Combined graph/relational database management system for calculated chemical reaction pathway data

Timur Gimadiev[1], Ramil Nugmanov[2], Dinar Batyrshin[2], Timur Madzhidov[2], Satoshi Maeda[1], Pavel Sidorov[1]*, Alexandre Varnek[1,3]*

[1] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan
[2] Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institute of Chemistry, Kazan Federal University, 18, Kremlyovskaya str., 420008 Kazan, Russia
[3] Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, 4, Blaise Pascal str., 67081 Strasbourg, France
*e-mail: pavel.sidorov@icredd.hokudai.ac.jp, varnek@unistra.fr

Abstract

Nowadays quantum chemical calculations are widely used to generate extensive datasets for machine learning applications, however, generally these sets only include information on equilibrium structures and some close conformers. Exploration of potential energy surface provides an important information on ground and transition states, but analysis of such data is complicated due to the number of possible reaction pathways. Here, we present RePathDB, a database system for managing 3D structural data for both ground and transition states resulted from quantum chemical calculations. Our tool allows to store, to assemble and to analyze reaction pathway data. It combines relational database CGR DB for handling compounds and reactions as molecular graphs with a graph database architecture for the pathway analysis by graph algorithms. Original Condensed Graph of Reaction Technology is used to store any chemical reaction as a single graph.

Introduction

Public and commercial chemical databases collect, curate and index thousands of chemical data records daily. They range from gigantic libraries like ZINC[1] (over 700M 2D structure records, as of 2018) or CAS registry[2] (over 163M small molecules) including purchasable chemical compounds, to smaller, more focused databases such as ChEMBL[3] (over 1.9M structures with bioactivity records) or DrugBank[4] (about 14 000 registered drugs, with biological targets and clinical information). Experimental reaction data is also quite abundant: CAS REACT[5] includes more than 128 million reactions and synthetic preparations, while Reaxys[6] has over 53M reactions. All these databases greatly assist organic and medicinal chemists, chemoinformaticians in their research.

However, conventionally the databases deal with the relational tables of 2D molecular structures, since this format is more convenient to handle and most algorithms such as structural search etc., are well implemented only for 2D structures. Needless to say, the sources of 3D structures are also quite scarce – they can be obtained either via crystallographic studies, or by quantum chemical calculations.

Datasets of 3D structures calculated by quantum chemistry have started emerging recently with the growing interest of application of machine learning methods to emulate QC. The most prominent QM7 and QM9 sets[7] are based on GDB virtual database which assembles combinatorially enumerated structures containing up to 7 or 9 heavy atoms, respectively. The PubchemQC[8] dataset was created in a similar manner: it contains extracted from Pubchem database structures for which the geometries and energies of the ground and excited states were calculated using TD-DFT method. The largest available set of structures calculated by QC is ANI set[9,10] containing > 5M structures.

Accurate predictions of ground states, however, has limited applications, notably if it concerns single molecules. Exploring the potential energy surface (PES) of molecular ensembles could lead to better understanding of intermolecular interaction and, especially, chemical transformations. With the advancement in computational power, it is now possible to explore PES for ground and transition states in extensive QC calculations[11–15] taking into account also the transformations occurring in molecular complexes. For example, recently developed Global Reaction Route Mapping (GRRM)[16,17] approach for a given set of molecules, provides with multiple reaction pathways at DFT level. However, the analysis of such data is challenging due to the large number of both equilibrium structures and transition states, as well as intermediate points on the slopes of PES. Conventional relational databases architecture can hardly be used for the storage and analysis of this information because (i) both 3D and 2D structures must be stored; (ii) equilibrium, transition and intermediate states should be stored in different relational tables; and, therefore, (iii) building reaction pathways would require multiple joinings of these tables. In this case, represent of the ensemble of data as a graph looks more convenient, because it allows to apply various graph theory algorithms (e.g., shortest path search) for the analysis of pathways. Recently, application of a graph database architecture for chemical data has been reported[18] as a tool for merging data from ChEMBL abd DrugBank.

In this work, we report a new reaction database management system combining relational and graph architectures which allow to consider both 2D and 3D objects. On one hand, all 3D structures calculated by QC are stored along with their energy and some other selected parameters. On the other, all conformations of a given reaction complex can be stored as a 2D molecular graph (or set of graphs), thus allowing to perform conventional structural or similarity searching. The treatment of chemical transformations is based on the Condensed Graph of Reaction (CGR) approach[19]. A CGR represent a single molecular graph resulted from superposition of related atoms of reactant and products. This graph contains both conventional chemical bonds and such called dynamical bonds and atoms characterizing chemical transformations (breaking, formation, change of order, etc.). This technology greatly facilitates the storage and analysis of chemical reaction data. In this work, we extend the CGR approach to three-dimensional space by encoding a given transition state structure by a 3D graph. Thus, resulting database accommodates 3D structures calculated by QC, related 2D structures together with reactions (CGRs) in a single graph. Special algorithms for a fast search of reaction paths between two selected species have been implemented. The developed database system is available as a Python library.

Methods

Condensed Graph of Reaction

In silico handling of chemical reactions can be significantly simplified by the Condensed Graph of Reaction (CGR) approach[19]. In this framework, the structures of reactants and products are merged into a single molecular graph (Figure 1). The CGR edges correspond either to standard chemical bonds or to “dynamic” bonds describing transformations. Similarly, dynamic atoms may also be designated as atoms that change a property (e.g., formal charge) during the transformation. In such a way, one can consider a CGR as a pseudomolecule for which some types of molecular descriptors can easily be computed followed by their application in data analysis and statistical modeling tasks.[20] This approach has been successfully applied to various tasks, such as similarity search in reaction databases,[19,21] building quantitative structure-reactivity models,[22–25] assessment of tautomer distributions,[26,27] prediction of activity cliffs,[28] classification of enzymatic transformations,[29] prediction of reaction conditions[30,31].

Here we use recently developed CGRtools Python library[32] as a base for storage and analysis of compound and reaction data.

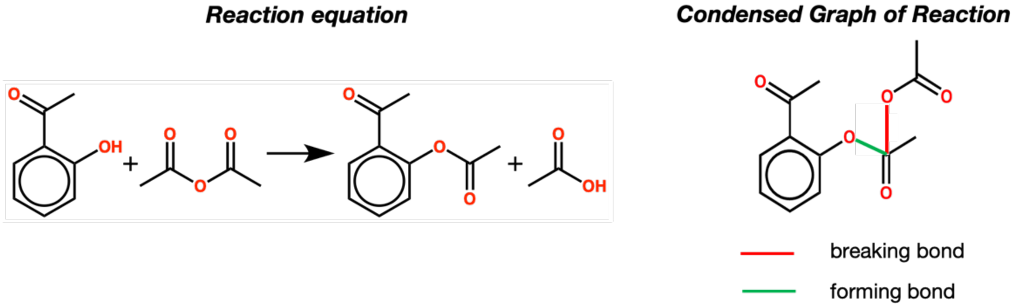

Figure 1. Traditional depiction of reaction equation (left) and the corresponding Condensed Graph of Reaction (right), exemplified by an esterification reaction. In addition to conventional bonds of different orders, CGR contains dynamical bonds (forming and breaking during the transformation), indicated here by green and red color, respectively.

Global Reaction Route Mapping and Artificial Force-Induced Reaction

The exploration of reaction pathways as parts of PES are performed here by Global Reaction Route Mapping program, version 2017 (GRRM17), with the Artificial Force-Induced Reaction (AFIR) methodology. AFIR methods introduces a special force to make slight deformations on the molecular structure in order to rapidly analyze the reaction pathways including both equilibrium structures and transition states. The calculations are performed using user-defined computational level (semi-empirical, DFT, etc.).

Here, as a case study we consider the Wohler's urea synthesis pathways data resulted from GRRM17 calculations[33]. The data set consists of >850 equilibrium structures and >5000 approximate and optimized transition states calculated at DFT level with ωB97X-D basis set (cf. original paper[33] for details).

Database management

All new types of entities in this project (see "Results and Discussion" section) are implemented as extensions of Neo4J database[34] (v. 3.3) nodes, via neomodel Python library (v. 3.3.2). CGRdb and CGRtools (v.4) Python libraries have been used for storing and management of molecular structures, encoding reactions as Condensed Graphs, and substructural and similarity searches. CGRdb is based on PostgreSQL. As the developed library requires an extended list of external DB tools to be properly installed and set up, it is also available as a Docker container.

Feeding the data to the database requires Neo4J, CGR DB to be launched on the server, and the following command to be run with appropriate information (hosts, ports, etc.):

```
   python -m repathdb
   -pg POSTGRES_CONNECTION_INFO ("//user:port@host:port/table")
   -nj BOLT_CONNECTION_INFO("bolt://login:pass@host:port")
populate -f FOLDER -s FILE_TYPE
```

"Populate" command above will launch he script that will search the specified folder (-f option) for files (-s option) and will parse these files to extract 3D coordinates and the energy of structures. Please note that the parser has been implemented specifically for the output files created by GRRM/AFIR software, so different formats will require the development of a new, specific parser.

Web interface

Web interface is implemented with Plotly Dash framework[35] as it is the most convenient modular framework that can include a wide variety of widgets on a web page. MarvinJS web plugin (v.20.2.0, 2020, ChemAxon, http://www.chemaxon.com) is used for the search query editor. 3D structure visualization is implemented as a custom JavaScript widget based on the 3DMol.js library[36].

Results and discussion

RePathDB architecture

The developed database and database management system combines two architectures – a relational database for molecular graphs (for both compounds and reactions) and related properties, implemented using CGR DB library, and a graph database to represent the relationships between the entities, built using Neo4J.

As a DBMS developed specifically for dealing with chemical reactions using the CGR concept, the purpose of CGR DB in here is two-fold. First, CGRtools library serves as the base of CGR DB and allows to parse a variety of chemical formats, including XYZ format storing 3D coordinates of all atoms of the system. Second, it is used to perform fast search by chemical structures (substructural or similarity search) for both molecules and reactions, represented as molecular graphs.

On the other hand, the relationships between entities in RePathDB are managed by a graph database, rather than by the conventional "key-attribute" paradigm of a relational table. This is especially convenient when the nature or the number of relations for each entity is not known beforehand. Another advantage is the accessibility of graph algorithms. The set of pathways of a reaction can be imagined as a graph, therefore, such architecture facilitates the exploration and analysis by graph algorithms. For example, the shortest path between a reactant and a product can be easily found via a graph.

To implement the graph architecture (see the schematic view on Figure 2), we introduce entities (graph nodes) of the following types:

1) ***Molecule*** (***M***) represents an individual chemical compound (e.g., NH_3, H_2O, etc.). *Molecules* are stored as 2D molecular graphs.

2) ***Complex*** (***C***) describes an ensemble of species involved in a system investigated by quantum calculations and consists of one or several ***M*** stored as 2D molecular graphs. Since the order of atoms may vary from one ***C*** to another one, a special tool assures atom-to-atom mapping of a given ***M*** into related ***C***.

3) ***Reaction*** (***R***) encodes a transformation of one ***C*** to another one by a Condensed Graph of Reaction (CGR). Atom-to-atom mapping of two *Complexes* ***C1*** and ***C2*** participating in ***R*** is stored. For each pair of related *Complexes*, two *Reactions* are considered: from ***C1*** to ***C2***, and vice versa (see Figure 1).

4) ***Equilibrium structure*** (***S***) is a particular 3D configuration of ***C*** corresponding to a minimum on the PES. Since several geometries may correspond to local minima, one ***C*** can be related to several ***S***. Atomic coordinates and energy of every ***S*** as well as atom-to-atom mapping of ***S*** into ***C*** are stored in the database.

5) Two types of ***Transition state*** (***TS***) are considered: true or approximate ones. Each ***TS*** relates two ***S***. In 2D perspective, ***TS*** corresponds to a *CGR* of encoding a given *Reaction*. 3D atomic coordinates and energies of transition states are stored in the entities.

6) ***Molecular formula (B)*** represents the full atomic composition of the studied system and is the same for all pathways considered in the given study. These pathways can be viewed as different parts of the same PES.

7) Single system (explored PES for one reaction or one reaction mapping) may include hundreds of single transformation paths and, therefore, will accommodate many connections C – R – C – R – C …, etc. Schematically, the architecture of the database for a single transformation is shown on Figure 2.

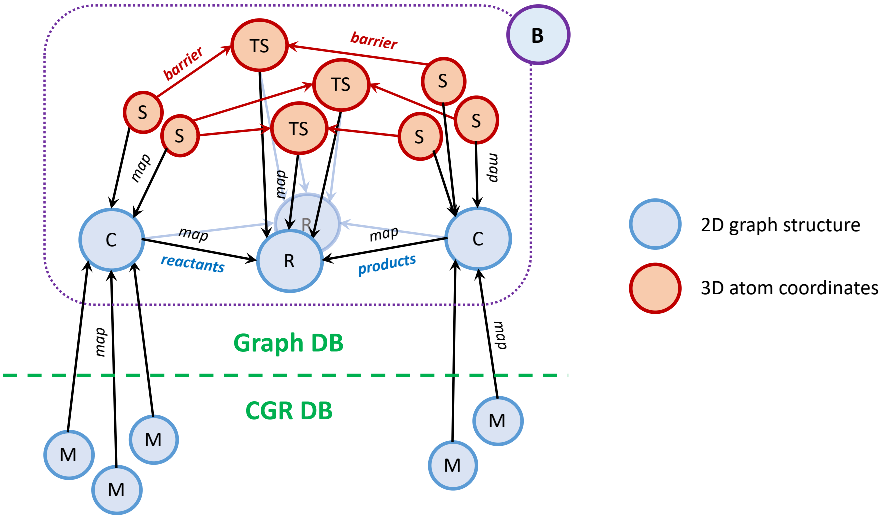

Figure 2. Schematic representation of the RePathDB architecture for transformation of one *Complex* to another one. Molecule nodes (***M***) represent individual chemical compounds. *Complex* (***C***) is a combination of several molecules. Each *Complex* is represented by one or several equilibrium structures (***S***). Transformation of one complex ***C*** to another corresponds to *Reaction* (***R***), stored as a Condensed Graph of Reaction. For each pair of related *Complexes*, both forward and backward reactions ***R*** are considered. From 3D perspective, each ***R*** is associated with one or several *Transition states* (***TS***), each connecting two ***S***. The arrows show atom-to-atom mapping ***M*** to ***C***, ***S*** to ***C*** and ***TS*** to ***R***. Any entity with the same atomic composition is also connected to the corresponding *Molecular formula* node (***B***), the connections are not shown here for easier reading.

3D structure parsing

Typical output of quantum chemical calculations contains a list of atoms together with their Cartesian coordinates, without any information concerning chemical bonds. The existence of a chemical bond is usually deduced from related interatomic distance. In order transform a 3D structure into molecular graph a bond type assignment (single, double, etc.) is also required. In order to determine both chemical bonds and their types for a given structure, the following rules have been applied:

1. A single bond between atoms *i* and *j* is assigned if related interatomic distance $D_{ij} < 1.25(R_i+R_j)$, where R_i and R_j are covalent radii of the corresponding atoms. Notice that all hydrogen atoms are explicitly accounted for.
2. All atoms are then checked for full valence (implementation by CGRtools library).
3. An atom failing the full valence check searches connected neighbors for other non-full valences. If found, related bond order is changed to double or triple.
4. If the valence check still fails and there are no neighboring atoms with vacancies, formal atomic charges are assigned instead.

Related software tool (XYZ parser) has been implemented in CGRtools library. The log file parser used to extract an information from the GRRM/AFIR output files is specialized for this format only. The parser searches files with specific headers and processes them only if they have, at least, three structures: initial and final equilibrium states, and the maximum energy point (real or approximate transition state). The parser checks the existence of transition state (maximum energy) and ignores the files lacking this information. For other data formats and purposes, the parser can further be customized.

Reaction pathway search

The main advantage of mapping 3D structures to 2D molecular graphs is the simplification of the underlying data structure. A single GRRM run may result in thousands of isolated equilibrium states and even more paths and transition states connecting them. Performing a shortest path search on such an extended graph becomes extremely cumbersome. Furthermore, as the number of nodes grows up, especially in the case of merging information from many different runs, an application of standard search algorithms becomes hardly possible. Several tests performed on some 200 000 nodes gathered from four different runs, demonstrated that the search could not be resolved in reasonable time.

The aggregation of several 3D structures with the help of corresponding 2D structure (either a molecular complex, or a CGR) significantly reduces the number of nodes in the final graph (see Figure 3). The search for the reaction pathway connecting two molecules for the reduced graph is implemented in RePathDB. Technically, a search can be launched using either command line or web interface. When web interface is used, a query involving given initial and final structures can be prepared with the MarvinJS structure editor. Structural keys (fragments) implemented in CGR DB are used in substructure and Tanimoto similarity searching algorithms.

Once both reactant and product molecules are detected, the graph database will find the most optimal path between them. The optimality metric is estimated as the sum of all energy barriers along the path, i.e. the differences of energies between a transition state and the preceding equilibrium state are summed up to form the score. Only the nodes belonging to the reduced graph are considered (see Figure 3). In this case, the lowest energies among corresponding 3D configurations are used to assess the energy barriers. Alternatively, each barrier can be ranked along the way, so that the lowest is selected at every step. A test search for the shortest path between two given molecules among 200 000 nodes corresponding to four different runs (therefore, four different atomic compositions) resolved in less than 5 minutes using a server with Intel Xeon 12-core processor and 64 GB RAM. Current implementation of the web interface shows five most optimal paths from reactant to product, their scores and the related energy diagrams to the user.

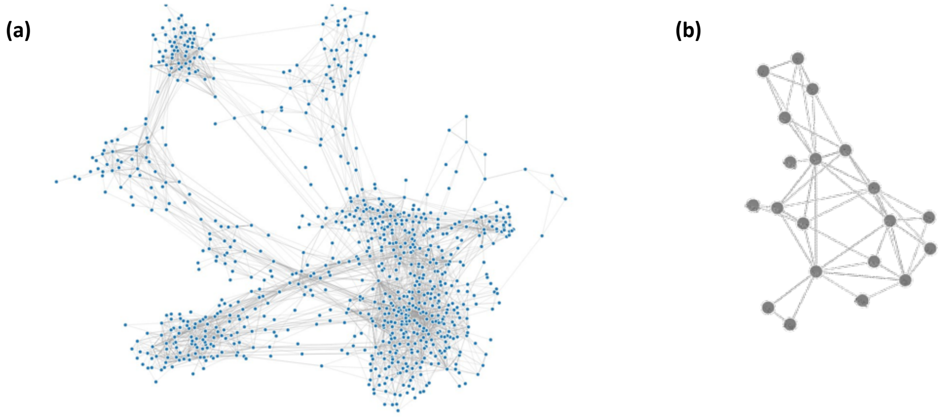

Figure 3. Full reaction pathway graph obtained from GRRM studies (a) and related reduced graph (b). Nodes in the full graph correspond to the equilibrium states discovered in QC calculations, edges correspond to true or approximate transition states. In the reduced graph, the nodes correspond to the 2D structure of the complexes (objects **C** in Figure 2), whereas the edges to reactions (objects **R** in Figure 2) linking two complexes. This allows to reduce the number of nodes from 852 in the full graph to 20 in the reduced graph, which leads to significant acceleration of the reaction pathways search. Graphs are visualized with D3.js library.

Transition states as 3D CGRs

GRRM/AFIR calculations allow to extensively explore the PES to find numerous transition states of a given system. The number of these TS may be great: e.g., for Wohler's urea synthesis pathways 152 true TS (representing slow processes like chemical bond rearrangements) and over 5000 approximate TS (representing fast processes like conformational changes) were found. On the other hand, they represent a smaller number of actual transformations (**R** in Figure 2). In the database, all reactions are encoded into Condensed Graphs of Reactions (CGR). Ensemble of transition states (***TS***) associated with a given **R** are simply different 3D geometries corresponding to the same transformation, and thus can be represented by a single 3D CGR (see example on Figure 4).

A special extension for 3D CGR has been developed and implemented into CGRtools library (also available in web interface as 3Dmol.js widget). The information about bond types in 3D CGR is not deduced from the atomic coordinates but is taken directly from 2D CGR.

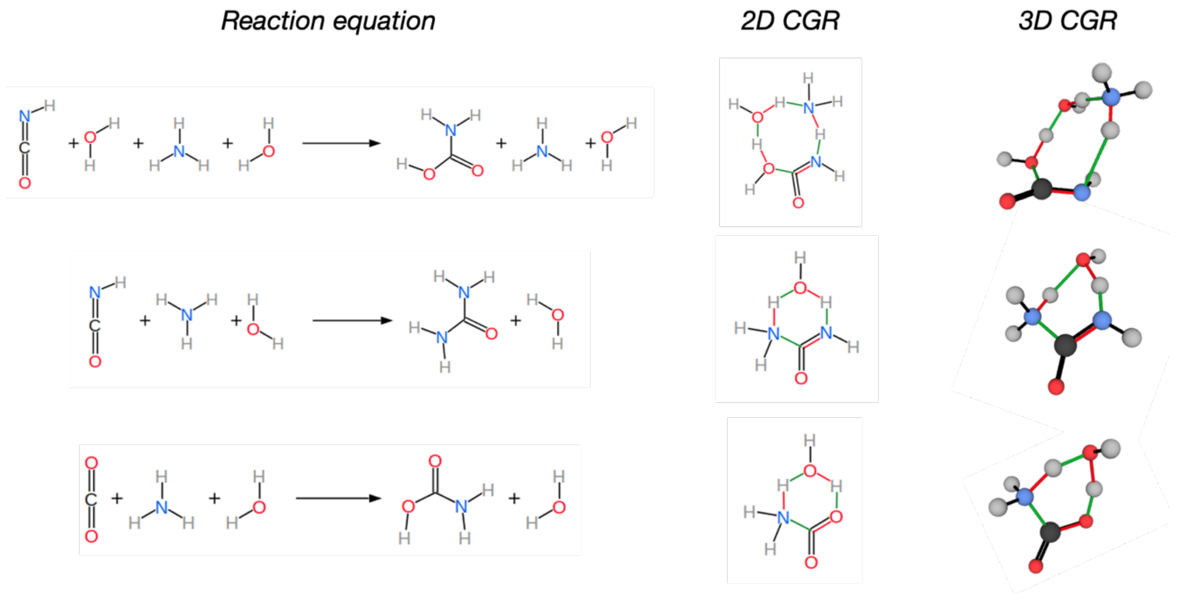

Figure 4. Examples of reactions along the urea synthesis pathway (left), corresponding 2D CGR (center) and related 3D CGR (right) describing the lowest-energy conformer of the transition state of each reaction. Notice that molecules are retained in the formal reaction equation whenever they are a part of the transition state structure. Conventional, forming and breaking bonds are shown in black, green and red, respectively. Superposition of red and black sticks corresponds to a double-to-single bond transformation.

Conclusions

In this paper we present a new database architecture for storage and analysis of 3D chemical structures calculated by quantum chemistry. While the management of data related to single molecules is possible with conventional relational databases, they are less suitable for chemical reaction data. Specifically, the pathway of chemical transformations during the exploration of PES is more easily understood as a graph. Therefore, we propose a database system that combines the graph and relational architectures and all their advantages.

Since the number of 3D structures obtained via the exploration of PES even for one reaction system may be overwhelming, our tool circumvents this by relating all equilibrium structures to molecular graphs, and all transition states to Condensed Graphs of Reactions. This allows to reduce the computational cost of searches and pathway scoring significantly. Furthermore, CGR concept allows to represent the 3D structure of a transition state in a new way, actually showing the transformations happening during the reaction. The developed system is implemented as a Python library.

Code availability

Described Python library is freely available at GitHub: https://github.com/icredd-cheminfo/RePathDB. We also use following freely available libraries: CGRtools (https://github.com/cimm-kzn/CGRtools), CGR DB (https://github.com/stsouko/CGRdb), Neo4J Graph Database (https://neo4j.com/neo4j-graph-database/), Plotly Dash (https://plotly.com/dash/), D3.js (https://d3js.org).

Data availability

Data used to illustrate the functionality of RePathDB were taken from reference[33], and are available from the authors on demand.

Acknowledgements

PS and TG express their gratitude to WPI-ICReDD for the financial support. TM and RN are grateful to Russian Science Foundation grant No 19-73-10137 for the support of the development of the CGRdb cartridge and CGRtools library.

References

(1) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(2) CAS Registry https://www.cas.org/support/documentation/chemical-substances.

(3) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(4) 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 Exploration. *Nucleic Acids Res.* **2006**, *34* (suppl_1), D668–D672. https://doi.org/10.1093/nar/gkj067.

(5) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30* (4), 394–399. https://doi.org/10.1021/ci00068a008.

(6) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49* (12), 2897–2898. https://doi.org/10.1021/ci900437n.

(7) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1* (1), 140022. https://doi.org/10.1038/sdata.2014.22.

(8) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **2017**, *57* (6), 1300–1308. https://doi.org/10.1021/acs.jcim.7b00083.

(9) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules. *Sci. Data* **2020**, *7* (1), 134. https://doi.org/10.1038/s41597-020-0473-z.

(10) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16* (7), 4192–4202. https://doi.org/10.1021/acs.jctc.0c00121.

(11) Hori, K.; Yamaguchi, T. Chemical Reaction Transition State Search System, Method, and Program. JP5164111B2, 2008.

(12) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.* **2017**, *13* (11), 5780–5797. https://doi.org/10.1021/acs.jctc.7b00764.

(13) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *WIREs Comput. Mol. Sci.* **2018**, *8* (2), e1354. https://doi.org/10.1002/wcms.1354.

(14) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123* (2), 385–399. https://doi.org/10.1021/acs.jpca.8b10007.

(15) Sumiya, Y.; Maeda, S. Paths of Chemical Reactions and Their Networks: From Geometry Optimization to Automated Search and Systematic Analysis. *Chem. Model.* **2019**, *15*, 28–69.

(16) Maeda, S.; Ohno, K.; Morokuma, K. Systematic Exploration of the Mechanism of Chemical Reactions: The Global Reaction Route Mapping (GRRM) Strategy Using the ADDF and AFIR Methods. *Phys. Chem. Chem. Phys.* **2013**, *15* (11), 3683–3701. https://doi.org/10.1039/c3cp44063j.

(17) Maeda, S.; Harabuchi, Y.; Takagi, M.; Saita, K.; Suzuki, K.; Ichino, T.; Sumiya, Y.; Sugiyama, K.; Ono, Y. Implementation and Performance of the Artificial Force Induced Reaction Method in the GRRM17 Program. *J. Comput. Chem.* **2018**, *39* (4), 233–251. https://doi.org/10.1002/jcc.25106.

(18) Murali, V.; Königs, C.; Deekshitula, S.; Nukala, S.; Santhi, M. D.; Athri, P. CompoundDB4j: Integrated Drug Resource of Heterogeneous Chemical Databases. *Mol. Inform.* **2020**, *39* (9), 2000013. https://doi.org/10.1002/minf.202000013.

(19) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. A Representation to Apply Usual Data Mining Techniques to Chemical Reactions — Illustration on the Rate Constant of SN2 Reactions in Water. *Int. J. Artif. Intell. Tools* **2011**, *20* (02), 253–270. https://doi.org/10.1142/S0218213011000140.

(20) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P.; Solov'ev, V. P.; Solov'ev, V. P. Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput. Aided. Mol. Des.* **2005**, *19* (9–10), 693–703. https://doi.org/10.1007/s10822-005-9008-0.

(21) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. A Representation to Apply Usual Data Mining Techniques to Chemical Reactions; 2010; pp 318–326. https://doi.org/10.1007/978-3-642-13025-0_34.

(22) Madzhidov, T. I.; Polishchuk, P. G.; Nugmanov, R. I.; Bodrov, A. V; Lin, A. I.; Baskin, I. I.; Varnek, A. A.; Antipin, I. S. Structure-Reactivity Relationships in Terms of the Condensed Graphs of Reactions. *Russ. J. Org. Chem.* **2014**, *50* (4), 459–463.

(23) Madzhidov, T. I. I.; Bodrov, A. V. V; Gimadiev, T. R. R.; Nugmanov, R. I. I.; Antipin, I. S. S.; Varnek, A. A. A. Structure-Reactivity Relationship in Bimolecular Elimination Reactions Based on the Condensed Graph of a Reaction. *J. Struct. Chem.* **2015**, *56* (7), 1227–1234. https://doi.org/10.1134/S002247661507001X.

(24) Gimadiev, T.; Madzhidov, T.; Tetko, I.; Nugmanov, R.; Casciuc, I.; Klimchuk, O.; Bodrov, A.; Polishchuk, P.; Antipin, I.; Varnek, A. Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis. *Mol. Inform.* **2019**, *38* (4), 1800104. https://doi.org/10.1002/minf.201800104.

(25) Glavatskikh, M.; Madzhidov, T.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Malakhova, D.; Marcou, G.; Varnek, A. Predictive Models for Kinetic Parameters of Cycloaddition Reactions. *Mol. Inform.* **2019**, *38* (1–2), 1800077. https://doi.org/10.1002/minf.201800077.

(26) Gimadiev, T. R. R.; Madzhidov, T. I. I.; Nugmanov, R. I. I.; Baskin, I. I. I.; Antipin, I. S. S.; Varnek, A. Assessment of Tautomer Distribution Using the Condensed Reaction Graph Approach. *J. Comput. Aided. Mol. Des.* **2018**, *32* (3), 401–414. https://doi.org/10.1007/s10822-018-0101-6.

(27) Gimadiev, T. R.; Nugmanov, R. I.; Madzhidov, T. I.; Polishchuk, P. G.; Petrovsky, A. S.; Baskin, I. I.; Antipin, I. S.; Varnek, A. Prediction of Tautomer Equilibrium Constants Using Condensed Graphs of Reaction. In *Second Kazan Summer School on Chemoinormatic*; Kazan, Russia, 2015; p 34.

(28) Horvath, D.; Marcou, G.; Varnek, A.; Kayastha, S.; de la Vega de León, A.; Bajorath, J. Prediction of Activity Cliffs Using Condensed Graphs of Reaction Representations, Descriptor Recombination, Support Vector Machine Classification, and Support Vector Regression. *J. Chem. Inf. Model. 56* (9), 1631–1640.

(29) Latino, D. A. R. S.; Aires-de-Sousa, J. Classification of Chemical Reactions and Chemoinformatic Processing of Enzymatic Transformations. *Methods Mol. Biol.* **2011**, *672*, 325–340.

(30) Madzhidov, T. I.; Afonina, V.; Delova, A.; Mukhametzyanova, D.; Nugmanov, R. I.; Mukhametgaliev, R.; Klimchuk, O.; Varnek, A. Artificial Neural Networks Model for Assessment of Optimal Conditions of Hydrogenation Reactions. In *In 22nd European Symposium on Quantitative Structure-Activity Relationships.*; Thessaloniki, Greece, 2018; p 186.

(31) Marcou, G.; Aires de Sousa, J.; Latino, D. A. R. S.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. Expert System for Predicting Reaction Conditions: The Michael Reaction Case. *J. Chem. Inf. Model.* **2015**, *55* (2), 239–250. https://doi.org/10.1021/ci500698a.

(32) Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A. CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. *J. Chem. Inf. Model.* **2019**, *59* (6), 2516–2521. https://doi.org/10.1021/acs.jcim.9b00102.

(33) Sumiya, Y.; Maeda, S. Rate Constant Matrix Contraction Method for Systematic Analysis of Reaction Path Networks. *Chem. Lett.* **2020**, *49* (5), 553–564. https://doi.org/10.1246/cl.200092.

(34) Neo4J Graph Database https://neo4j.com/neo4j-graph-database/.

(35) Plotly Technologies, I. Collaborative Data Science. Plotly Technologies Inc.: Montreal, QC 2015.

(36) Rego, N.; Koes, D. 3Dmol.Js: Molecular Visualization with WebGL. *Bioinformatics* **2015**, *31* (8), 1322–1324. https://doi.org/10.1093/bioinformatics/btu829.