Discovery of Novel Chemical Reactions by Deep Generative Recurrent Neural Network

William Bort[1], Igor I. Baskin[2,3], Timur Gimadiev[4], Artem Mukanov[2], Ramil Nugmanov[2], Pavel Sidorov[4], Gilles Marcou[1], Dragos Horvath[1], Timur Madzhidov[2] and Alexandre Varnek[1,4] *

[1] Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France

[2] Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya str. 18, 420008 Kazan, Russia

[3] Faculty of Physics, M.V. Lomonosov Moscow State University, Leninskie Gory, 119991 Moscow, Russia

[4] Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, 001-0021 Sapporo, Japan

Abstract

Here, we report an application of Artificial Intelligence techniques to generate novel chemical reactions of the given type. A sequence-to-sequence autoencoder with bidirectional Long Short-Term Memory layers was trained on the USPTO reaction database. Each reaction in this database was converted into a single Condensed Graph of Reaction (CGR), followed by their translation into on-purpose developed SMILES/GGR text strings, which are then fed as such to the autoencoder. The autoencoder latent space was visualized on the two-dimensional generative topographic map, from which some zones populated by Suzuki coupling reactions were targeted. These served for the generation of novel reactions by sampling the latent space points and decoding them to SMILES/CGR. Among generated reactions many displayed reaction centers not seen in training set reactions. These pertinent suggestions can be critically analyzed by the expert, cleaned of chemically irrelevant functional groups in order to be experimentally attempted and validated (or discarded), herewith enlarging the synthetic purpose of popular synthetic pathways.

Keywords: chemical reaction generation, sequence-to-sequence autoencoder, deep learning, Condensed Graph of Reaction, Generative Topographic Mapping

Introduction

The discovery of new organic reactions has always been in the focus of synthetic organic chemistry. Each new reaction enriches the arsenal of synthetic tools and opens new horizons in the development and optimization of new drugs and materials. Such reactions are often given the names of their discoverers, which is the highest recognition of their contribution to organic chemistry. Most of the new reactions have been discovered by plain luck, and it has been up to the chemists to notice the discovery and apply their "chemical intuition" to study it in detail.[1] The beginning of a systematic approach to the search for new reactions was laid in 1967 by Balaban, who applied the graph theory for systematical enumeration of pericyclic reactions proceeding through a 6-membered transition state.[2] In the 1970s, these studies were significantly expanded by Hendrickson[3], Arens[4–6], Zefirov, and Tratch[7,8] who considered various formal schemes describing bonds redistribution for different types of pericyclic reactions. Another approach implemented in the IGOR[1,9] and IGOR2[10] programs concerned the algebraic model of constitutional chemistry developed by Dugundji and Ugi[11]. This approach supports the hierarchical representation of organic reactions and deals explicitly with heteroatoms and charges, keeps track of rings in molecules.[10] Its application led to the discovery of previously unknown reactions: the thermal decomposition of α-formyl-oxy ketones,[1,9] and the formation of a cage molecule from N-methoxycarbonyl homopyrrole and tropone.[10] Then, an alternative method based on the generation of the complete sets of non-isomorphic spanning subgraphs of a given graph was suggested. With the help of this approach, new carbene reaction[12] and two new elimination reactions leading to the formation of synthetically important dienes [13] were discovered. The formal-logical approach to organic reactions [7] implemented in the SYMBEQ[14] and ARGENT[15,16] software was used to discover substituted furans.[14]

Despite great expectations, no significant progress in computer-aided reaction design was achieved; approaches, algorithms, and software tools reported so far have not found any widespread popularity among organic chemists. The work with those tools required both extensive knowledge in synthetic organic chemistry and a well-developed intuition in order to turn abstract schemes of bonds redistribution into specific chemical reactions with particular reagents, catalysts, and experimental conditions. This explains why all reactions computationally discovered so far were relatively simple (mainly thermal pericyclic reactions).

We believe that real progress in the discovery of new chemical reactions can be achieved with the help of deep learning methods supported by big data.[17] Recently, Segler et al. reported a chemical synthesis planning system based on the use of the deep neural networks and symbolic AI trained on a big collection of known synthetic reactions.[18] This tool, however, implements automatic extraction of transformation rules (patterns) from known chemical reactions and therefore, in principle, cannot lead to the discovery of new chemical reactions. New reactions might, in principle, be discovered using template-free approaches in which reaction products are directly related from reactants or *vice versa*. Such template-free approaches were successfully implemented in recurrent neural networks operating in sequence-to-sequence mode,[19] in which SMILES of products were directly predicted from SMILES of reactants [20,21] and *vice versa* [22,23]. This approach, however, is narrowly aimed at predicting reactants for given products or products for given reactants, and, therefore, can hardly be used in the discovery of new types of chemical reactions.

Generative models based on recurrent deep neural networks are successfully used to generate novel chemical structures.[24–31] Recently, we have demonstrated that the structures of molecules possessing desirable properties could be generated using a combination of autoencoder with Generative Topographic Map built on the latent vectors. [26] In order to apply this approach to chemical reactions, the latter must be encoded by SMILES strings. However, conventional reaction SMILES can hardly be used because: (*i*) they are too long and have more complicated semantics since every atom is present in both reactants and products; (*ii*) introduction of atom-to-atom mapping (AAM) needed to the identification of the reaction type makes this semantics even more complex. In this case, the autoencoder needs to learn not only semantics and syntax of SMILES but also the AAM rules.

Earlier, we have demonstrated that processing information on chemical reactions complexity can significantly be simplified by the Condensed Graph of Reaction (CGR) approach,[32] in which 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 chemical transformations. 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.[33] Thus, this approach was successfully applied to similarity searching in reaction databases,[32,34] building quantitative structure-reactivity models,[35–38] assessment of tautomer distributions,[39,40] prediction of activity cliffs,[41] classification of enzymatic transformations,[42] prediction of reaction conditions,[43,44] etc. Here, for the first time, we describe dedicated SMILES strings encoding CGRs (SMILES/CGR). Unlike the canonical reaction equation, each atom in CGR is present only once, which significantly reduces the length of the string. Moreover, the CGR (and, hence, SMILES/CGR) contains information about the reaction center and its close neighborhood[45]. The conventional reaction equation can be easily derived from CGR.

The key idea of the present paper is to perform the generation of chemical reactions by generating their CGRs using methods commonly used for generating chemical structures. The plan of the article is the following. First, we describe the methodology of CGR, CGR-based SMILES representation for reactions, and implementation of the autoencoder model. Then, we analyze and discuss the reactions generated this way. We focused on the generation of Suzuki coupling reactions because of their importance and popularity in chemical synthesis.

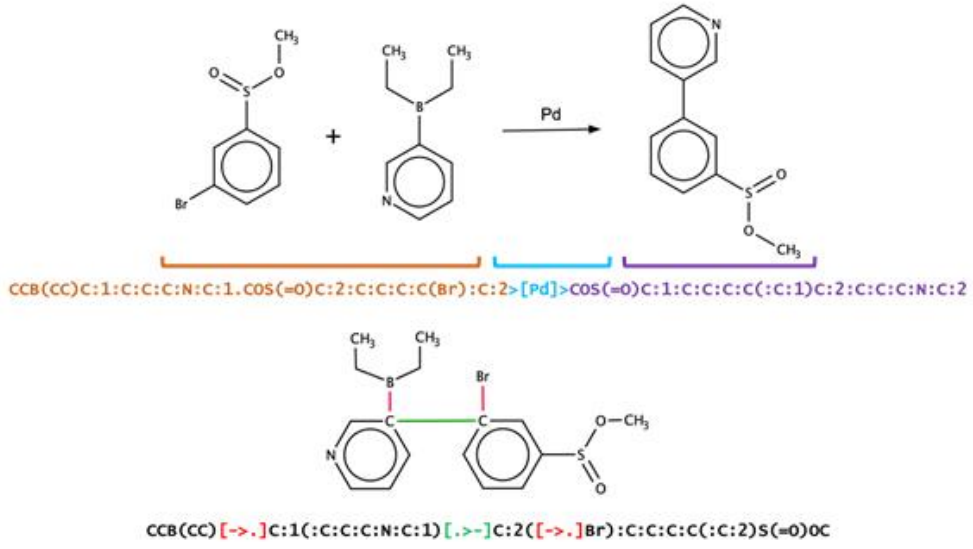

CH3
O
S
O
Br
+
CH3
CH3
B
N
Pd
N
O
S
O
CH3
CCB(CC)C:1:C:C:C:N:C:1.COS(=O)C:2:C:C:C:C(Br):C:2>[Pd]>COS(=O)C:1:C:C:C:C(:C:1)C:2:C:C:C:N:C:2
CH3
CH3
B
Br
C
C
N
O—CH3
S
O
CCB(CC)[->.]C:1(:C:C:C:N:C:1)[.>-]C:2([->.]Br):C:C:C:C(:C:2)S(=O)OC

Figure 1. Example of Suzuki coupling reaction (*top*) and related Condensed Graph of Reaction (CGR, *bottom*). Reaction SMILES and SMILES/CGR are given under reaction equation and CGR, respectively. Reaction SMILES consists of three parts corresponding to reactants (highlighted in orange), reagents (such as catalysts or solvent, in blue), and products (in purple). Here, atom-to-atom mapping is not included in the reaction SMILES. SMILES/CGR contains special features to label dynamic bonds and atoms characterizing chemical transformations. Thus, broken single bonds are encoded as [->.] (in red), while the created C-C bond is encoded as [.>-] (in green). The *CGRTools* library represents aromatic bonds as a colon (:) in both formats.

Results and discussion

Modeling and postprocessing workflows

The modeling workflow included a sequence-to-sequence neural network with Bidirectional Long Short-Term Memory layers trained on special SMILES strings for Condensed Graph of Reaction (SMILES/CGR). Generative Topographic Mapping (GTM) was used to visualize the autoencoder latent space on the two-dimensional map on which the areas mostly populated with Suzuki coupling reactions were detected (Figure 3). Then, the targeted map zone was used to generate virtual chemical reactions by sampling associated latent space points and decoding them to SMILES/CGR. Recently, a similar workflow was successfully used for the generation of novel molecular structures possessing desirable biological activity [26].

The SMILES/CGR strings were prepared using the in-house CGRtools library[45]. Unlike conventional reaction SMILES strings, SMILES/CGR depicts a pseudo-molecule with extended features corresponding to dynamic bonds and dynamic atoms, see Figure 1. A detailed description of SMILES/CGR features is given in the Methods section.

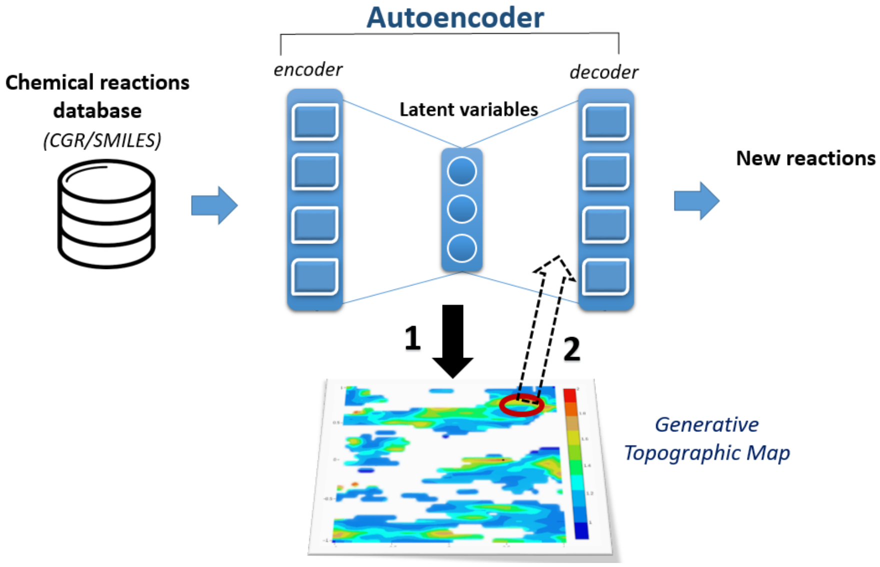


Figure 2. The modeling workflow used in this work.

A set of 2.5 million reactions, extracted and curated from USPTO database [46], was transformed to CGRs and then into SMILES/CGR strings used to feed the autoencoder. The latter was trained on 2 M reactions and validated on 0.45 M reactions. The reconstruction rate was 98.4% and 97.8% at the training and validation stage, respectively. The latent vectors for 100 000 randomly selected reactions were used to construct a Generative Topographic Map (GTM) using in-house software [47]. Then the entire USPTO database was projected onto the map, on which several zones predominantly populated by Suzuki reactions were identified, as it is shown in Figure 3. Random latent vectors were sampled from one of these zones with the highest relative population of Suzuki reactions. As expected, the sampling procedure led to virtual transformations of similar type.

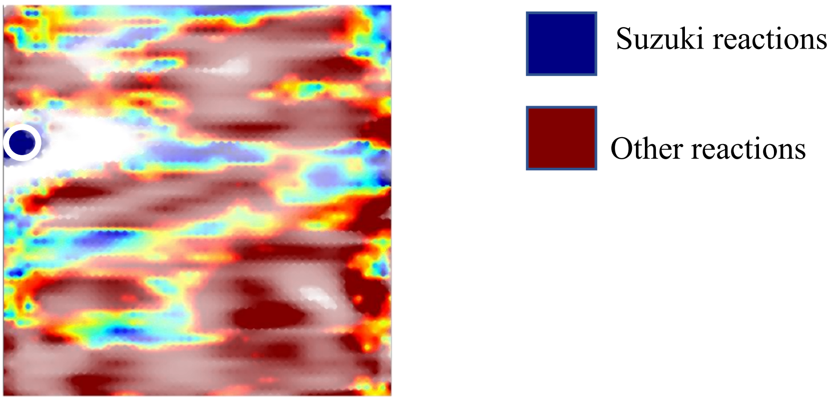


Figure 3. Generative Topographic Map of USPTO reactions encoded by the autoencoder latent variables. Larger transparency levels correspond to smaller data density. The color code characterizes different classes of reactions. Thus, zones in dark blue are exclusively populated by Suzuki reactions, zones in brown are exclusively populated by other types of reactions; in yellow, green, etc zones, the

mixed population is observed. The white circle indicates a zone from which virtual Suzuki reactions were sampled. More

In total, 10,000 text strings have been generated, 7226 (72.2%) of which were valid SMILES/CGR. Valence and aromaticity checking procedures implemented in CGRtools.v3 reduced this number to 1099 structurally correct reactions, out of which 466 "novel" reactions were found (Figure 4).

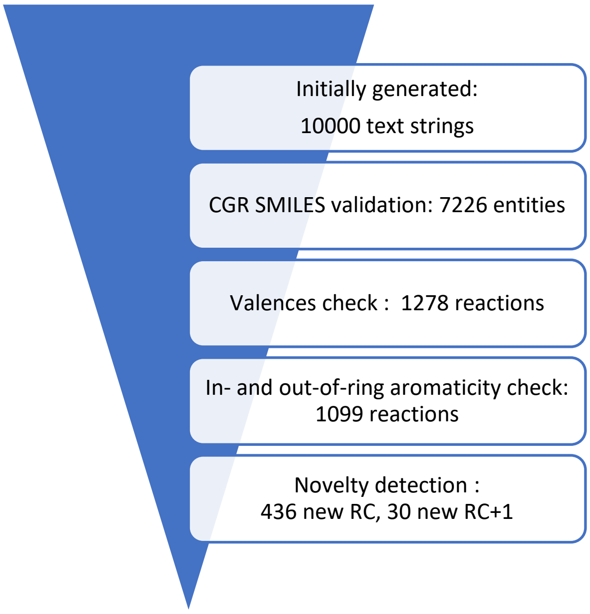

Initially generated:
10000 text strings
CGR SMILES validation: 7226 entities
Valences check : 1278 reactions
In- and out-of-ring aromaticity check:
1099 reactions
Novelty detection :
436 new RC, 30 new RC+1

Figure 4. Postprocessing workflow for the SMILES/CGR generated by the trained autoencoder. This included: (*i*) SMILES validation consisting of the check for SMILES string errors (unclosed cycles, bonds not terminating with atoms, etc.) (*ii*) valences check including aromaticity verification, (*ii*) in- and aromaticity check, and (*iv*) the novelty detection procedure. Steps (*i*)- (*ii*) were performed with the help of the CGRtools.v3 library.

Reaction Novelty Analysis

The main interest of reaction generation is the detection of novel reactions among those generated by the model. However, unlike individual compounds, where novelties can be identified as unique scaffolds or particular structural motifs[26], the definition of reaction novelty was not discussed in the literature. The most descriptive part is the reaction center (***RC***), i.e. atoms and bonds directly involved in the transformation. Thus, we consider two types of reaction novelty: (*i*) the reaction center is unknown (not present in training set); (*ii*) reaction center is known, but its closest neighborhood (1st atoms and bonds near the RC, ***RC+1***) is new. The latter can be extended to a more distant neighborhood (*n* atoms and bonds away, ***RC+n***), but in this work, we only focus on the reaction center and the closest neighbors. To decide whether a reaction is novel, these substructural reaction motifs are encoded by a hashing function as reaction signatures and are compared to all signatures extracted from the initial dataset.

Analysis of generated reactions.

Among 1099 reactions selected using the post-processing workflow (Figure 4), 436 contain new reaction center ***RC*** and 30 reactions contain known ***RC*** but with new first neighborhood ***RC+1***. Some generated reactions have two or more distinct reaction centers, i.e. represent multistep transformations. Note that "novelty" defined as absence from the training set data is *per se* meaningful, as an illustration of the "creativity" of this Artificial Intelligence, *i.e* the ability of the tool to "extrapolate" to configurations that are formally correct – *i.e.* to generate really original configurations which can be submitted for empirical feasibility assessment to human experts. Unfortunately, "novelty" as absence from both the training set and public reaction databases is not easy to interpret, for it may both mean that (a) such reactions were tried, but fail and thus were not published or (b) reactions were never explored, thus represent a real asset of innovation. The choice not to publish failed reactions is a major drawback in training reactivity models [44].

Reactions with new reaction centers

Three interesting "Suzuki-like" subtypes of reactions with new reaction centers (***RC***) have been suggested by the tool:

1. C-Si coupling reactions (**1-4**, Table 1). The substrate in reaction **1** has a Si = N bond experimentally observed by Wiberg et al [48]. This may not be of direct synthetic interest, for such compounds are very unstable, but it has the merit illustrate the "creativity" of the tool, in the above-mentioned sense. Substrates in **2** and **3** have Si-Br bond. A similar reaction with the Si-Cl bond was mentioned in the patent by Kim et al [49] (Figure).

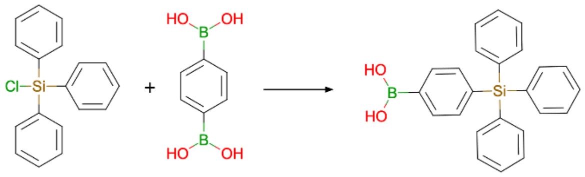


Figure 5. Example of Suzuki reaction with the substrate bearing Si-halogen bond [49]

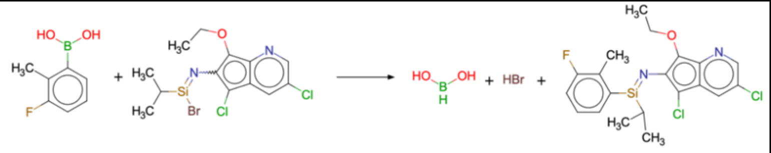
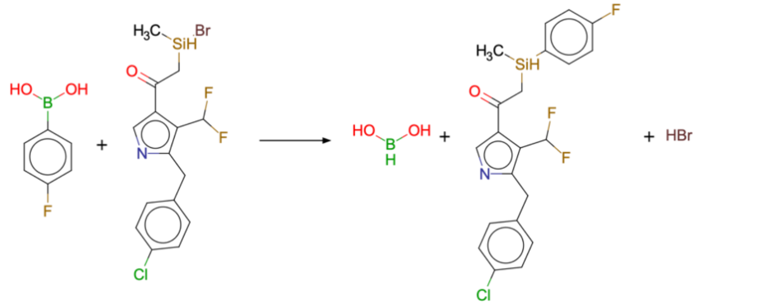
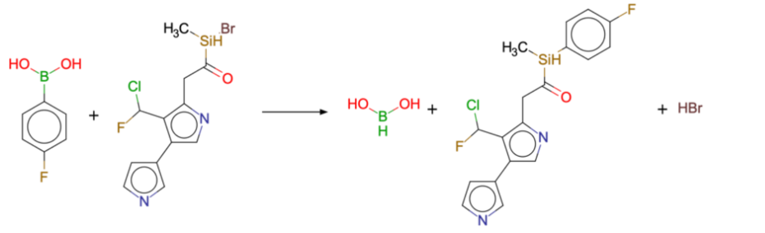
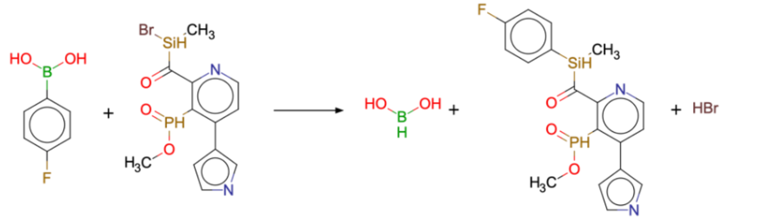
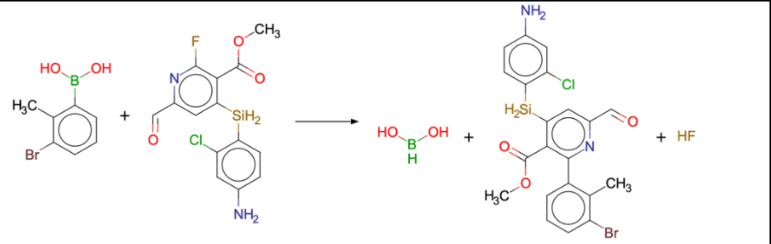
Although a substrate in reaction **4** bearing RP(H)(O)=O group looks unstable due to easy oxidation, we found in Reaxys more than 5000 compounds with a similar functional groups.

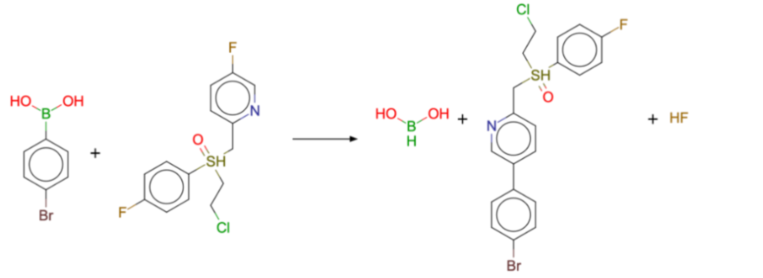
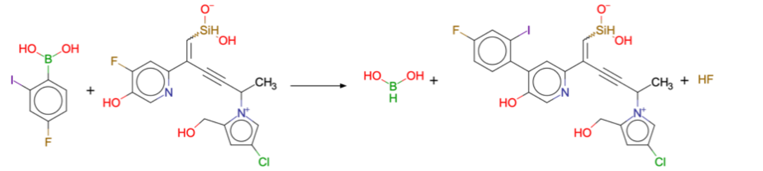
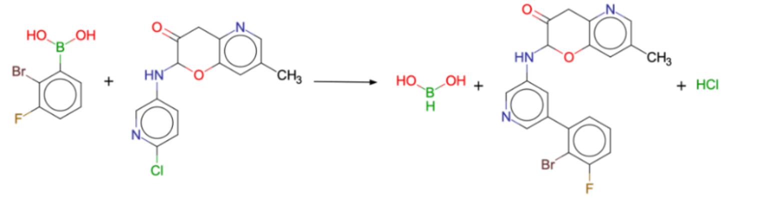
2. Suzuki coupling reactions with unconventional leaving groups, reactions **5-7** (Table 1). In these reactions, fluorine is a leaving group, while no examples of such reactions were present in the initial dataset. Suzuki reactions with fluorine as a leaving group [51] are known, however, Cl and Br are clearly more reactive Note that in some cases the proposed reaction suggests the substitution of F even though more reactive halogens exist elsewhere in the species. This not unexpected from a stochastic navigator focused on a generic "Suzuki-like" zone of the reaction space, not including any chemical "intelligence" beyond whatever could be learned from the training pool of reactions (where F-substitutions were absent altogether). Given the focus on the latent space zone, suggested CGRs should display Suzuki-like (but occasionally original) reaction centers. The RS(H)(R)(R)(=O) group in reaction **6** apparently including a hexacoordinated S atoms makes sense if it is

interpreted as the (hypothetic) conjugated acid of (putatively real) sulfoxonium compound (the error of the decoder being the failure to add formal charges). On the other hand, reactions leading to products with RSi(O$^-$)(OH)H group (reaction **7**) were reported in the literature.[52]

3. Reaction **8** (Table 1) represents a typical error in the generated reactions: an incorrect placement of the substituent in the product which leads to two separate reaction centers (formally, the -Cl "migrates" from ortho to para position with respect to the pyridine N, prior to reaction). This reflects a common error of aromatic atoms assignment for Suzuki reactions in atom-to-atom mapping. For this reason, in some training set reactions, two reaction centers instead of a single-center one were detected. This shows the impact of AAM on the generation of valid reactions.

Table 1. Selected virtual reactions with a new reaction center.

N	Reaction	Reaction center SMILES
1		`BC.[Si]Br>>C[Si].Br.B`
2		`BC.[Si]Br>>C[Si].Br.B`
3		`BC.[Si]Br>>C[Si].Br.B`
4		`BC.[Si]Br>>C[Si].Br.B`
5		`BC.CF>>B.CC.F`

6		BC.CF>>B.CC.F
7		BC.CF>>B.CC.F
8		BC.C>>CC.B and CCl>>C.Cl

It should be noted that most of problems related to reactions **1-7** are related to the presence of reactive groups in organohalide substrate. However, if one focuses just on the reaction center, generalized equations for reactions **1-4** and **5-7** look quite reasonable (Figure 6). Several examples of these types of reactions were found in SciFinder Scholar.

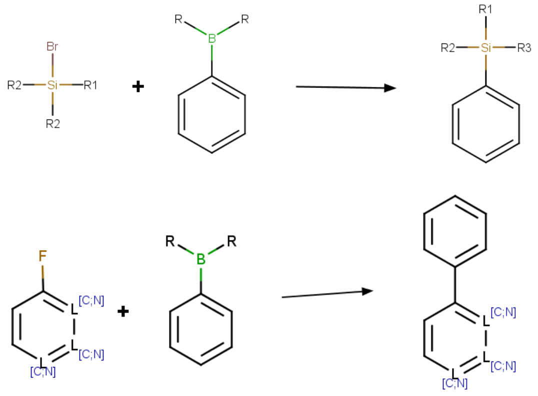


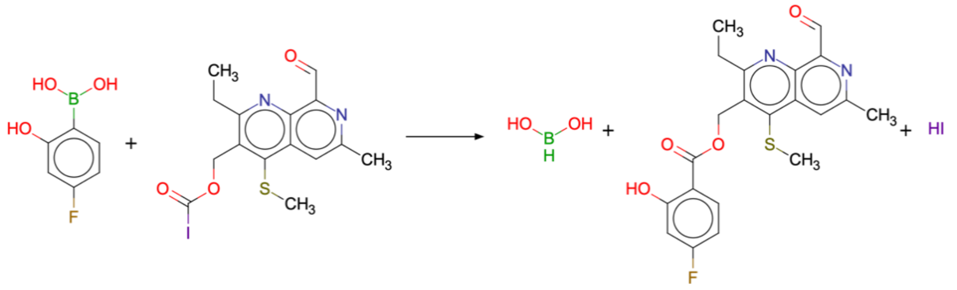
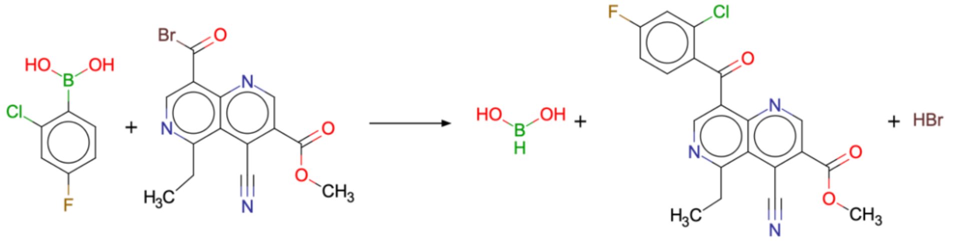
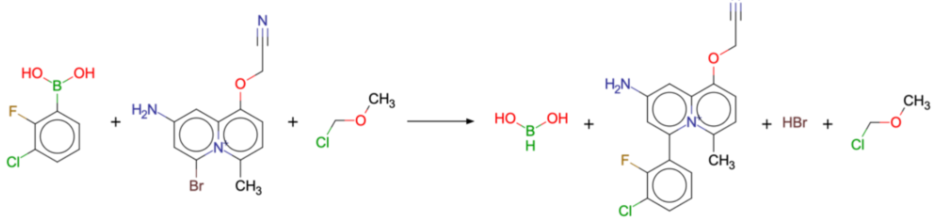
Figure 6. Generalized equations of reactions **1-4** (on the top) and **5-7** (on the bottom) in Table 1

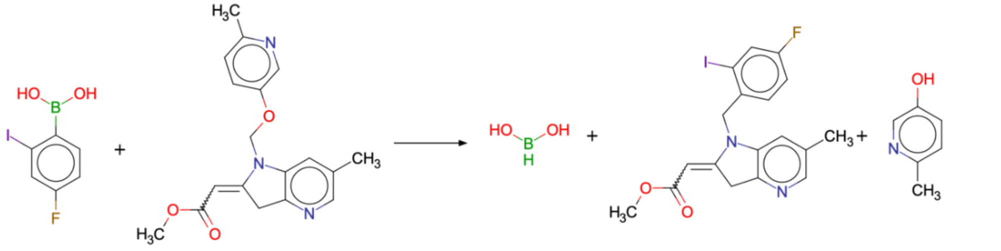
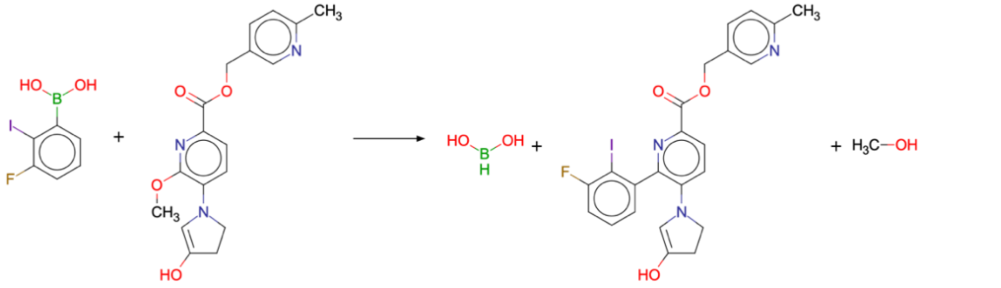
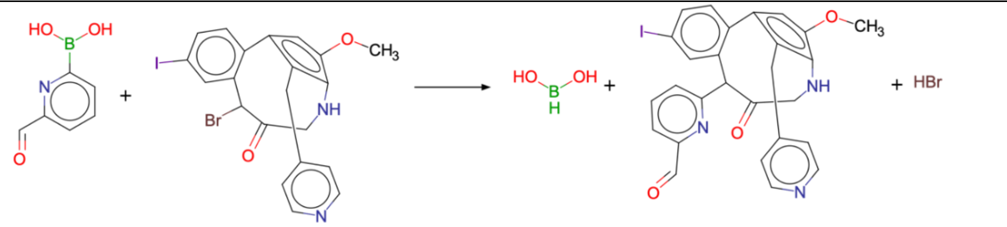
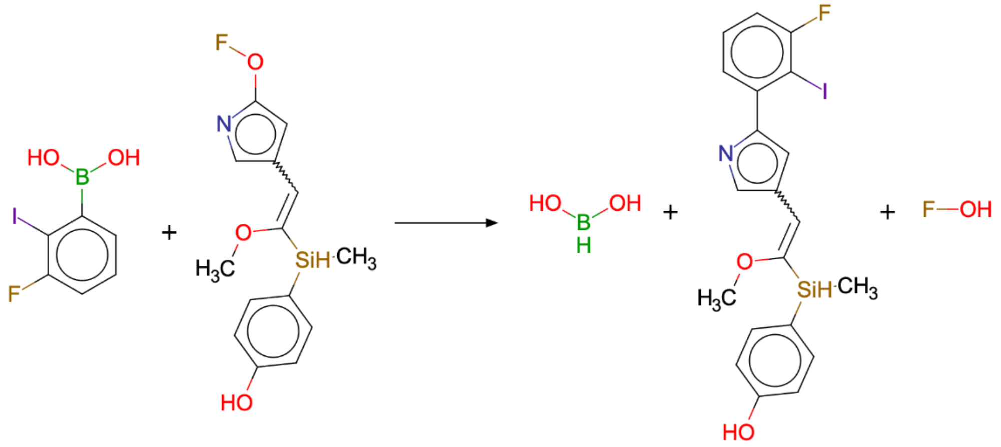
Reactions with a new environment of known reaction centers (RC+1)

Following the novelty detection procedure, 30 reactions that have known reaction centers but an original first environment (RC+1) were detected. Among those, seven reactions have a single reaction center (Table 2). All these reactions can be divided into 4 groups:

1. Suzuki reaction with iodine in an unconventional environment. Reaction **1** looks reasonable, except for an unstable iodoformate group. However, examples for compounds with common formula RCOC(=O)I were found in the Reaxys and PubChem databases and the compound CC(C)(C)COC(=O)I is commercially available (Achemica, “carboniodidic acid, 2,2-dimethylpropyl ester”).
2. Suzuki reaction with bromine in an unconventional environment. In reaction **2**, the BrC(=O)R group is involved. Reaction **3** involves bromine attached to an unusual heterocycle with charged nitrogen. The latter, however, was described in the literature.[53]
3. Suzuki reaction with substitution of O-R group. Generally, such reactions are well studied. Reactions **4-5** are correct from a formal point of view but unfeasible since both leaving groups can hardly be cleaved, 3-hydroxypyridine derivative can serve as a weak leaving group in reaction **4** but methylate in reaction **5** is too weak. Regioselectivity is also unfavorable: the substitution will likely occur on I (reactions **4** and **5**) than involve sp^3-C (reaction **4**) and replace methylate (reaction **5**).
4. Generated structures look unstable or even impossible. Thus, reaction **6** involves a sterically hindered 9-membered ring. In reaction **7**, RO-F bond is extremely unlikely for organic compounds.

Table 2. Examples of generated Suzuki reactions with the **RC** found in the training set. All examples are divided into 4 types: (1) substitution of Iodine; (2) substitution of Bromine; (3) substitution of O-R group; (4) reactions involving reactants with irrealistic structure.

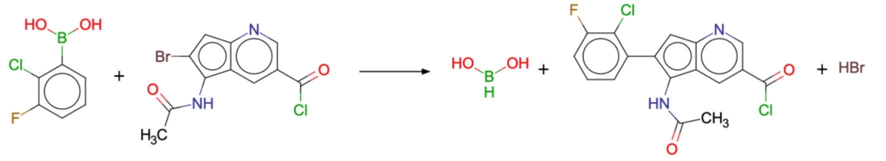
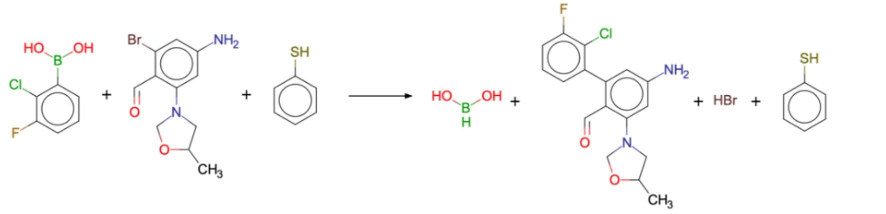
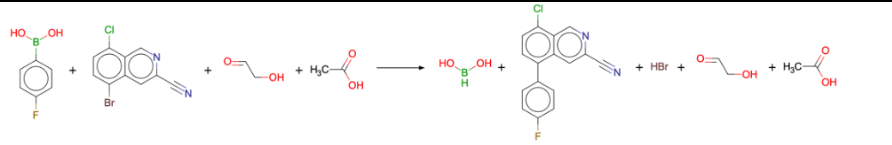
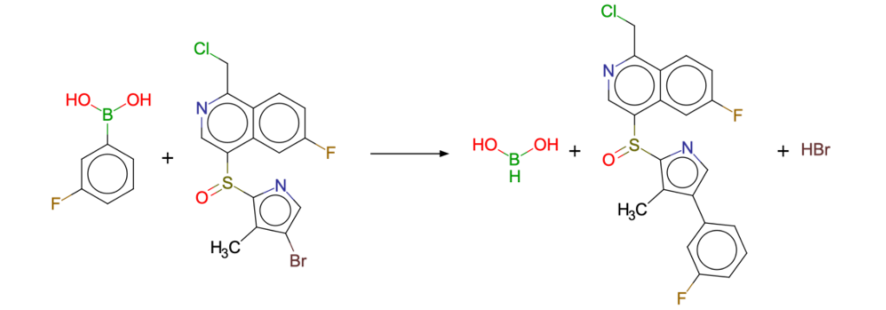
N	Structure	Type
1		1
2		2
3		2

4		3
5		3
6		4
7		4

Reactions with known RC and ***RC+1.***

All reactions with known ***RC*** and ***RC+1*** belong to the Suzuki coupling type, as exemplified in Table 3. For them, the percentage of unstable structural moieties was much lower than for novel reactions. Classical -Cl and -Br substitutions present in the training data are sufficiently well represented for the neural network to tentatively learn selectivity rules. Unlike with the original outputs proposing F substitution (*vide supra*), in the presence of concurrent groups (Br and Cl in Reaction **1**, **3** and **4**) the preferred regioselectivity is predicted correctly.

Table 3. Examples of generated reactions with reaction centers and their first environment present in the training set.

N	Reaction	Reaction center
1		`BC.CBr>>CC.B.Br`
2		`BC.CBr>>CC.B.Br`
3		`BC.CBr>>CC.B.Br`
4		`BC.CBr>>CC.B.Br`

Conclusions

Here we present the first attempt to generate new chemical reactions using a combination of Condensed Graph of Reaction, Generative Topographic Mapping, and sequence-to-sequence autoencoder. In order to feed the autoencoder, special reaction SMILES strings (SMILES/CGR) were suggested. Among generated Suzuki coupling reactions, some species have particular structural motifs (reaction center solely of the reaction center with its close environment) which don’t occur in the training set reactions. These reactions look feasible for real synthesis.

The generative model’s “creativity” depends on the training set. If the latter contains erroneous structures or its size is not big enough, the model may produce wrong or unstable structures. Even if a SMILES/CGR syntax is correctly learnt, the model is not able to capture important information about chemical reactivity. The latter also leads to generation of synthetically unfeasible species. However, if one focuses not on exact reaction equations, but on the novelties like new reaction center or reaction center with its environment, this opens a way to discover new types of chemical reactions.

Method

Datasets and data curation

The dataset we use in this project comes from United States Patents and Trademark Office database (1976 to 2016) extracted by Lowe[46]. It contains about 3.5 million reactions. The initial dataset was preprocessed with *in-house* scripts based on the CGRtools library.[45] The curation includes the standardization (aromatization and functional group standardization), removal of empty reactions (those where the products and reactants are exactly the same, or no reactants or products are recorded) and reactions with valence errors. For curated reactions, atom-to-atom mapping (AAM) was performed using the ChemAxon Automapper tool which is a part of the JChem toolkit.[54] Mapped reactions were converted into CGRs and their reaction centers were extracted with the CGRtools. In total, 165 879 different reaction centers were obtained. Since AAM errors lead to incorrect reaction centers, which are usually rare, only highly populated reaction centers were selected. Thus, the resulting dataset consisted of some 2.5 million reactions (approximately 70% of the initial dataset) which corresponds to 300 most frequent reaction centers.

Reaction data treatment

CGRtools library (version 3)[45] was used for the reactions cleaning, their transformation to CGRs, conversion of CGRs into SMILES/CGR and processing of generated SMILES/CGR back into reactions.

SMILES/CGR notation

Generally, SMILES/CGR follows the OpenSMILES rules[55]. Unlike regular Daylight SMILES, in OpenSMILES, the ring closure number is given after bond order – it gives an opportunity to easily operate with more than 9 rings and has two-digit ring closure symbols. In the given version of SMILES/CGR, instead of specification of aromatic atoms in lowercase, we used colon for aromatic bonds; the aromatic atoms are given in uppercase. It reduces the diversity of symbols and is much more convenient for the specification of aromatic atoms involved in reaction centers in CGR, especially in cases when aromatic atoms are changed to aliphatic or *vice versa*.

Upon SMILES/CGR generation, the following convention is used: any expression given in squared brackets is considered one symbol. Thus, two symbol atoms (e.g. [Co]), charged atoms [N+], etc. are considered as special types of atoms. This convention is used by the tokenizer, and it also reduces the complexity of SMILES/CGR generation by autoencoder.

Dynamic bond labels and dynamic atoms are also specified within squared brackets. Dynamic bonds in CGR have special labels representing changes in bond orders. The list of available dynamic bond labels corresponding to bond order changes is given in **Erreur ! Source du renvoi introuvable.**. Dynamic atom corresponds to change of formal charge or radical state of this atom in reaction. Their labels are also given in brackets, including the atom symbol and text keys for atomic property in reactant and product, separated by symbol >. The list of available text keys is given in Table 5. For a neutral atom A gaining a positive charge *+n* in reaction dynamic atom will be encoded as [A0>+*n*]. In the case of charges +1 and -1, the number 1 is omitted. Properties for charges and radicals may be combined consecutively within one pair of brackets, e.g. [A0>-^>*] stands for an atom which becomes an anion-radical. An example of a Suzuki reaction encoded as SMILES/CGR together with the corresponding reaction is given in **Erreur ! Source du renvoi introuvable.**(b).

Table 4. SMILES/CGR text representations for bonds. The corresponding bond type in reactants is shown in rows (From), and the bond type in products is in columns (Into). The main diagonal represents the non-dynamic bond text key for the corresponding bond type.

Into From	No bond	Single bond	Double bond	Triple bond	Aromatic bond	Any bond
No bond	.	[.>-]	[.>=]	[.>#]	[.>:]	[.>~]
Single bond	[->.]	-[1]	[->=]	[->#]	[->:]	[->~]
Double bond	[=>.]	[=>-]	=	[=>#]	[=>:]	[=>~]
Triple bond	[#>.]	[#>-]	[#>=]	#	[#>:]	[#>:]
Aromatic bond	[:>.]	[:>-]	[:>=]	[:>#]	:	[:>~]
Any bond	[~>.]	[~>-]	[~>=]	[~>#]	[~>:]	~

[1] Usually omitted.

Table 5. Text keys for properties of dynamic atoms in SMILES/CGR.

Property	Uncharged	Positively charged	Negatively charged	Non-radical	Radical
Text key	0[1]	+n	-n	^	*

[1] Omitted for conventional atoms, used only for dynamic atoms.

SMILES/CGR generation and parsing, including preparation of canonic SMILES/CGR, are implemented into CGRtools Python library [45]. Since generation rules of molecular SMILES represent a subset of the SMILES/CGR rules, the same algorithm was used for SMILES and SMILES/CGR.

Reaction generation algorithm

The network architecture previously applied for molecular SMILES generation[26] has been used in this study. It is based on the autoencoder architecture introduced by Xu et al.[56]. SMILES/CGR transformed into sequences of one-hot encoded characters with padding to constant length (256) were used to feed the encoder. Symbols within square brackets (conventional or dynamic atoms or dynamic bonds) were considered as a single symbol within tokenization. The encoder consists of two bidirectional Long Short-Term Memory (LSTM) layers (128 nodes each), while the decoder is composed of two forward LSTM layers (256 nodes each). The bottleneck dense layer between the encoder and the decoder transforms the states of the encoder LSTMs into latent variables to subsequently feed them to the decoder; it consists of 128 nodes. Finally, the decoder outputs are transformed back to one-hot encoded characters via a single dense layer.

To generate latent variable vectors for eventual decoding, we use the Generative Topographic Mapping method. It is a non-linear dimensionality reduction method that has been successfully used for chemical space analysis,[47,57–64] comparison of chemical libraries,[65] building classification[37,57–59,62,66] and regression[67,68] models via activity landscapes, as well as for solving the "inverse" QSAR problem[69]. The GTM algorithm operates by embedding a nonlinear two-dimensional manifold into a D-dimensional descriptor space and calculating the distribution of objects of initial space on these two dimensions. In this work, we utilize the autoencoder's latent vectors as an initial descriptor space. Once a map for the entire USPTO database was constructed, the zones corresponding to the desired reaction type (Suzuki reaction) were located, from which the latent vectors for virtual reactions were sampled. These new vectors fed the trained decoder resulting in new SMILES/CGR strings.

Novelty detection

Novelty detection is based on the comparison of hashed reaction signatures corresponding to reaction centers (**RC**) and their environment between the database of known reaction (here, USPTO database) and the reactions generated by the autoencoder (**Erreur ! Source du renvoi introuvable.**). Encoding chemical reactions by CGR significantly simplifies **RC** detection. Thus, substructural motifs involving the reaction center (**RC**, **RC+1**, **RC+2**, …) can easily be extracted from CGR (see **Erreur ! Source du renvoi introuvable.**). Since any operations with molecular graphs are time-consuming, each substructural motif was encoded by a unique hash code[45] – a reaction signature uniquely identifying given transformation. In this case, the novelty detection is reduced to the comparison of signature (hash code) of a generated reaction with those of known reactions (Figure). The suggested procedure assures fast and precise novelty detection.

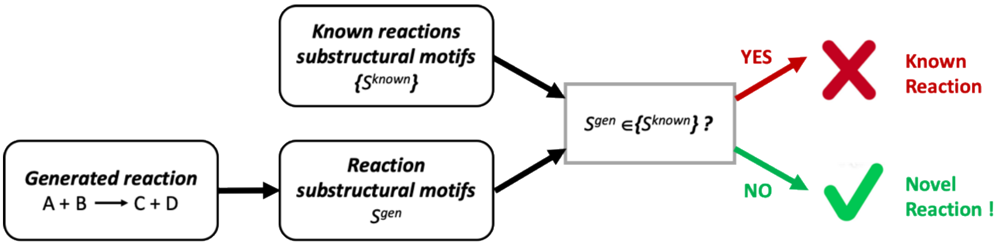

Known reactions substructural motifs {S^{known}}
Generated reaction
A + B ⟶ C + D
Reaction substructural motifs S^{gen}
$S^{gen} \in \{S^{known}\}$?
YES
NO
Known Reaction
Novel Reaction !

Figure 6. Reactions novelty detection workflow. Substructural motifs S^{gen} (**RC**, **RC+1**, **RC+2**, …) are extracted from the query CGR and compared with those for known reactions {S^{known}}. In such a way, motifs belonging to novel reactions will easily be identified.

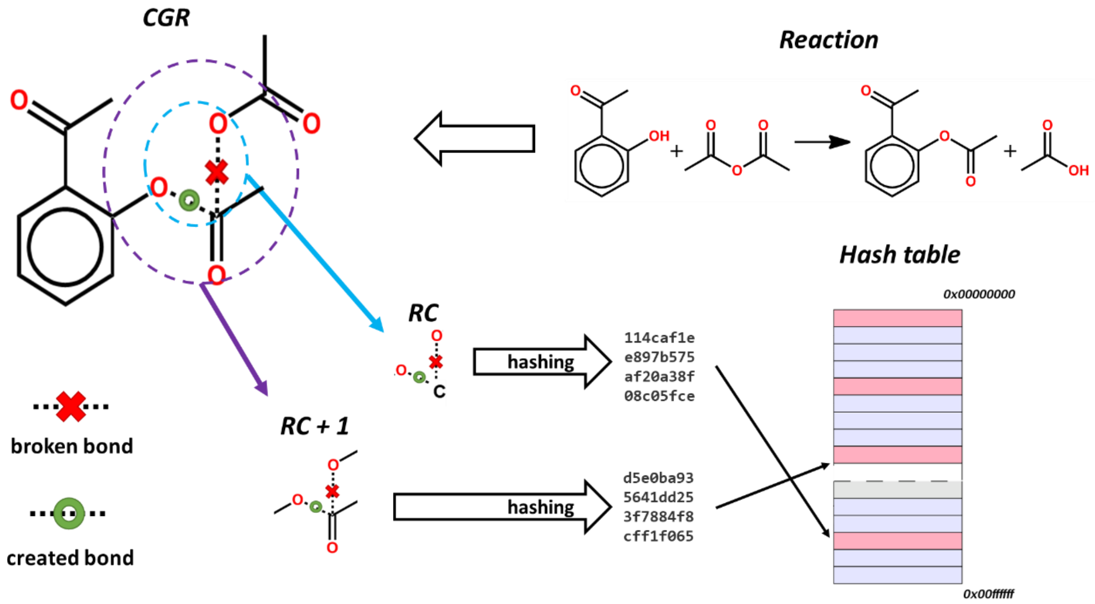


Figure 7. Preparation of a collection of reaction signatures as hash codes. From a CGR generated from a given reaction, substructural motifs containing reaction center (RC), or reaction center with *n* neighboring bonds and atoms (RC+*n*, here *n*=1) can be extracted. Each motif is encoded by a hashing function into a unique hash code – reaction signature. Ensemble of unique hash codes for all reactions in the database is stored in the hash table.

References

1. Herges, R. Reaction planning: Computer-aided reaction design. *Tetrahedron Comput. Methodol.* **1**, 15–25 (1988).
2. Balaban, A. T. Chemical graphs. 3. Reactions with cyclic 6-membered transition states. *Rev. Roum. Chim.* **12**, 875–902 (1967).
3. Hendrickson, J. B. The Variety of Thermal Pericyclic Reactions. *Angew. Chemie Int. Ed. English* **13**, 47–76 (1974).
4. Arens, J. F. A formalism for the classification and design of organic reactions. I. The class of (− +)n reactions. *Recl. des Trav. Chim. des Pays-Bas* **98**, 155–161 (1979).
5. Arens, J. F. A formalism for the classification and design of organic reactions. II. The classes of (+ −)n + and (− +)n − reactions. *Recl. des Trav. Chim. des Pays-Bas* **98**, 395–399 (1979).
6. Arens, J. F. A formalism for the classification and design of organic reactions III. The class of (+ -)nC reactions. *Recl. des Trav. Chim. des Pays-Bas* **98**, 471–483 (1979).
7. Zefirov, N. S. & Tratch, S. S. Formal-Logical Approach to Multicentered Processes with Cyclic Electron Transfer. *Match* 263–264 (1977).
8. Zefirov, N. S., Tratch, S. S. & Trach, S. S. Systematization of tautomeric processes and formal-logical approach to the search for new topological and reaction types of tautomerism. *Chem. Scr.* **15**, 4–12 (1980).
9. Bauer, J., Herges, R., Fontain, E. & Ugi, I. IGOR and computer assisted innovation in chemistry. *Chimia (Aarau).* **39**, 43–53 (1985).
10. Bauer, J. IGOR2: a PC-program for generating new reactions and molecular structures. *Tetrahedron Comput Methodol* **2**, 269–280 (1989).
11. Dugundji, J. & Ugi, I. An algebraic model of constitutional chemistry as a basis for chemical computer programs. in *Computers in Chemistry. Fortschritte der Chemischen Forschung* **39**, 19–64 (Springer , 1973).
12. Herges, R. Reaction planning: prediction of new organic reactions. *J Chem Inf Comput Sci* **30**, 377–383 (1990).
13. Herges, R. & Hoock, C. Reaction planning: computer-aided discovery of a novel elimination reaction. *Science* **255**, 711–713
14. Zefirov, N. S., Baskin, I. I. & Palyulin, V. A. SYMBEQ Program and Its Application in Computer-Assisted Reaction Design. *J. Chem. Inf. Comput. Sci.* **34**, 994–999 (1994).
15. Zefirov, N., Tratch, S. & Molchanova, M. The argent program system: A second-generation tool aimed at combinatorial search for new types of organic reactions. *1 Main concepts potentialities MatchCommunications Math. Comput. Chem. 46* **2002 SRC**, 253–273
16. Molchanova, M. S., Tratch, S. S. & Zefirov, N. S. Computer-aided design of new organic transformations: exposition of the ARGENT-1 program. *J. Phys. Org. Chem.* **16**, 463–474 (2003).
17. Baskin, I. I., Madzhidov, T. I., Antipin, I. S. & Varnek, A. A. Artificial intelligence in synthetic chemistry: achievements and prospects. *Russ. Chem. Rev.* **86**, 1127–1156 (2017).
18. Segler, M. H. S. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604 (2018).
19. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. *Proc. 27th Int. Conf. Neural Inf. Process. Syst. Montr. Canada pp* **2 SRC-B**, 3104–3112 (2014).
20. Nam, J. & Kim, J. Linking the Neural Machine Translation and the Prediction of Organic

Chemistry Reactions. (2016).

21. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
22. Liu, B. *et al.* Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
23. Karpov, P., Godin, G. & Tetko, I. V. A Transformer Model for Retrosynthesis. in 817–830 (2019). doi:10.1007/978-3-030-30493-5_78
24. Xue, D. *et al.* Advances and challenges in deep generative models for de novo molecule generation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **9**, e1395 (2019).
25. Xu, Y. *et al.* Deep learning for molecular generation. *Future Med. Chem.* **11**, 567–597
26. Sattarov, B. *et al.* De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* just accepted (2019). doi:10.1021/acs.jcim.8b00751
27. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
28. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inform.* **37**, (2018).
29. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning:Generative models for matter engineering. *Science* **361**, (2018).
30. Jørgensen, P. B., Schmidt, M. N. & Winther, O. Deep Generative Models for Molecular Science. *Mol. Inform.* **37**, 1700133 (2018).
31. Gupta, A. *et al.* Generative Recurrent Networks for De Novo Drug Design. *Mol. Inform.* **37**, 1700111 (2018).
32. 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* **20**, 253–270 (2011).
33. Varnek, A., Fourches, D., Hoonakker, F. & Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided. Mol. Des.* **19**, 693–703 (2005).
34. Hoonakker, F., Lachiche, N., Varnek, A. & Wagner, A. A Representation to Apply Usual Data Mining Techniques to Chemical Reactions. in 318–326 (2010). doi:10.1007/978-3-642-13025-0_34
35. Madzhidov, T. I. *et al.* Structure-reactivity relationships in terms of the condensed graphs of reactions. *Russ. J. Org. Chem.* **50**, 459–463 (2014).
36. Madzhidov, T. I. *et al.* Structure–reactivity relationship in bimolecular elimination reactions based on the condensed graph of a reaction. *J. Struct. Chem.* **56**, 1227–1234 (2015).
37. Gimadiev, T. *et al.* Bimolecular Nucleophilic Substitution Reactions: Predictive Models for Rate Constants and Molecular Reaction Pairs Analysis. *Mol. Inform.* minf.201800104 (2018). doi:10.1002/minf.201800104
38. Glavatskikh, M. *et al.* Predictive Models for Kinetic Parameters of Cycloaddition Reactions. *Mol. Inform.* **38**, 1800077 (2019).
39. Gimadiev, T. R. *et al.* Assessment of tautomer distribution using the condensed reaction graph approach. *J. Comput. Aided. Mol. Des.* **32**, 401–414 (2018).
40. Gimadiev, T. R. *et al.* Prediction of tautomer equilibrium constants using condensed graphs of reaction. in *Second International School-Seminar 'From empirical to predictive chemistry'* 11 (2015).
41. Horvath, D. *et al.* 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**, 1631–1640 (2016).

42. Latino, D. A. R. S. & Aires-de-Sousa, J. Classification of chemical reactions and chemoinformatic processing of enzymatic transformations. *Methods Mol. Biol.* **672**, 325–340 (2011).
43. Madzhidov, T. *et al.* Artificial neural networks model for assessment of optimal conditions of hydrogenation reactions. in *Abstract book of 22nd European Symposium on Quantitative Structure-Activity Relationships* 186 (2018).
44. Marcou, G. *et al.* Expert System for Predicting Reaction Conditions: The Michael Reaction Case. *J. Chem. Inf. Model.* **55**, 239–250 (2015).
45. Nugmanov, R. I. *et al.* CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. *J. Chem. Inf. Model.* **59**, 2516–2521 (2019).
46. Lowe, D. M. Extraction of chemical structures and reactions from the literature. (University of Cambridge, 2012). doi:https://doi.org/10.17863/CAM.16293
47. Gaspar, H. A. *et al.* Generative Topographic Mapping Approach to Chemical Space Analysis. in 211–241 (2016). doi:10.1021/bk-2016-1222.ch011
48. N.Wiberg *et al.* Preparation and Structure of a Stable Molecule containing a Silicon Nitrogen Double Bond and of its Tetrahydrofuran Adduct. *J. Chem. Soc., Chem. Commun.,* 591 (1986)
49. Kim, S. *et al.* Condensed cyclic compound and organic light-emitting device including the same. US Patent 2016/0072072 A1.
50. Corriu, R. J. P., Mazhar, M., Poirier, M. & Royo, G. Arigid pentacoordinate silicon structure generalization of pseudorotation. *J. Organomet. Chem.* **306**, C5–C9 (1986).
51. Burrueco, M. I., Mora, M., Jiménez-Sanchidrián, C. & Ruiz, J. R. Hydrotalcite-supported palladium nanoparticles as catalysts for the Suzuki reaction of aryl halides in water. *Appl. Catal. A Gen.* **485**, 196–201 (2014).
52. Chandler, M. L., Krahnke, R. H. & LeVier, R. R. Anticonvulsant phenylsilanes. (1976).
53. Jones, G. Aromatic Quinolizines. in 1–62 (1982). doi:10.1016/S0065-2725(08)60395-5
54. ChemAxon. Chemical Structure Representation Toolkit. (2019).
55. James, C. A. OpenSMILES specification. *www.opensmiles.org* (2016). Available at: http://opensmiles.org/opensmiles.html.
56. Xu, Z., Wang, S., Zhu, F. & Huang, J. Seq2seq Fingerprint. in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17* 285–294 (ACM Press, 2017). doi:10.1145/3107411.3107424
57. Gimadiev, T. R., Madzhidov, T. I., Marcou, G. & Varnek, A. Generative Topographic Mapping Approach to Modeling and Chemical Space Visualization of Human Intestinal Transporters. *Bionanoscience* **6**, 464–472 (2016).
58. Klimenko, K., Marcou, G., Horvath, D. & Varnek, A. Chemical Space Mapping and Structure–Activity Analysis of the ChEMBL Antiviral Compound Set. *J. Chem. Inf. Model.* **56**, 1438–1454 (2016).
59. Sidorov, P., Gaspar, H., Marcou, G., Varnek, A. & Horvath, D. Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds. *J. Comput. Aided. Mol. Des.* **29**, 1087–1108 (2015).
60. Maniyar, D. M., Nabney, I. T., Williams, B. S. & Sewing, A. Data Visualization during the Early Stages of Drug Discovery. *J. Chem. Inf. Model.* **46**, 1806–1818 (2006).
61. Owen, J. R., Nabney, I. T., Medina-Franco, J. L. & López-Vallejo, F. Visualization of Molecular Fingerprints. *J. Chem. Inf. Model.* **51**, 1552–1563 (2011).
62. Kireeva, N. *et al.* Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **31**, 301–312 (2012).

63. Glavatskikh, M. *et al.* Visualization and Analysis of Complex Reaction Data: The Case of Tautomeric Equilibria. *Mol. Inform.* **37**, 1800056 (2018).
64. Horvath, D., Marcou, G. & Varnek, A. Generative Topographic Mapping Approach to Chemical Space Analysis. in *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences* 167–199 (2017). doi:10.1007/978-3-319-56850-8_6
65. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. Chemical Data Visualization and Analysis with Incremental Generative Topographic Mapping: Big Data Challenge. *J. Chem. Inf. Model.* **55**, 84–94 (2015).
66. Gaspar, H. A. *et al.* Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS). *J. Chem. Inf. Model.* **53**, 3318–3325 (2013).
67. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. GTM-Based QSAR Models and Their Applicability Domains. *Mol. Inform.* **34**, 348–356 (2015).
68. Baskin, I. I., Solov'ev, V. P., Bagatur'yants, A. A. & Varnek, A. Predictive cartography of metal binders using generative topographic mapping. *J. Comput. Aided. Mol. Des.* **31**, 701–714 (2017).
69. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. Stargate GTM: Bridging Descriptor and Activity Spaces. *J. Chem. Inf. Model.* **55**, 2403–2410 (2015).