when evaluating a chemical compound's profile. We demonstrate our method on three safety end points: mutagenicity, carcinogenicity, and AhR, which are of prominent concern from a health and regulatory standpoint.⁵³

Mutagenicity End Point. Mutagenicity is the ability of a substance to induce mutations to DNA, which can be measured with the Ames *Salmonella*/microsome mutagenicity assay (AMES test).²⁴ Ames mutagenicity data from a study by Kazius et al.²⁵ was downloaded from (<http://cheminformatics.org/datasets/bursi/>) as a structure-data file (SD file) containing the chemical structures and mutagenicity classification stored as a property. The data set contained 4337 chemical structures of which 2401 were classified as *mutagen*. Hydrogens were removed, and CDK fingerprints, standard InChI, and signatures were added as properties, calculated using the Chemistry Development Kit (CDK)^{14,15} version 1.3.8, jniinchi version 0.8, and java-signatures version 1.0. An exact match and a nearest neighbor model were defined, using the implementations from *net.bioclipse.ds.common*, and a structural alerts model using the toxicophores defined in Kazius et al.²⁵ was also constructed, represented as a list of SMARTS.

A QSAR model was also constructed from the data set. The chemical structures were described using the signature molecular descriptor of height 0 to 3, and an SVM classification model was built using the plugin *net.bioclipse.ds.libsvm*, with the objective to

classify new compounds as either *mutagen* or *nonmutagen*. The QSAR model was bundled together with the AMES data set and other models as a plugin for Bioclipse, *net.bioclipse.ds.ames*.

Carcinogenicity End Point. A carcinogen is a type of mutagen that specifically leads to cancer, and this is another important end point in safety assessment. Carcinogenicity data originating from the Carcinogenic Potency Database (CPDB)²⁶ was downloaded from the distributed structure-searchable toxicity (DSSTox) public database network²⁷ versioned 5d (20 Nov 2008). The data set contained 892 chemical structures of which 470 were annotated as *active*. CDK fingerprints, standard InChI, and molecular signatures were calculated and added as properties in the same way as for the mutagenicity end point. An exact match and a nearest neighbor model were defined, using the molecular signatures and CDK Fingerprint implementations from the plugin *net.bioclipse.ds.common*.

Structural alerts were calculated for the data set based on atom signatures ranging from height 0 to 5, using the method described in Ahlberg Helgee et al.¹⁷ A QSAR model was also constructed from the data set. Similar to the mutagenicity end point, the chemical structures were described using the signature molecular descriptor of height 0 to 3, and an SVM classification model was built using the plugin *net.bioclipse.ds.libsvm* with the CPDB property *ActivityOutcome_CPDGBAS_Rat* as response property. The QSAR model was bundled together with the CPDB data set and other models as a plugin for Bioclipse, *net.bioclipse.ds.cpdb*.

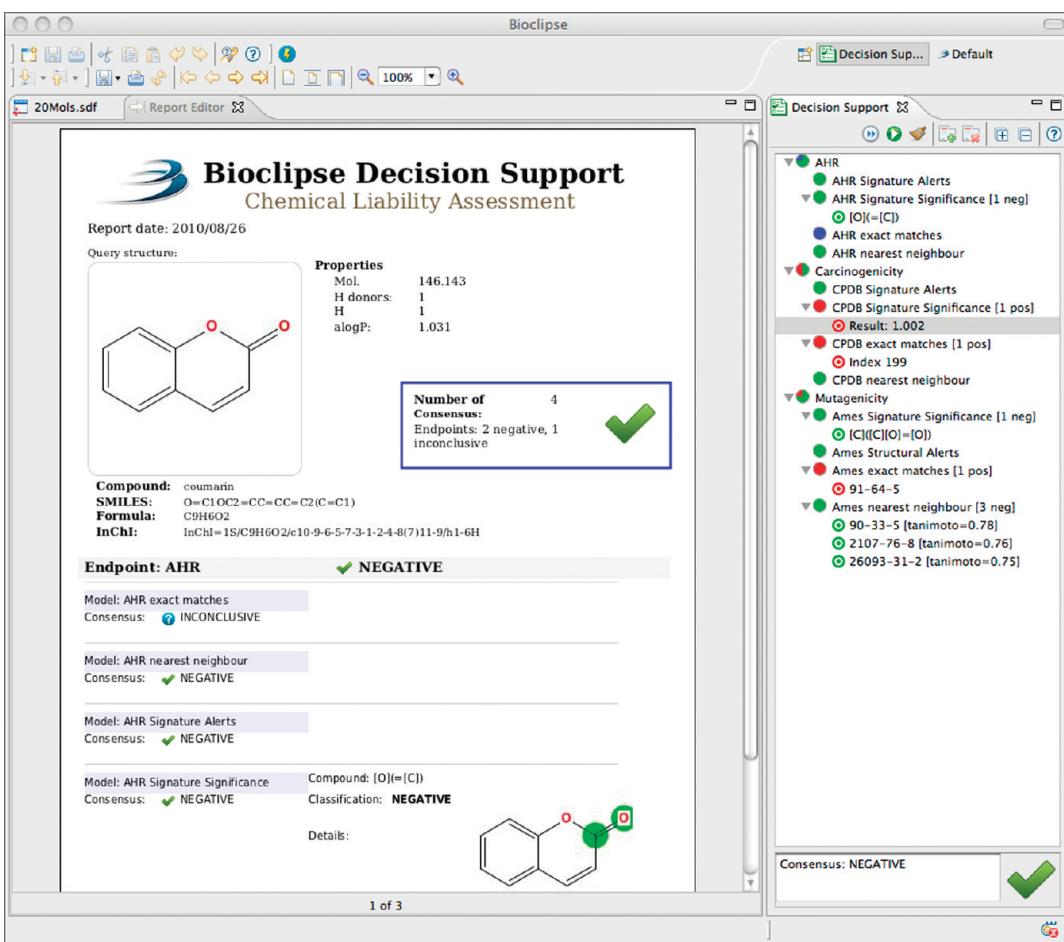


Figure 4. A generated report summarizing the chemical liability models for the compound Coumarin. The report can be printed and exported in various formats, including PDF, RTF, XLS, CSV, XML, and HTML. It is easy to add new report templates and customize existing reports using a graphical report designer.

Aryl Hydrocarbon Receptor End Point. The aryl hydrocarbon receptor (AhR) is a ligand-activated transcription factor involved in the biological response to aromatic hydrocarbons and regulates the expression of xenobiotic-metabolizing enzymes such as cytochrome P450, aldehyde dehydrogenase, quinone reductase, and other phase I and phase II detoxification genes. The ability of the AhR to bind and be activated by a range of structurally divergent chemicals suggests that the AhR contains a rather promiscuous ligand binding site.²⁸

The PubChem BioAssay 2796, named *Luminescence-based cell-based high-throughput confirmation assay for activators of the Aryl Hydrocarbon Receptor*, was downloaded from PubChem. The data were filtered for unknown CDK atom types and contained after this filtering 15951 compounds where 7971 were annotated as *Active*. The data set was converted to an SD-file using Bioclipse, and CDK fingerprints, standard InChI, and molecular signatures were calculated and added as properties in the same way as for the mutagenicity and carcinogenicity end points. An exact match and a nearest neighbor model were defined, using the InChI and CDK Fingerprint implementations from the plugin *net.bioclipse.ds.common*.

Structural alerts were calculated for the data set using the same procedure as for the carcinogenicity end point, and a

QSAR classification model was built using the same procedure as for the mutagenicity end point, capable of classifying new structures as either *Active* or *Inactive*. The QSAR model was bundled together with the AhR data set and other models as a plugin for Bioclipse, *net.bioclipse.ds.ahr*.

Model Validation. All QSAR models were built and validated according to the following procedure:

- 1 Split data set into a training set (80%) and an external test set (20%)
- 2 Use the training set to find optimal values for C and gamma by a grid-search with 5-fold cross-validated model accuracy as objective function. Build model A using the optimal values of C and gamma.
- 3 Predict class belonging of the compounds in the external test set using model A - this produces an external accuracy estimate
- 4 Use the entire data set to find optimal values for C and gamma by a grid-search with 5-fold cross-validated model accuracy as objective function. Build model B using the optimal values of C and gamma - this is the model included in Bioclipse Decision Support.

The values of external accuracy was for the AMES mutagenicity QSAR model: 0.85, the CPDB carcinogenicity QSAR model: 0.68, and the AhR QSAR model: 0.83. External classification

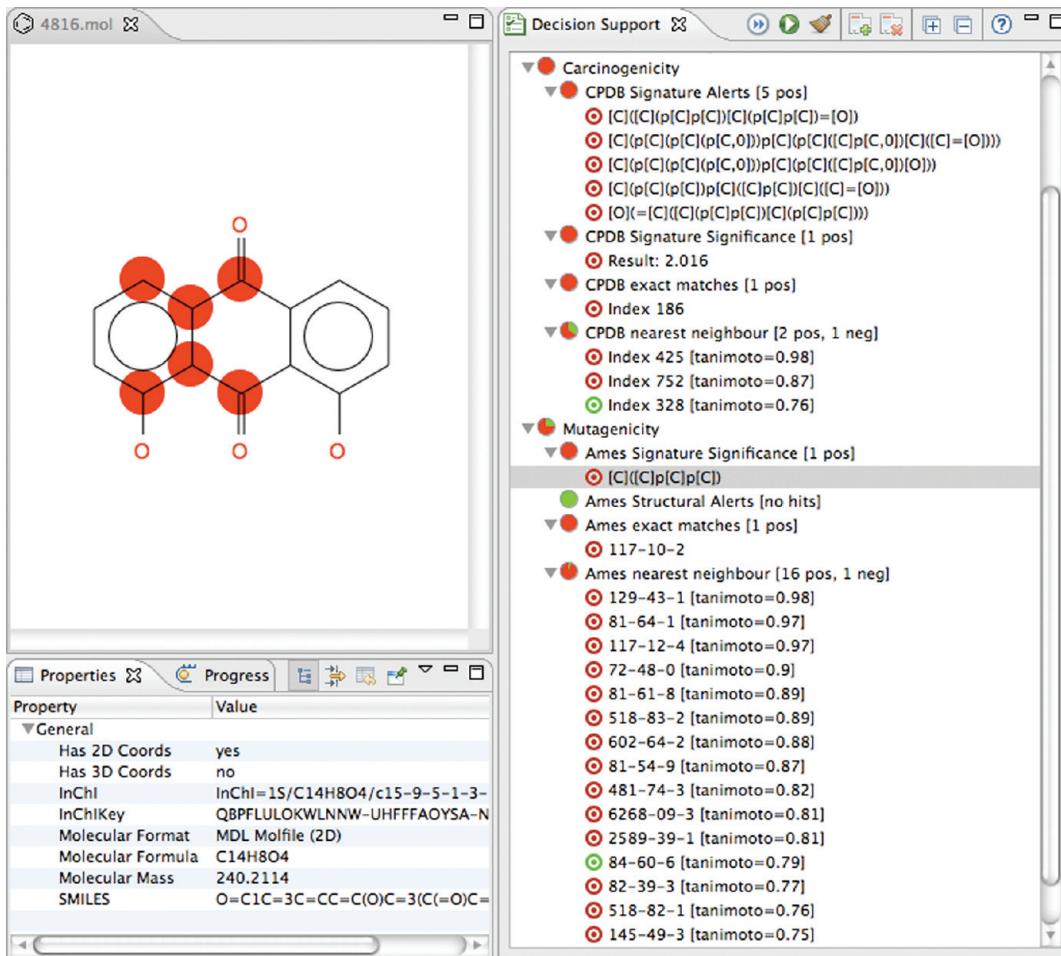


Figure 5. The compound *Danthon* which was withdrawn from the market in 1998 due to genotoxicity. We see that the AMES predictions, even if the exact match is omitted, give a clear indication of possible genotoxicity, regardless of the fact that there are no matches for the manually constructed structural alerts. We also see that the compound is predicted to have carcinogenic properties with support from several models. The atoms which are highlighted in red contributed the most to the positive mutagenicity prediction, according to the QSAR model which is selected in the right panel.

accuracy is communicated to the users in Bioclipse Decision Support as a property which is displayed when the model is selected in the GUI.

■ DISCUSSION

We used the Bioclipse decision support system to study molecules available in DrugBank.²⁹ Figure 2 shows the compound *Toremifene*, with a structural alert for mutagenicity (aliphatic halide) but near neighbors and a QSAR model (Ames Signature Significance) displaying negative results. This example shows the importance of having several models available simultaneously. Figure 5 shows the withdrawn drug *Dantron* with hazard indications and good agreement between models. An example of how structural modifications affect predictions is demonstrated in Figure 6 with the approved drug *Candersartan*. Figure 3 shows a collection of compound evaluated with the Mutagenicity models. Applying multiple predictive models to compound collections gives the ability to single out problematic compounds in a collection and ranks molecules according to model results. Benigni has shown that such ranking of collections of chemicals can be favorable to predicting liabilities for individual compounds.³⁰

The Bioclipse decision support system is a flexible system which provides a framework where new end points, data, and implementations can be easily added or reused in different contexts and enabling secure combination of public- and private data and models. The workbench can be customized for individual user needs, for example some users could choose to install only certain end points and data sets and obtain a decision-support system suitable for their specific needs. A key point of the implementation is the encapsulation of functionality into reusable plugins, which can be applied to any data (see Figure 1). An example is the implementation of the signature descriptor in a plugin (*net.bioclipse.ds.signatures*), which can be used when predicting any QSAR model which is based on this descriptor. A second example is the SMARTS matching implementation in the plugin *net.bioclipse.ds.common*, which can be used to query any data set for a list of structural alerts described as SMARTS. The flexible Eclipse Public License (EPL) allows for both open source and commercial plugins to extend the system.

Using interpretable chemical descriptors in the QSAR models has the advantage that results can be related back to the chemical structure. In the Bioclipse-DS system, both the carcinogenicity, mutagenicity, and AhR QSAR models are capable of coloring the chemical substructures based on prediction results and suggest

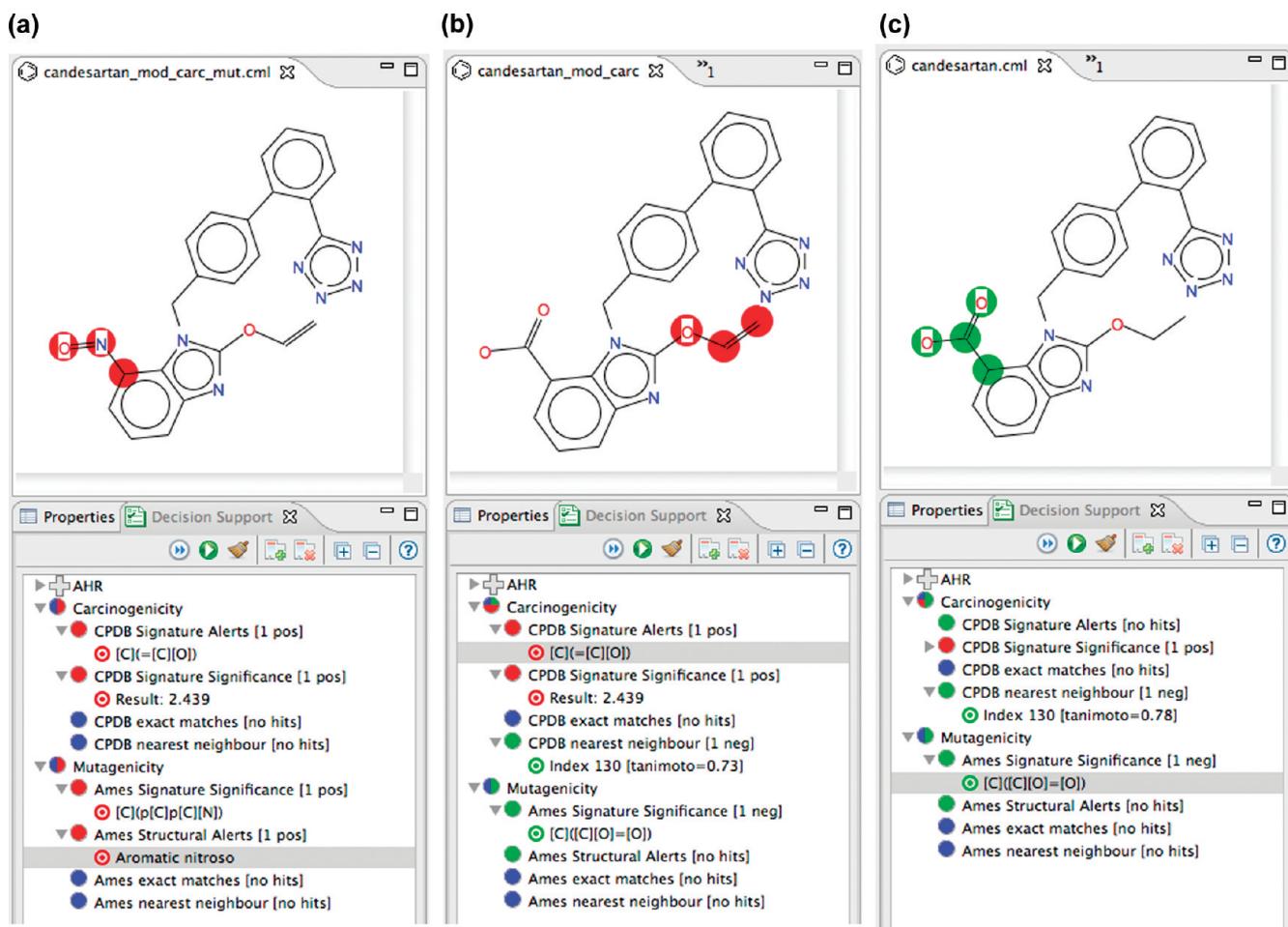


Figure 6. a) A hypothetical example showing a chemical structure similar to the approved drug Candesartan. We note that we have positive predictions for both mutagenicity and carcinogenicity, with the mutagenicity structural alert *aromatic nitroso* highlighted. b) Here the aromatic nitroso has been replaced with a carboxyl group, and we see that the mutagenicity model is now predicting a negative result. We still note that the carcinogenicity model indicates a structural alert for the signature $[C](=[C])[O]$. c) Here we have replaced the double bond in the structural alert with a single bond, and we see overall negative predictions with the highlighted carboxyl group indicating a large contribution to the negative mutagenicity prediction. This is in fact the chemical structure of Candesartan. Note that this is a hypothetical example but demonstrates a possible workflow in the Bioclipse Decision Support system.

where it would be favorable to modify a structure (in the case of a prediction with positive result) or which substructures should be kept unchanged (in the case of a prediction with negative result). A feature when evaluating individual chemical compounds is the ability to turn on automatic calculations and get instant predictions when editing the chemical structure. This allows for testing different hypotheses and explore effects of structural modifications, which is deemed as an important property for *in silico* drug safety system.⁷ Users can also inspect the accuracy of QSAR models, which is presented in the form of classification accuracy based on an external test set. The ability to pinpoint structurally important parts of molecules also makes it possible to optimize the molecules for a certain predictive model.³¹ As models for optimization are computationally intensive, we envision them to be deployed on high performance computing facilities or computational clouds³² and invoked by asynchronous Web services.³³

We conclude that Bioclipse-DS exposes all the necessary features that are required by an advanced decision support system. The responsive models with very fast results have the

possibility to change how people work with testing different hypothesis regarding effects of structural changes. Although we have demonstrated our method on standard data sets which are publicly available and used standard approaches to modeling (such as SVM for QSAR modeling and fingerprints for similarity searches), the approach is general and could be applied to any data.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ola.spjuth@farmbio.uu.se.

■ ACKNOWLEDGMENT

This work was supported by the Swedish VR (04X-05957) and Uppsala University (KoF 07). The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010009.

■ REFERENCES

- (1) Hughes, B. 2009 FDA drug approvals. *Nat. Rev. Drug. Discovery* **2010**, *9*, 89–92.
- (2) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug. Discovery* **2010**, *9*, 203–14.
- (3) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.
- (4) C., H. A Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232–239.
- (5) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.
- (6) Frijters, R.; Verhoeven, S.; Alkema, W.; van Schaik, R.; Polman, J. Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics* **2007**, *8*, 1521–1534.
- (7) Egan, W. J.; Zlokarnik, G.; Grootenhuis, P. D. In silico prediction of drug safety: despite progress there is abundant room for improvement. *Drug Discovery Today: Technol.* **2004**, *1*, 381–387.
- (8) Spjuth, O.; Helmus, T.; Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. S. Bioclipse: an open source workbench for chemoand bioinformatics. *BMC Bioinf.* **2007**, *8*, 59.
- (9) Spjuth, O.; Alvarsson, J.; Berg, A.; Eklund, M.; Kuhn, S.; Mäšak, C.; Torrance, G.; Wagener, J.; Willighagen, E. L.; Steinbeck, C.; Wikberg, J. E. S. Bioclipse 2: a scriptable integration platform for the life sciences. *BMC Bioinf.* **2009**, *10*, 397.
- (10) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (11) Basak, S. C.; Grunwald, G. D. Molecular Similarity and Estimation of Molecular Properties. *J. Chem. Inf. Comput. Sci.* **1995/05/01/**, *35*, 366–372.
- (12) Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* **2005**, *3*, 1832–1834.
- (13) Faulon, J.-L.; Collins, M. J.; Carr, R. D. The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 427–436.
- (14) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–20.
- (15) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- (16) Ashby, J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ. Mutagen.* **1985**, *7*, 919–921.
- (17) Ahlberg Helgee, E.; Carlsson, L.; Boyer, S. Identification of Toxicologically Relevant Substructures in Large Datasets Using Atom Signatures. *Manuscript in preparation*
- (18) Dearden, J. C. In silico prediction of drug toxicity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 119–127.
- (19) Valerio, L. G. J. In silico toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol.* **2009**, *241*, 356–370.
- (20) Faulon, J.-L.; Visco, D. P. J.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.
- (21) Carlsson, L.; Ahlberg Helgee, E.; Boyer, S. Interpretation of nonlinear QSAR models applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2551–2558.
- (22) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. 2001. Software available at <http://www.csie.ntu.edu.tw/cjlin/libsvm> (accessed 2011–06–30).
- (23) Richard, A. M. Structure-based methods for predicting mutagenicity and carcinogenicity: are we there yet? *Mutat. Res.* **1998**, *400*, 493–507.
- (24) Mortelmann, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **2000**, *455*, 29–60.
- (25) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
- (26) Fitzpatrick, R. B. CPDB: Carcinogenic Potency Database. *Med. Ref. Serv. Q.* **2008**, *27*, 303–311.
- (27) Richard, A. M.; Williams, C. R. Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat. Res.* **2002**, *499*, 27–52.
- (28) Denison, M. S.; Nagy, S. R. Activation of the aryl hydrocarbon receptor by structurally diverse exogenous and endogenous chemicals. *Annu. Rev. Pharmacol. Toxicol.* **2003**, *43*, 309–34.
- (29) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploratio. *Nucleic Acids Res.* **2006**, *34*, D668–72.
- (30) Benigni, R. Chemical structure of mutagens and carcinogens and the relationship with biological activity. *J. Exp. Clin. Cancer Res.* **2004**, *23*, 5–8.
- (31) Ahlberg Helgee, E.; Carlsson, L.; Boyer, S. A method for automated molecular optimization applied to Ames mutagenicity data. *J. Chem. Inf. Model.* **2009**, *49*, 2559–2563.
- (32) Dudley, J. T.; Butte, A. J. In silico research in the era of cloud computing. *Nat. Biotechnol.* **2010**, *28*, 1181–5.
- (33) Wagener, J.; Spjuth, O.; Willighagen, E. L.; Wikberg, J. E. S. XMPP for cloud computing in bioinformatics supporting discovery and invocation of asynchronous web services. *BMC Bioinf.* **2009**, *10*, 279.