



**HAL**  
open science

## Discovery of novel chemical reactions by deep generative recurrent neural network

William Bort, Igor Baskin, Timur Gimadiev, Artem Mukanov, Ramil Nugmanov, Pavel Sidorov, Gilles Marcou, Dragos Horvath, Olga Klimchuk, Timur Madzhidov, et al.

### ► To cite this version:

William Bort, Igor Baskin, Timur Gimadiev, Artem Mukanov, Ramil Nugmanov, et al.. Discovery of novel chemical reactions by deep generative recurrent neural network. *Scientific Reports*, 2021, 11 (1), 10.1038/s41598-021-81889-y . hal-03371808

**HAL Id: hal-03371808**

**<https://hal.science/hal-03371808>**

Submitted on 8 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



OPEN

## Discovery of novel chemical reactions by deep generative recurrent neural network

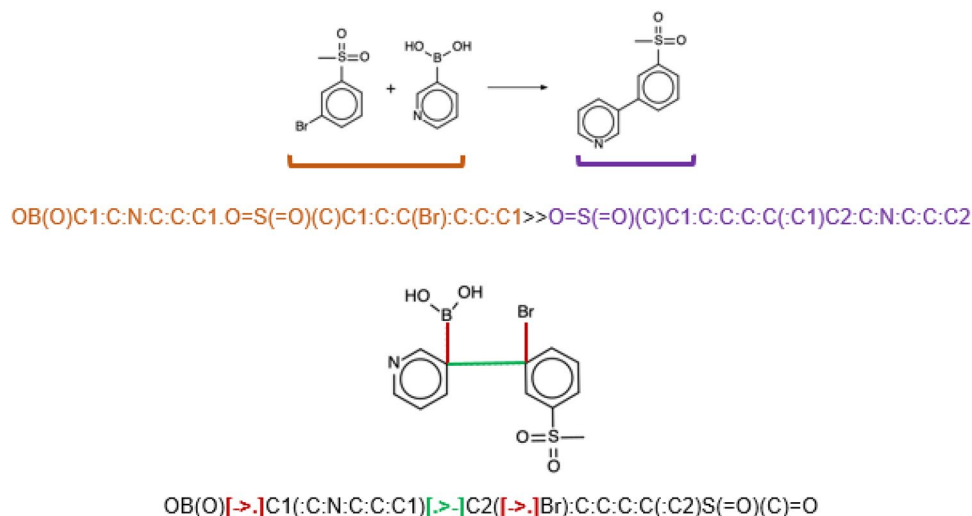
William Bort<sup>1</sup>, Igor I. Baskin<sup>1,2,4</sup>, Timur Gimadiev<sup>3</sup>, Artem Mukanov<sup>2</sup>, Ramil Nugmanov<sup>2</sup>, Pavel Sidorov<sup>3</sup>, Gilles Marcou<sup>1</sup>, Dragos Horvath<sup>1</sup>, Olga Klimchuk<sup>1</sup>, Timur Madzhidov<sup>2</sup> & Alexandre Varnek<sup>1,3</sup>✉

The “creativity” of Artificial Intelligence (AI) in terms of generating de novo molecular structures opened a novel paradigm in compound design, weaknesses (stability & feasibility issues of such structures) notwithstanding. Here we show that “creative” AI may be as successfully taught to enumerate novel *chemical reactions* that are stoichiometrically coherent. Furthermore, when coupled to reaction space cartography, de novo reaction design may be focused on the desired reaction class. A sequence-to-sequence autoencoder with bidirectional Long Short-Term Memory layers was trained on on-purpose developed “SMILES/CGR” strings, encoding reactions of the USPTO database. The autoencoder latent space was visualized on a generative topographic map. Novel latent space points were sampled around a map area populated by Suzuki reactions and decoded to corresponding reactions. These can be critically analyzed by the expert, cleaned of irrelevant functional groups and eventually experimentally attempted, herewith enlarging the synthetic purpose of popular synthetic pathways.

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<sup>1</sup>. 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<sup>2</sup>. In the 1970s, these studies were significantly expanded by Hendrickson<sup>3</sup>, Arens<sup>4–6</sup>, Zefirov, and Tratch<sup>7,8</sup> who considered various formal schemes describing bonds redistribution for different types of pericyclic reactions. Another approach implemented in the IGOR<sup>1,9</sup> and IGOR2<sup>10</sup> programs concerned the algebraic model of constitutional chemistry developed by Dugundji and Ugi<sup>11</sup>. This approach supports the hierarchical representation of organic reactions and deals explicitly with heteroatoms and charges, keeps track of rings in molecules<sup>10</sup>. Its application led to the discovery of previously unknown reactions: the thermal decomposition of  $\alpha$ -formyl-oxy ketones<sup>1,9</sup>, and the formation of a cage molecule from N-methoxycarbonyl homopyrrole and tropone<sup>10</sup>. 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<sup>12</sup> and two new elimination reactions leading to the formation of synthetically important dienes<sup>13</sup> were discovered. The formal-logical approach to organic reactions<sup>7</sup> implemented in the SYMBEQ<sup>14</sup> and ARGENT<sup>15,16</sup> software was used to discover substituted furans<sup>14</sup>.

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 to turn abstract schemes of bonds redistribution into specific chemical reactions with particular reagents, catalysts, and experimental conditions. This explains why all reactions computationally

<sup>1</sup>Laboratory of Chemoinformatics, UMR 7140 CNRS, University of Strasbourg, 1, rue Blaise Pascal, 67000 Strasbourg, France. <sup>2</sup>Laboratory of Chemoinformatics and Molecular Modeling, Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya str. 18, 420008 Kazan, Russia. <sup>3</sup>Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, Sapporo 001-0021, Japan. <sup>4</sup>Department of Materials Science and Engineering, Technion – Israel Institute of Technology, 3200003 Haifa, Israel. ✉email: varnek@unistra.fr



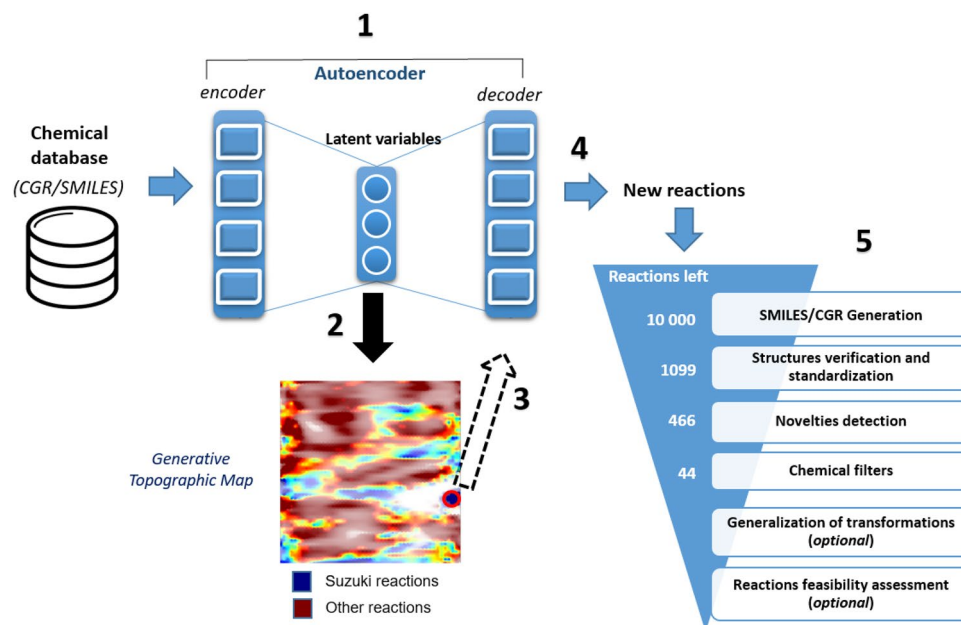
**Figure 1.** An example of Suzuki coupling reaction (*top*) and its condensed graph (CGR, *bottom*). Reaction SMILES and SMILES/CGR are given underneath. The reaction SMILES features reactants (in orange), and products (in purple). Atom-to-atom mapping is not provided. In the SMILES/CGR broken single bonds are encoded as [ $>$ .] (in red), while the created C–C bond is [ $>$ .] (in green). The colon (:) represents aromatic bonds. See Supporting Information for the details.

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 by deep learning from big data<sup>17</sup>. Recently, Segler et al. reported a chemical synthesis planning system based on deep neural networks and symbolic AI trained on a big collection of known synthetic reactions<sup>18</sup>. This tool, however, implements automatic extraction of transformation rules (“templates”) from known chemical reactions and therefore, in principle, cannot “suggest” not yet seen transformations. Several template-free techniques based on recurrent neural networks and transformers were successfully implemented. They operate in sequence-to-sequence translation mode<sup>19</sup>, in which SMILES of products were directly predicted from SMILES of reactants<sup>20,21</sup> and vice versa<sup>22–24</sup>. Interesting chemistry knowledge driven approaches aiming to predict organic reactions outcomes from given reactants were proposed by Coley et al.<sup>25</sup> and in Baldi’s group<sup>26,27</sup>. Although, discovery of new chemical transformations cannot be excluded, this is not an objective of such type of calculations. To our knowledge, no new types of chemical reactions resulted from the “reactants-to-products” models were reported in the literature so far.

Generative models based on recurrent deep neural networks were successfully used to generate novel chemical structures<sup>28–37</sup>. 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<sup>26</sup>. In order to apply this approach to chemical reactions, they must be encoded by SMILES strings. However, conventional reaction SMILES can hardly be used because: (i) they are much longer, and (ii) atom-to-atom mapping (AAM) needed for reaction center identification, adds a further layer of complexity. The autoencoder would have to learn not only semantics and syntax of SMILES but also the AAM rules.

Earlier, we showed that *in silico* chemical reaction handling can be significantly simplified by the Condensed Graph of Reaction (CGR) approach<sup>38</sup>, in which the structures of reactants and products are merged into a single graph (Fig. 1). The CGR edges correspond either to standard chemical bonds or to “dynamic” bonds describing 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<sup>39</sup>. Thus, this approach was successfully applied to similarity searching in reaction databases<sup>38,40</sup>, building quantitative structure–reactivity models<sup>41–44</sup>, assessment of tautomer distributions<sup>45,46</sup>, prediction of activity cliffs<sup>47</sup>, classification of enzymatic transformations<sup>48</sup>, prediction of reaction conditions<sup>49,50</sup>, etc. Here, for the first time, we introduce dedicated SMILES strings encoding CGRs (SMILES/CGR), see their detailed description in Supporting Information. Moreover, the CGR (and, hence, SMILES/CGR) contains information about the reaction center and its close neighborhood<sup>51</sup>.

Basically, CGRs are nothing but “molecules” with “exotic” bond orders for the changing bonds—thus, let us “teach” *de novo* molecular design tools on how to generate new reactions! Following a workflow recently used for the generation of novel molecular structures potentially possessing desirable biological activity<sup>30</sup>, we have chosen to focus here on the generation of “Suzuki-like” putative chemical transformations. The Suzuki coupling reaction was chosen because this reaction is widely used in organic synthesis, and, therefore, its new variants implying different leaving groups and reaction centers could be of interest for synthetic chemists. From the technical point of view, Suzuki reactions constitute a sizeable part of the USPTO database which assures satisfactory knowledge extraction upon the model training. Reaction center of Suzuki reaction can be represented by a SMILES string BC.QL>>B.CQ.L (where **Q** = C, N, O, S, Si and **L** is a leaving group). In our simulation we expect that AI may suggest realistic and unseen combinations of **Q** and **L**.



**Figure 2.** Modeling workflow for generation of new reactions consists of five main steps: (1) training sequence-to-sequence autoencoder on the USPTO database of chemical reactions; (2) building of Generative Topographic Map (GTM) using the autoencoder latent variables and preparation of GTM class landscape; (3) selecting on GTM a zone populated to Suzuki coupling reactions and identification of related autoencoder latent vectors; (4) sampling from the autoencoder latent space and generation of new reactions; and, (5) post-processing step. On the Generative Topographic Map, larger transparency levels correspond to lower density. The color code renders the (binary: Suzuki vs Other) reaction class distribution. Thus, zones in dark blue are exclusively populated by Suzuki reactions, zones in dark red are exclusively populated by other types of reactions; while intermediate colors correspond to reaction space areas hosting both categories, in various ratios. The red circle indicates the zone from which virtual Suzuki reactions were sampled.

A sequence-to-sequence neural network with Bidirectional Long Short-Term Memory<sup>52</sup> layers trained on SMILES/CGR achieved the ability to convert SMILES/CGR to their latent vectors (“encode”) and back (“decode”). Generative Topographic Mapping (GTM) was used to visualize the latent space in 2D and to detect a cluster mostly populated with Suzuki reactions (Fig. 2). Then, virtual chemical reactions were generated by sampling the targeted zone followed by the decoding of associated latent vectors to SMILES/CGR. Notice that visualization is not strictly required for clusters identification, but may significantly help to choose a cluster from which the sampling is performed.

## Results

**Reaction sampling from generative topographic map.** A set of 2 424 306 reactions, extracted and curated from the USPTO database<sup>53</sup>, was rendered as CGRs and then as SMILES/CGR strings used to feed the autoencoder. The latter was trained on some 2 million reactions and validated on 450 thousand reactions. The reconstruction rate (a ratio of correctly reconstructed SMILES/CGR) was 98.4% and 97.8% at the training and validation stage, respectively. This is slightly less than reconstruction rates of plain molecular SMILES by state-of-the-art encoders/decoders, but it can be explained by larger complexity and length of SMILES/CGR and an additional source of error: the errors of atom-to-atom mapping in some entries. SMILES/CGR is intrinsically more difficult to learn, with dynamical bonds, dynamical atoms and formal coordination numbers exceeding atomic valency representing novel degrees of freedom in the syntax. Unbalanced or erroneous entries may pass the standardization protocols and thus negatively impact generated SMILES/CGR quality. Nevertheless, reconstruction rates are robust and although LSTM has relatively short memory compared to some other neural networks architectures like transformers and, therefore, may fail to learn relatively complex structural motifs, the bidirectional LSTM used in our work seems to perform acceptably well.

The latent vectors for 100 000 randomly selected reactions were used to construct a Generative Topographic Map (GTM) using in-house software<sup>54</sup>. Then the entire USPTO database was projected onto the map, on which several zones predominantly populated by Suzuki reactions were identified, as shown in Fig. 2.

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 a similar type. Finally, 10,000 text strings have been generated, followed by their analysis using a complex post-processing protocol (Fig. 2). At the structures verification and standardization stage, the CGRtools.v3 tool was used to discard invalid SMILES/CGR and to perform valence and aromaticity check. This reduced the dataset to 1099 reactions (some 11% of generated text strings) in which structures of reactants and products were correct. This value is similar to that

(15–20%) observed for the SMILES strings in our previous studies devoted to generation of individual molecules. Clearly, not every latent space vector corresponds to a valid structure. However, since invalid SMILES/CGR can be discarded algorithmically, they are not a liability but a manageable consequence of exploratory sampling.

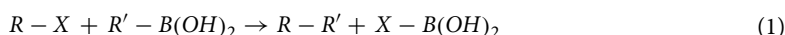
Also, the USPTO reactions are unevenly distributed in terms of types. Deep learning typically focuses on dense clusters allowing it to extensively capture their associated syntactic rules. The reality of different combinations of leaving group **L** and a coupling partner **Q** in the reaction center  $BC.QL > B.CQ.L$  suggested by AI depends not only on the training set size but also on its diversity, i.e., on the presence in the training set related examples. The USPTO dataset contains very few reactions with **Q** = O, S and Si which may explain the relatively high rate of invalid SMILES/CGR strings.

**Reaction novelty analysis.** The main interest of *in silico* reaction generation is the proposal of novel reactions that a human mind would not spontaneously think of. However, unlike individual compounds, where novelties can be identified as unique scaffolds or particular structural motifs<sup>30</sup>, the definition of reaction novelty was not discussed in the literature. The most descriptive part is the reaction center (**RC**)<sup>55</sup>, i.e. atoms and bonds directly involved in the transformation. Thus, we consider two levels of reaction novelty: (i) the reaction center is unknown (not present in the training set); (ii) reaction center is known, but its closest neighborhood (1<sup>st</sup> 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.

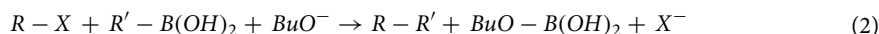
Among 1099 reactions selected using the post-processing workflow (Fig. 2), 436 contain new reaction center **RC** and 30 reactions are novel at first neighborhood level **RC + 1**. Some generated reactions have two or more distinct reaction centers, i.e. represent multistep transformations. Note that “novelty” defined as the absence of reaction center from the training set data is per se meaningful, as an illustration of the “creativity” of this Artificial Intelligence, i.e. its ability to generate original reaction centers which can be submitted for empirical feasibility assessment to human experts. Unfortunately, “novelty” as the absence of reaction center 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 failed 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<sup>44</sup>.

**Reactions curation and generalisation.** A close look at the generated reactions reveals several serious drawbacks: (i) unbalanced reaction equations, (ii) presence of likely unstable groups (e.g.,  $R_3S(=O)H$  and  $R-PH(=O)-OR'$ ), and, (iii) transformations which require harsh reaction conditions (e.g., breaking of a C–C bond), or kinetically unfavorable reactions (e.g., cleavage of a leaving group with carbon at attachment point). Some reactions can be corrected or discarded using some heuristic rules (“Chemical Filters”).

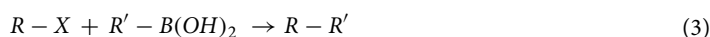
Output of unbalanced reactions is a direct consequence of the training set composition: almost all USPTO reactions are also unbalanced, e.g., leaving groups are almost never reflected in the reaction equation present in the database. The application of the CGR technology may implicitly solve that problem. Indeed, within the CGR formalism, heavy atoms in reactants and products are implicitly conserved—as the same graph is simply interpreted differently in terms of dynamical bond status in order to convert it to reagents or to products, respectively. Even if the initial reaction was not stoichiometrically balanced (see example in SI), its CGR representation will be—in so far the conversion of an unbalanced transformation to CGR succeeds to produce the correct CGR of the balanced process. However, as the exact state of the leaving group cannot be deduced from the training set, *in silico* generated CGRs may occasionally decode into reactions by simply substituting a broken bond by a hydrogen atom, leading to a disbalance in terms of implicit hydrogens. This is seen in the example from Fig. 3A in which the products contain 2 hydrogens more than reactants. Furthermore, the postulated product  $BH(OH)_2$  is highly reactive, thus unrealistic. Formally, this is a rather “creative” *in silico* interpretation of the Suzuki reaction pattern, in which the organic halide  $R-X$  is replaced by an amide group: the acyl fragment is assimilated to “R” while the benzylamine is the leaving group X. Formally, a balanced Suzuki process could be formulated as either



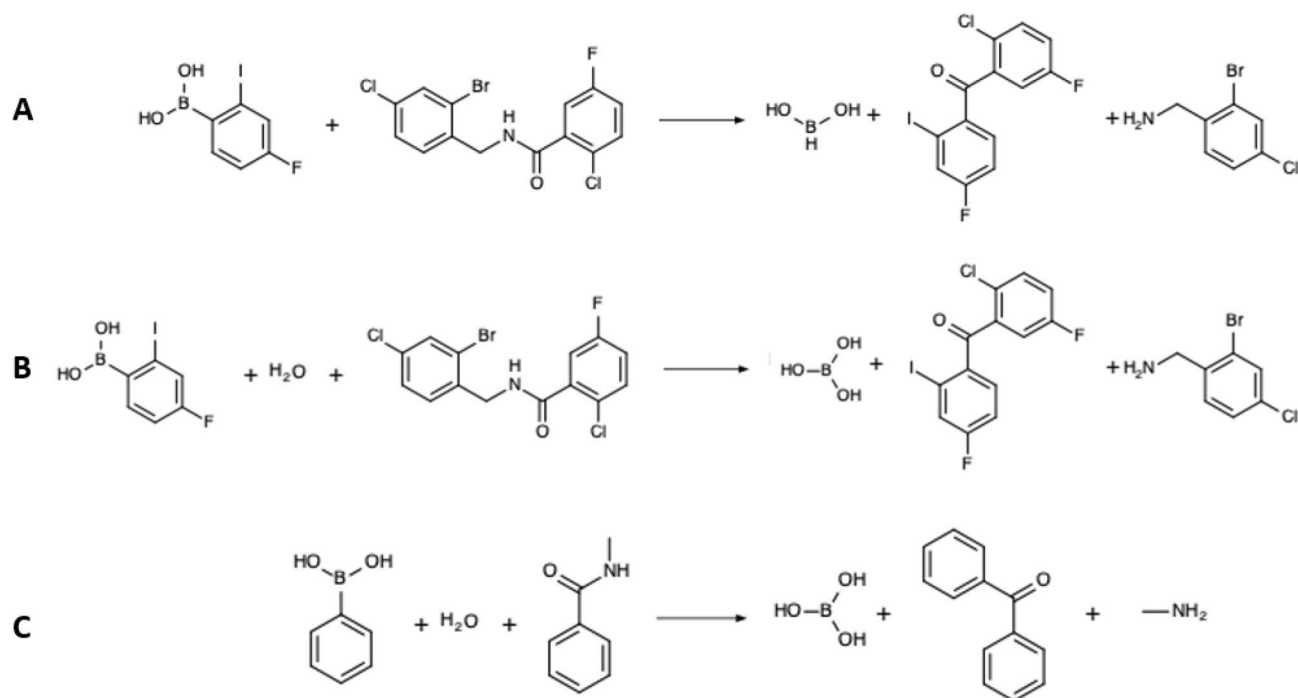
or, more realistically, with inclusion of the required alkoxy base, typically  $BuO^-$ :



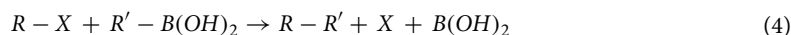
Unfortunately, a sketchily written Suzuki transformation, carelessly ignoring the inorganic leaving groups:



converts to a CGR corresponding to



**Figure 3.** Example of generated chemical reaction with a new reaction center as is (A), balanced by the addition of a water molecule as a reactant (B), and its simplified form (C). Notice that the aminobenzylic leaving group suggested by the autoencoder for generated reaction looks unrealistic.

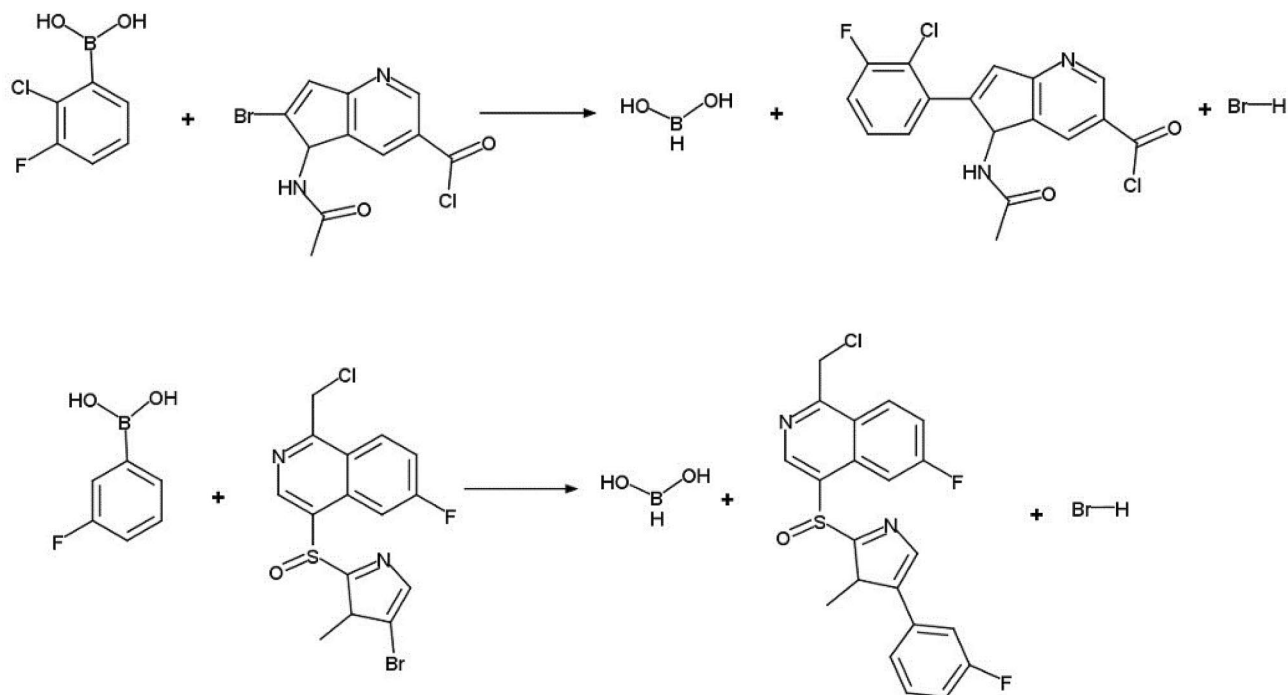


in which the unsatisfied valences of X and B are interpreted by cheminformatics tools as implicit hydrogens. This explains why the AI tool is inclined to generate formal reactions of type (4), which are nothing but a biased interpretation of Suzuki processes, corrupted by intrinsic representation errors in USPTO database entries. The addition of a water molecule as a “formal” basic species leads to a fully balanced reaction (Fig. 3B). Although water is not a perfect base for the Suzuki reaction, it helps to correctly represent the boron-containing leaving group in reaction equations. In silico generated reactions that cannot be “corrected” in this way have been discarded.

We also decided to discard the unfeasible under normal conditions transformations consisting in the cleavage of a C–C bond and assuming a carbon-centric leaving group. Application of these heuristics reduced a considered set of novel reactions to 44 including 31 reactions with new **RC** and 13 reactions with new **RC + 1**.

The question arises whether we need to consider explicit chemical structures of generated reactants and products. In our opinion, this is not firmly required if one focuses on the detection of new reaction transformations identified by **RC** or **RC + 1** structural motifs. In this case, a “simplified” reaction in which substrates contain only atoms of reaction center and their closest environment (including second neighbors) could be sufficient, see Fig. 3C. Notice that such simplified reactions correspond to general reactivity patterns. Particular reactants can be selected by chemists as a function of availability, intended conditions, reactivity concerns, etc. For example, the reaction in Fig. 3C looks unfeasible, but it becomes more realistic if the amine leaving group were strengthened by binding to strong electron acceptors (for example, trifluoromethanesulfonyl) or by quaternization.

Notice that the majority of generated reactions have known **RC** and **RC + 1** motifs. All belong to the Suzuki coupling type, as exemplified below.

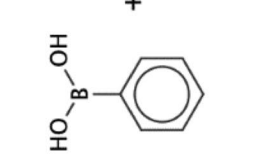
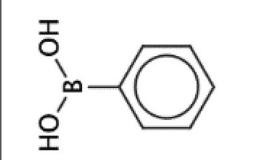
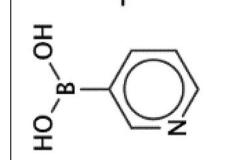
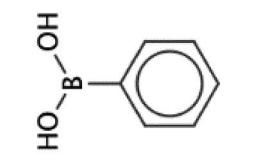


**Reactions with new reaction centers.** 31 reactions featuring a total of 13 distinct reaction centers not seen in USPTO were generated (see Table 1 and Table S3 in Supporting Information). Substructural searching with *RC* as a query in the much larger CAS REACT database (SciFinder) resulted in retrieval of several reactions similar to those “discovered” by Artificial Intelligence. In particular, this concerns reactions with C–Br<sup>56</sup> bond formation and C–Si<sup>57</sup> coupling, and C–C coupling with N-containing<sup>58</sup> and F<sup>59</sup> leaving groups, as well as C–O coupling with organosilicon leaving groups. In total, 5 out of 13 new reaction centers discovered computationally, were found in SciFinder reactions (Table S5 in Supporting Information). Since none of them were used for the autoencoder training, these generated reactions were pure “imagination” of AI. Thus, several “novel” reactions (4 in Table 1, 5, and 7 in Table S5) correspond to a quite interesting C–N bond cleavage with amine as leaving group. A similar reaction has recently been discovered experimentally by Weires et al.<sup>60</sup> who shown that the formation of amides facilitated nickel-catalyzed cleavage of C–N bonds accompanied by C–C coupling (reaction 12 in Table S3). Experimental analogues of C–Si coupling reactions generated by the model (reactions 8 and 9 in Table 1, reactions 19–26 in Table S3) were found in SciFinder (reaction 9 in Table 1 and 19 in Table S5). In the experiment, bromotriarylsilane was used as template<sup>57</sup> (reaction 19 in Table S5) whereas our tool proposed less stable di-substituted silane bromide possibly with heteroatoms surrounding silicon (reactions 8 and 9 in Table 1). The organosilicon leaving group proposed for the C–O coupling (reaction 11 in Table 1) is very similar to that reported by Kori et al.<sup>61</sup> Fluoro–Suzuki reaction proposed by the autoencoder (reaction 5 in Table 1) was observed experimentally in the study by Chi et al.<sup>62</sup> Reaction 7 in Table 1 is not a coupling but boron substitution by bromine; it has been experimentally discovered using N-bromosuccinimide as a donor of bromine in the study by Thiebes et al.<sup>56</sup> (reaction 17 in Table S5). However, from the structural point of view, its reaction center looks similar to “classical” Suzuki type reactions (boron substitution by carbon or heteroatom).

Some of the reactions still look unfeasible, e.g., the O–I compound seems quite unstable (reaction 2 in Table 1). Nonetheless, such compounds are listed as commercially available (e.g. CAS Nos 3240-34-4, 1338247-47-4). Sulfur-containing compounds are generally unsuitable for Suzuki catalysts. Their generation can be explained by an excessive model’s “creativity”, which can be hardly controlled in the employed neural network architecture.

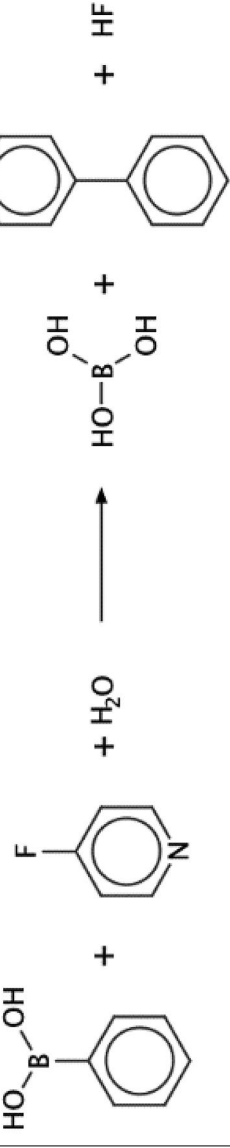
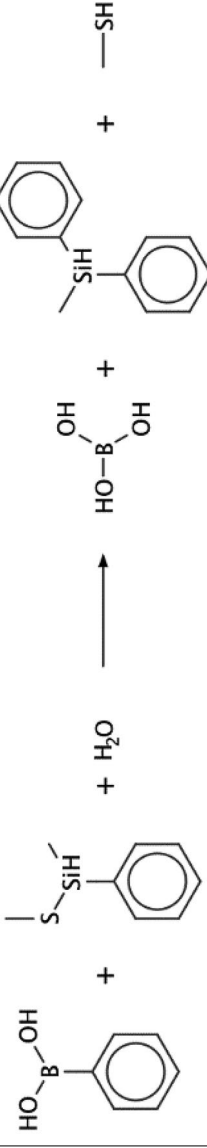

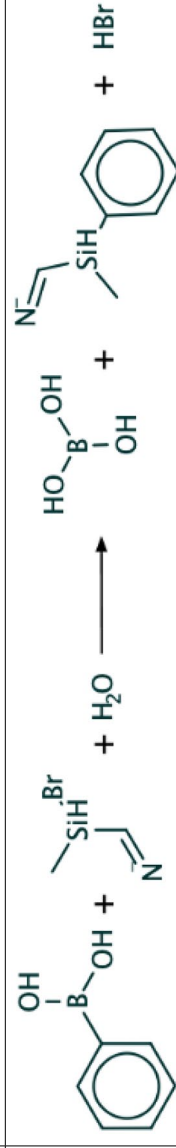
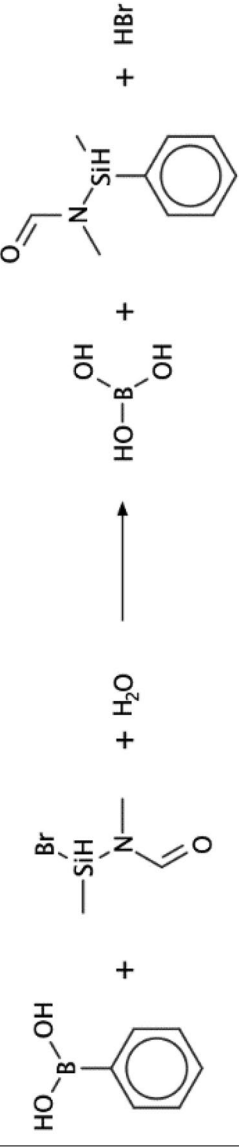
**Reactions with a new environment of known reaction centers (RC + 1).** Following the novelty detection procedure, 13 reactions that correspond to 3 known reaction centers but an original first environment (*RC + I*) were detected, see Table 2 and Table S4 in SI. Two similar reactions have been found in SciFinder. Although the simplified reaction 1 in Table 2 looks unfeasible, a more suitable leaving group might render it possible. For instance, in hydrogenation conditions, a catalyst can facilitate reductive cleavage of C–O bond (in esters, carbamates, benzyl ethers, etc.) followed by a coupling (as in reaction 10 in Table S6).

The use of alkyl and acyl bromides in C–C coupling in reaction 3 (Table 2), was observed experimentally (see reaction 13 in Table S6 in SI). Reaction 2 in Table 2 looks quite feasible because synthesis of acyl iodides was reported in the literature (e.g., CAS 191340-22-4 and CAS 1332596-80-1) whereas carboniodidates can be provided by some vendors (e.g., Enamine BBV-109267542 or BBV-109267541). Notice that similar reactions with chloroformates have been also found in Reaxys<sup>63</sup>.

Reaction center SMILES	Simplified reaction	References
1 O.BC.N>>OB.CN		a
2 O.BC.OI>>OB.CO.I		a
3 O.BC.CS>>OB.CC.S		a
4 O.BC.CN>>OB.CC		58

Continued



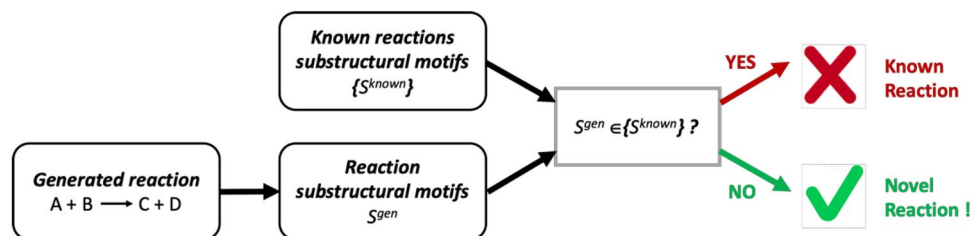
Reaction center SMILES	Simplified reaction	References
5 O.BC.CF>>OB.CC.F		59
6 O.BC.[Si]S>>OB.C[Si].S		a
7 O.BC.BrN>>BO.CBr.N		56
8 O.BC.[Si]Br>>OB.C[Si].Br		a
9 O.BC.[Si]Br>>OB.C[Si].Br		57
Continued		

Reaction center SMILES	Simplified reaction	References
10 O.BC.SBr > > OB.CS.Br		<sup>a</sup>
11 O.BC.O[Si] > > OB.CO.[Si]		<sup>61</sup>

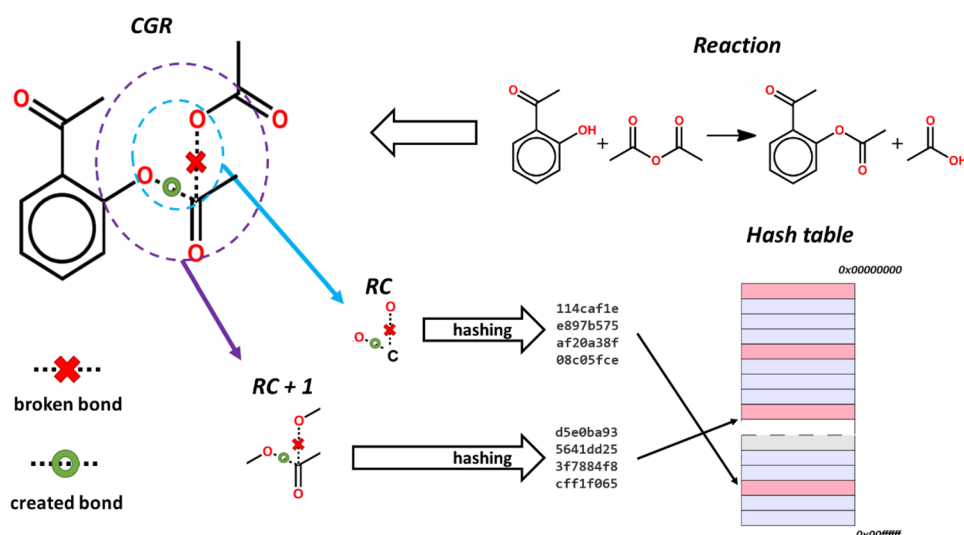
**Table 1.** Examples of simplified Suzuki-type reactions with a new reaction center together with corresponding conventional reaction SMILES notation. The right column refers to similar types of reactions found in SciFinder. A complete list of simplified reactions is given in SI. <sup>a</sup>Reaction centers for which no information in the literature was found.

Reaction center SMILES	Simplified reaction	References
1 <chem>BC.O.CO&gt;&gt;O.CC.BO</chem>		64
2 <chem>BC.O.Cl&gt;&gt;I.CC.BO</chem>		63
3 <chem>BC.O.CBr&gt;&gt;Br.CC.BO</chem>		65

**Table 2.** Examples of simplified Suzuki-type reactions with the RC present in the training set but in a new chemical environment. The right column refers to similar types of reactions found in SciFinder. A complete list of simplified reactions is given in SI.



**Figure 4.** 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 5.** 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. The ensemble of unique hash codes for all reactions in the database is stored in the hash table.

**Reactions feasibility assessment.** Strictly speaking, reaction feasibility is defined by both kinetic and thermodynamic factors. However, according to the Bell-Evans-Polanyi principle<sup>66,67</sup>, in a series of similar reactions, the trend of activation energies follows the trend of reaction enthalpies. Thus, favorable thermodynamics, namely reaction enthalpy ( $\Delta H$ ), can be considered as weak proof of reaction feasibility. A series of gas-phase DFT calculations were performed to assess  $\Delta H$  for all simplified reactions with new  $RC$  and  $RC + 1$ . According to our estimations, almost all reactions are exothermic except for four reactions with Si-containing substrates in which  $\Delta H$  is positive but close to zero (see Tables S3 and S4 in Supporting Information). This shows that all new computer-generated reactions are feasible, at least, as far as DFT-based thermodynamics estimates can tell. Since DFT is a rather time-consuming method and can hardly be applied for thousands of generated reactions, we also performed a rough estimation of  $\Delta H$  using the tabulated bond energies in reactants and products<sup>68–70</sup>. Although calculated in such a way reaction enthalpies poorly correlate with the DFT values, they generally provide similar conclusions concerning reaction feasibility (see Tables S3 and S4 in Supporting Information).

## Conclusion

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. To feed the autoencoder, special reaction SMILES strings (SMILES/CGR) were conceived and implemented. In order to discard the seemingly unfeasible reactions, a special 4-steps post-processing procedure has been implemented. It includes: (i) stoichiometric balancing of reaction equations, (ii) reduction of substrates structure to their simplified form, (iii) discarding chemically infeasible transformations using suggested heuristics (“Chemical filters”), and (iv) assessment of synthetic feasibility using quantum mechanics calculations. The effectiveness of the suggested approach was demonstrated on the example of Suzuki-like coupling reactions. Among generated reactions we discovered transformations with 13 new reaction centers which did not occur in the training set. Five out of 13

transformations were then found in the reaction databases (not used in the model training), thus showing the reliability of our approach to generate new synthetically feasible reactions.

This study reveals that creativity of Artificial Intelligence is rather limited. Deep learning neural networks, at least, in their current state, are not able to invent completely new type of chemical transformations but rather propose unseen and sometimes not trivial variations of existing ones. Thus, in this study novel (in the context of the training set) C–N, C–O, C–S and C–Si bond formation reactions, as well as nitrogen- and sulfur-containing leaving groups have been suggested by the model. We believe that this opens a way to propose putatively new synthetic pathways in a way that is not affected by the bias of human expertise—with all the benefits and the pitfalls this may bring. It should also be noted that compared to theory-driven quantum chemical models, data-driven DNN is much less time consuming and, practically, is not limited by the reactants size. The more data are used in the neural network training, the more realistic the predicted reactions are. Since the sizes of reaction databases are rapidly growing up, deep learning approach has an obvious perspective as a tool for discovery of novel reactions.

## Methods

**Datasets and data curation.** The dataset used in this project comes from United States Patents and Trademark Office database (1976 to 2016) extracted by Lowe<sup>53</sup>. It contains about 3.5 million reactions. The initial dataset was preprocessed with *in-house* scripts based on the CGRtools library<sup>51</sup>. The curation includes the standardization (aromatization and functional group standardization), removal of empty reactions (those where the products and reactants are 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<sup>71</sup>. The 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.

According to our estimations<sup>72</sup>, the ChemAxon Automapper tool leads to the erroneous AAM for some 25% of USPTO reactions. Most of those concern cycloadditions with complex reaction centers. As far as Suzuki coupling is concerned, this error is around 3%.

Notice that practically all USPTO reactions are stoichiometrically unbalanced. This doesn't prevent to build Condensed Graph of Reaction, but, in some cases, may lead to erroneous atom-to-atom mapping.

**Reaction data treatment.** CGRtools library (version 3)<sup>51</sup> 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<sup>73</sup>. They differ from regular Daylight SMILES in terms of aromatic atoms and ring closure specification and introduce special “dynamic” bonds and atoms characterizing chemical transformations. Dynamic bonds in CGR characterizing chemical transformations have special labels representing changes in bond orders. Dynamic atom corresponds to change of formal charge or radical state of this atom in reaction. Detailed information about SMILES/CGR syntax is given in Supporting Information. SMILES/CGR generation and parsing, including preparation of canonic SMILES/CGR, are implemented into CGRtools Python library<sup>51</sup>.

**Reaction generation algorithm.** The network architecture previously applied for molecular SMILES generation<sup>30</sup> has been used in this study. It is based on the autoencoder architecture introduced by Xu et al.<sup>74</sup>. 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)<sup>52</sup> 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.

The autoencoder was trained in batch mode, where batches of “one-hot”-encoded sequences were generated on-the-fly from training set SMILES/ CGR strings. The Adam optimizer was used for training, initial learning rate was set to 0.005, and batch size was set to 256 samples per batch. The learning rate was reduced during training if there were no improvement in the validation loss for two epochs. The training was terminated after 34 epochs when no improvements in test set reconstruction accuracy was observed. 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<sup>54,74–82</sup>, comparison of chemical libraries<sup>83</sup>, building classification<sup>43,74–77,80,84</sup>, and regression<sup>85,86</sup> models via activity landscapes, as well as for solving the “inverse” QSAR problem<sup>87</sup>. 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 (Fig. 4). 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 Fig. 5). Since any operations with molecular graphs are time-consuming, each substructural motif was encoded by a unique hash code<sup>51</sup>—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 (Fig. 4). The suggested procedure assures fast and precise novelty detection.

**Reaction enthalpy calculations.** The difference in energies between reactants and products is calculated in several steps<sup>88</sup>. First, a conformer with the lowest energy is generated for each compound in the reaction using the ChemAxon cxcalc module. Then, the geometry of each compound was optimized using the Priroda16 program with PBE exchange and correlation functional<sup>89</sup>, and the built-in triple-zeta split valence basis set 3z, which is equivalent to Schäfer's TZVP basis set<sup>90</sup>. Relativistic and solvent effects were neglected. The Priroda16 program was chosen as it is one of the fastest DFT software due to the efficient evaluation of density functional exchange–correlation terms based on the electron density expansion<sup>91</sup>. Final energy values were extracted for optimized structures and used for calculation of reaction enthalpy. The additive scheme for estimating reaction enthalpies was implemented using the tabulated chemical bonds increments<sup>68–70</sup>.

### Data availability

The dataset used in this project comes from the publicly available United States Patents and Trademark Office database (Lowe, <https://doi.org/10.17863/CAM.16293>). Curated USPTO dataset is available on GitHub: <https://github.com/Laboratoire-de-Chemoinformatique>. All data preprocessing procedures are described in the Methods section and are based on freely available CGRtools library.

### Code availability

CGRtools library is used for data preprocessing and creation and treatment of chemical reactions as SMILES/CGR, and is freely available (<https://github.com/cimm-kzn/CGRtools>). Autoencoder model code and the ISIDA/GTM tool are available upon request.

Received: 8 August 2020; Accepted: 6 January 2021

Published online: 04 February 2021

### References

- Herges, R. Reaction planning: Computer-aided reaction design. *Tetrahedron Comput. Methodol.* **1**, 15–25 (1988).
- Balaban, A. T. Chemical graphs. 3. Reactions with cyclic 6-membered transition states. *Rev. Roum. Chim.* **12**, 875–902 (1967).
- Hendrickson, J. B. The variety of thermal pericyclic reactions. *Angew. Chem. Int. Ed. English* **13**, 47–76 (1974).
- 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).
- 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).
- 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).
- Zefirov, N. S. & Tratch, S. S. Formal-logical approach to multicentered processes with cyclic electron transfer. *Match* **3**, 263–264 (1977).
- Zefirov, N. S. S., Tratch, S. 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).
- Bauer, J., Herges, R., Fontain, E. & Ugi, I. IGOR and computer assisted innovation in chemistry. *Chimia (Aarau)*. **39**, 43–53 (1985).
- Bauer, J. IGOR2: A PC-program for generating new reactions and molecular structures. *Tetrahedron Comput. Methodol.* **2**, 269–280 (1989).
- Dugundji, J. & Ugi, I. An algebraic model of constitutional chemistry as a basis for chemical computer programs. In *Computers in Chemistry* 19–64 (Springer-Verlag, Berlin, 1973).
- Herges, R. Reaction planning: Prediction of new organic reactions. *J. Chem. Inf. Comput. Sci.* **30**, 377–383 (1990).
- Herges, R. & Hoock, C. Reaction planning: Computer-aided discovery of a novel elimination reaction. *Science* **255**, 711–713 (2020).
- 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).
- Zefirov, N., Tratch, S. & Molchanova, M. The argent program system: A second-generation tool aimed at combinatorial search for new types of organic reactions. *Math. Comput. Chem.* **46**, 253–273 (2002).
- 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).
- 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).
- Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604 (2018).
- Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. *Adv. Neural. Inf. Process. Syst.* **4**, 3104–3112 (2014).
- Nam, J. & Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. Preprint at *arXiv* <http://arxiv.org/abs/1612.09529> (2016).
- 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).
- Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2020).
- Karpov, P., Godin, G. & Tetko, I. V. A transformer model for retrosynthesis. *Lect. Notes Comput. Sci.* **11731**, 817–830 (2019).

24. Schwaller, P. *et al.* Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy. (2019) doi:<https://doi.org/10.26434/chemrxiv.9992489.v1>.
25. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
26. Fooshee, D. *et al.* Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* **3**, 442–452 (2018).
27. Kayala, M. A. & Baldi, P. ReactionPredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. *J. Chem. Inf. Model.* **52**, 2526–2540 (2012).
28. 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).
29. Xu, Y. *et al.* Deep learning for molecular generation. *Fut. Med. Chem.* **11**, 567–597 (2019).
30. Sattarov, B. *et al.* De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping. *J. Chem. Inf. Model.* **59**, 1182–1196 (2019).
31. 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).
32. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. Application of generative autoencoder in de novo molecular design. *Mol. Inform.* **37**, 1700123 (2018).
33. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science (80-)* **361**, 360–365 (2018).
34. Jørgensen, P. B., Schmidt, M. N. & Winther, O. Deep generative models for molecular science. *Mol. Inform.* **37**, 1700133 (2018).
35. Gupta, A. *et al.* Generative recurrent networks for de novo drug design. *Mol. Inform.* **37**, 1700111 (2018).
36. Segler, M. H. S. & Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chem. A Eur. J.* **23**, 6118–6128 (2017).
37. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
38. 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).
39. Varnek, A., Fourches, D., Hoonakker, F. & Solovev, V. P. Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided. Mol. Des.* **19**, 693–703 (2005).
40. Hoonakker, F., Lachiche, N., Varnek, A. & Wagner, A. A representation to apply usual data mining techniques to chemical reactions. *Lect. Notes Comput. Sci.* **6097**, 318–326 (2010).
41. Madzhidov, T. I. *et al.* Structure-reactivity relationships in terms of the condensed graphs of reactions. *Russ. J. Org. Chem.* **50**, 459–463 (2014).
42. 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).
43. Gimadiev, T. *et al.* Bimolecular nucleophilic substitution reactions: Predictive models for rate constants and molecular reaction pairs analysis. *Mol. Inform.* **38**, 1800104 (2019).
44. Glavatskikh, M. *et al.* Predictive models for kinetic parameters of cycloaddition reactions. *Mol. Inform.* **38**, 1800077 (2019).
45. Gimadiev, T. R. *et al.* Assessment of tautomer distribution using the condensed reaction graph approach. *J. Comput. Aided. Mol. Des.* **32**, 401–414 (2018).
46. Gimadiev, T. R. *et al.* Prediction of tautomer equilibrium constants using condensed graphs of reaction. in *Second Kazan Summer School on Chemoinformatics* 34 (2015).
47. 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).
48. 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).
49. Madzhidov, T. I. *et al.* Artificial neural networks model for assessment of optimal conditions of hydrogenation reactions. in *12th European Symposium on Quantitative Structure-Activity Relationships*. 186 (2018).
50. Marcou, G. *et al.* Expert system for predicting reaction conditions: The Michael reaction case. *J. Chem. Inf. Model.* **55**, 239–250 (2015).
51. 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).
52. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
53. Lowe, D. M. M. Extraction of chemical structures and reactions from the literature. *Doctoral Thesis* (University of Cambridge, 2012). doi:<https://doi.org/https://doi.org/10.17863/CAM.16293>.
54. Gaspar, H. A. *et al.* Generative topographic mapping approach to chemical space analysis. *ACS Symp. Ser.* **1222**, 211–241 (2016).
55. Chen, W. L., Chen, D. Z. & Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 560–593 (2013).
56. Thiebes, C., Thiebes, C., Prakash, G. K. S., Petasis, N. A. & Olah, G. A. Mild preparation of haloarenes by ipso-substitution of arylboronic acids with N-halosuccinimides. *Synlett* **2**, 141–142 (1998).
57. Park, J. *et al.* Indole compound, compound for organic electric element containing derivative thereof, organic electric element using same, and corresponding electronic device. PCT/KR2013/003289. (2013).
58. Zong, Y., Hu, J., Sun, P. & Jiang, X. Synthesis of biaryl derivatives via a magnetic Pd-NPs-catalyzed one-pot diazotization-cross-coupling reaction. *Synlett* **23**, 2393–2396 (2012).
59. Luo, Z.-J., Zhao, H.-Y. & Zhang, X. Highly selective Pd-catalyzed direct C–F bond arylation of polyfluoroarenes. *Org. Lett.* **20**, 2543–2546 (2018).
60. Weires, N. A., Baker, E. L. & Garg, N. K. Nickel-catalyzed Suzuki–Miyaura coupling of amides. *Nat. Chem.* **8**, 75–79 (2016).
61. Kori, M. *et al.* Fused thiaziazine derivatives as AMPA receptor potentiators and their preparation and use for the treatment of diseases. *PCT Int. Appl.* **16**, 2012020848 (2012).
62. Chi, Y. & Lin, J. Iridium complex, OLED using the same, and nitrogen-containing tridentate ligand having carbene unit. *Faming Zhuanli Shenqing* 106928281 <https://patents.google.com/patent/US10153442B2> (2017).
63. Duan, Y.-Z. & Deng, M.-Z. Palladium-catalyzed cross-coupling reaction of arylboronic acids with chloroformate or carbamoyl chloride. *Synlett* **02**, 355–357 (2005).
64. Dindarloo Inaloo, I., Majnooni, S., Eslahi, H. & Esmaeilpour, M. Nickel(II) Nanoparticles Immobilized on EDTA-Modified Fe<sub>3</sub>O<sub>4</sub>. SiO<sub>2</sub> Nanospheres as Efficient and Recyclable Catalysts for Ligand-Free Suzuki–Miyaura Coupling of Aryl Carbamates and Sulfamates. *ACS Omega* **5**, 7406–7417 (2020).
65. Chakraborty, J., Nath, I. & Verpoort, F. Pd-nanoparticle decorated azobenzene based colloidal porous organic polymer for visible and natural sunlight induced Mott-Schottky junction mediated instantaneous Suzuki coupling. *Chem. Eng. J.* **358**, 580–588 (2019).
66. Bell, R. P. & Hinshelwood, C. N. The theory of reactions involving proton transfers. *Proc. R. Soc. London. Ser. A Math. Phys. Sci.* **154**, 414–429 (1936).
67. Evans, M. G. & Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans. Faraday Soc.* **32**, 1333–1360 (1936).

68. Cottrell, T. L. *The strengths of chemical bonds*. (Butterworths Scientific Publications, 1958).
69. Darwent, B. deB. *Bond dissociation energies in simple molecules*. (1970).
70. Benson, S. W. III. Bond energies. *J. Chem. Educ.* **42**, 502 (1965).
71. ChemAxon. Chemical Structure Representation Toolkit. (2019).
72. Lin, A. I. *et al.* Atom-to-Atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. <https://doi.org/10.26434/chemrxiv.13012679.v1> (2020).
73. James, C. A. OpenSMILES specification. [www.opensmiles.org](http://www.opensmiles.org) (2016).
74. 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:<https://doi.org/10.1145/3107411.3107424>.
75. 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).
76. 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).
77. 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).
78. 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).
79. 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).
80. 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).
81. Glavatskikh, M. *et al.* Visualization and analysis of complex reaction data: The case of tautomeric equilibria. *Mol. Inform.* **37**, 1800056 (2018).
82. Horvath, D., Marcou, G. & Varnek, A. Generative topographic mapping approach to chemical space analysis. 167–199 (2017). doi:[https://doi.org/10.1007/978-3-319-56850-8\\_6](https://doi.org/10.1007/978-3-319-56850-8_6).
83. 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).
84. 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).
85. 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).
86. Baskin, I. I., Solovev, V. P., Bagaturyants, A. A. & Varnek, A. Predictive cartography of metal binders using generative topographic mapping. *J. Comput. Aided. Mol. Des.* **31**, 701–714 (2017).
87. 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).
88. Gimadiev, T. R., Klimchuk, O., Nugmanov, R. I., Madzhidov, T. I. & Varnek, A. Sydnone-alkyne cycloaddition: Which factors are responsible for reaction rate? *J. Mol. Struct.* **1198**, 126897 (2019).
89. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
90. Schäfer, A., Huber, C. & Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.* **100**, 5829–5835 (1994).
91. Laikov, D. N. Fast evaluation of density functional exchange-correlation terms using the expansion of the electron density in auxiliary basis sets. *Chem. Phys. Lett.* **281**, 151–156 (1997).

## Acknowledgements

RN, IB, TM are grateful to Russian Science Foundation (Project No 19-73-10137) for the support of CGRtools library development.

## Author contributions

All authors contributed to the code development, obtaining the results and preparation the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-81889-y>.

**Correspondence** and requests for materials should be addressed to A.V.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021