Heterogeneous Catalysts in Grammar School

Johannes T. Margraf,[*,†] Zachary W. Ulissi,[‡] Yousung Jung,[¶] and Karsten Reuter[*,†]

[†]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany*

[‡]*Chemical Engineering Department, Carnegie Mellon University, Pittsburgh, PA 15217, United States of America*

[¶]*Department of Chemical and Biomolecular Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea*

Received December 3, 2021; E-mail: margraf@fhi.mpg.de; reuter@fhi.mpg.de

Abstract: The discovery of new catalytically active materials is one of the holy grails of computational chemistry as it has the potential to accelerate the adoption of renewable energy sources and reduce the energy consumption of chemical industry. Indeed, heterogeneous catalysts are essential for the production of synthetic fuels and many commodity chemicals. Consequently, novel catalysts with higher activity and selectivity, increased sustainability and longevity, or improved prospects for rejuvenation and cyclability are needed for a diverse range of processes. Unfortunately, computational catalyst discovery is a daunting task, among other reasons because it is often unclear whether a proposed material is stable or synthesizable. This perspective proposes a new approach to this challenge, namely the use of generative grammars. We outline how grammars can guide the search for stable catalysts in a large chemical space and sketch out several research directions that would make this technology applicable to real materials.

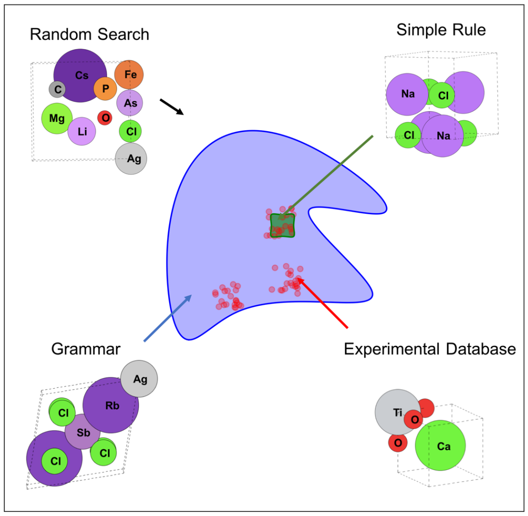

Figure 1. Cartoon depiction of different screening spaces. Random search (white) covers a wide range of candidates but includes many unphysical structures. A simple rule (green) defines a very restrictive space. Experimental or computational databases can be highly diverse but are also biased and incomplete, potentially missing entire classes of interesting materials. A formal grammar could in principle cover a large screening space without including unphysical candidates. Note that the proportions are arbitrary. In particular, the space of random structures is much larger than depicted and mostly consists of nonsensical structures.

Heterogeneous catalysis is an essential technology for enabling sustainable economic development.[1,2] On one hand, chemical processes like ammonia-synthesis require massive amounts of energy and are thus substantial greenhouse gas emitters. On the other hand, the long term storage of renewable energy in synthetic fuels is itself a catalytic process. In both cases, new and improved catalysts would therefore yield large benefits towards reducing global net carbon emissions. While new catalysts have historically often been found by serendipity or empirical insight, theoretical understanding has played an increasingly significant role over the last decades. Indeed, not least the fundamental theoretical understanding of catalyst functionality based on scaling relations (limited as it may practically be) has led to the emergence of an entire field of computational screening based catalyst discovery.[3–5]

Such a computational catalyst screening requires first to define a library of candidates (i.e. a chemical space, see Fig. 1). This space is typically constructed according to some simple rules (e.g. the set of ordered metals or solid solution alloys in a fixed lattice) or taken from some predefined experimental or computational database (e.g. the Materials Project[6]). Once this space is defined, the screening itself consists of computationally estimating the catalytic activity of all candidates (or representative samples) contained therein.

Using a predefined database to span the chemical space has the advantage that all candidates fulfill certain requirements (implicitly) set upon construction of the database. For instance, they correspond to known, stable structures if the database is constructed from experimental data. On the flipside, this means that there is a strong selection bias and the screening will not be able to discover new, unexpected materials. A rule-based definition of the library is in principle less biased, as it allows enumerating all possible structures within its constraints, not just known systems. Most catalyst screening studies typically use very simple rules, however, so that these screenings can be equally restrictive. Alternatively, one could imagine a third strategy, namely the completely random sampling of atomic arrangements. While this approach would definitely be unbiased and unconstrained, it would also lead to mostly unphysical structures, making the screening extremely inefficient.

The above paragraph reveals some crucial desiderata for a catalyst screening library, namely that it should be unbiased, extensive and exclusively contain valid samples. As we shall see, what this means in practice depends somewhat on the context. Nevertheless, these (partly competing) goals should always be taken into account when designing a screening study. It is the purpose of this perspective to argue that a good way to balance these requirements could be to define the chemical space of interest via *formal gram-*

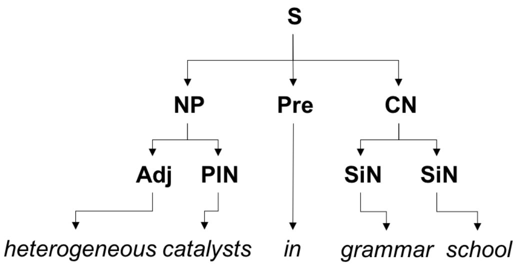

Figure 2. A grammatical derivation of the title of this article.

mars. In the following we briefly give a general introduction into this concept before discussing how it can be useful for catalyst discovery.

Formal grammars: Formal grammars were originally developed in the field of theoretical linguistics, where they describe how syntactically valid sentences can be formed from a language's words.[7,8] An example of this is shown in Fig. 2. Here, the title of this article is formed by creating a sentence of the form 'noun phrase'+'preposition'+'compound noun' (abbreviated as **NP**, **Pre** and **CN**, respectively). Subsequently, **NP** and **CN** are further specified: The former consists of an adjective (**Adj**) and a plural noun (**PlN**), while the latter consists of two singular nouns (**SiN**). Finally, these placeholders are replaced by actual english words so that **Pre** becomes *in*, **Adj** becomes *heterogeneous*, etc.

The power of formal grammars does not just lie in the analysis of given sentences, however. Instead a grammar can be used to generate *all* syntactically valid sentences in a language. To see how this works, we must first understand what the components of such a grammar are. To this end, a simple toy grammar is introduced in Fig. 3, based on the sets N, Σ and P. The first of these collects all so-called *non-terminal symbols*, which are placeholders for certain types of words or phrases (i.e. **Adj** or **SiN**, in the example above). Here, we also include the starting symbol **S**, which marks the start of every new sentence derivation. The second set contains all *terminal symbols*, which are the actual words of the language (i.e. *heterogeneous* or *school*). Finally, the set P contains the production rules of the grammar. Each production rule is a prescription of how the non-terminal symbols of a language can be replaced or modified. For example, one rule specifies that the non-terminal symbol **Adj** can be replaced by one of the terminal symbols *heterogeneous*, *young* and *promising*.

As exemplified in Fig. 2, sentences can be generated from this grammar by applying the production rules (depicted as branched arrows) in a sequential manner. Starting from the non-terminal symbol **S**, only a single production rule is available (**S** → **NP Pre CN**). The final sentence is reached when no non-terminal symbols are left and therefore no more production rules can be applied. Importantly, this sentence is only one of many that can be generated by the grammar in Fig. 3. In Table 1, some other examples are shown, along with random sentences constructed by combining arbitrary words from the dictionary Σ. These sentences also serve to illustrate the distinction between syntax and semantics: The grammar generates sentences that are syntactically valid. This does not mean that these sentences are necessarily meaningful (i.e. semantically valid). We are however much more likely to generate a meaningful sentence with the grammar than with the random generator, which produces complete gibberish in most cases.

$$N = \{\boldsymbol{S}, \boldsymbol{NP}, \boldsymbol{Pre}, \boldsymbol{CN}, \boldsymbol{Adj}, \boldsymbol{PlN}, \boldsymbol{SiN}\}$$

$$\Sigma = \left\{ \begin{array}{c} heterogeneous, young, promising, \\ catalysts, violinists, signals, \\ in, grammar, band, news, school, camp, room \end{array} \right\}$$

$$P = \left\{ \begin{array}{c} \boldsymbol{S} \rightarrow \boldsymbol{NP}\ \boldsymbol{Pre}\ \boldsymbol{CN} \\ \boldsymbol{NP} \rightarrow \boldsymbol{Adj}\ \boldsymbol{PlN} \\ \boldsymbol{CN} \rightarrow \boldsymbol{SiN}\ \boldsymbol{SiN} \\ \boldsymbol{Adj} \rightarrow [heterogeneous; young; promising] \\ \boldsymbol{PlN} \rightarrow [catalysts; violinists; signals] \\ \boldsymbol{SiN} \rightarrow [grammar; band; news; school; camp; room] \\ \boldsymbol{Pre} \rightarrow [in] \end{array} \right\}$$

Figure 3. Components of a formal grammar. The set N contains non-terminal symbols (i.e. placeholders for certain types of words or phrases). The set Σ contains terminal symbols (i.e. the words of the language). The set P contains production rules, which define how non-terminal symbols can be modified and replaced.

Table 1. Example sentences generated with the grammar in Fig. 3 and by randomly combining five words from the corresponding dictionary. The grammar leads to *syntactically* valid sentences but not necessarily *semantically* valid (i.e. meaningful) ones. The random generation meanwhile produces complete gibberish.

grammar

heterogeneous catalysts in grammar school
young violinists in band camp
promising signals in news room
heterogeneous signals in school band
young catalysts in grammar news

random

grammar catalysts school news in
young in promising signals camp
catalyst signals camp news young
camp signals news school band
violinists camp in heterogeneous camp

Catalyst grammars: While this may seem far removed from heterogeneous catalysis, string-based representations and concomitant grammars have actually already found wide application in the not so distant field of organic chemistry, e.g. in the form of SMILES strings or the more recent SELFIES grammar.[9,10] The dictionaries of these chemical languages consists of atoms and bonds that are combined to form strings representing molecules. Importantly, the corresponding syntax imposes physical and chemical constraints into what kinds of molecules can be formed. For instance, the SELFIES grammar is constructed such that all generated strings by definition fulfill the valence rules of organic chemistry.[10]

Despite their undisputed importance in organic chemistry, strings and grammars are much less developed for the inorganic and condensed-phase systems of interest in heterogeneous catalysis, however. Arguably, this results from the much higher complexity and variability of the corresponding extended materials. Before entering a more differentiated discussion onto this matter, let us first further motivate why striving for such grammars could be a worthwhile endeavor.

$$N = \{\boldsymbol{S}, \widehat{D_1}, \widehat{A_1}, D_1, D_2, A_1, A_2\}$$

$$\Sigma = \left\{ \begin{matrix} Li^{+}, Na^{+}, K^{+}, Rb^{+} \\ Be^{2+}, Mg^{2+}, Ca^{2+}, Sr^{2+} \\ O^{2-}, S^{2-}, Se^{2-}, Te^{2-} \\ F^{-}, Cl^{-}, Br^{-}, Rb^{-} \end{matrix} \right\}$$

$$P = \left\{ \begin{matrix} \boldsymbol{S} \rightarrow \widehat{D_1}\widehat{A_1} \\ \boldsymbol{S} \rightarrow D_2 A_2 \\ \widehat{D_1} \rightarrow D_1 \\ \widehat{D_1} \rightarrow D_2 A_1 \\ \widehat{A_1} \rightarrow A_1 \\ \widehat{A_1} \rightarrow D_1 A_2 \\ D_1 \rightarrow [Li^{+}; Na^{+}; K^{+}; Rb^{+}] \\ D_2 \rightarrow [Be^{2+}, Mg^{2+}, Ca^{2+}, Sr^{2+}] \\ A_2 \rightarrow [O^{2-}, S^{2-}, Se^{2-}, Te^{2-}] \\ A_1 \rightarrow [F^{-}, Cl^{-}, Br^{-}, Rb^{-}] \end{matrix} \right\}$$

Figure 4. Definition of a simple grammar for ionic compositions. The production rules ensure that only charge balanced compositions with at most four elements can be generated.

To this end, we consider the simple toy problem of finding stable ionic material compositions out of the 28 main group elements in periods 2-5 and groups 1-17 (i.e. from Lithium to Iodine). For simplicity, we assume that each element only occurs in oxidation states that lead to the closest noble gas configuration (e.g. Li^{+1}, Mg^{+2}, Cl^{-1}, etc.). To screen potential catalysts from these elements we could consider as a simple rule only those binaries of the type AB with balanced charges (e.g. $Na^{+1}Cl^{-1}$, $Mg^{+2}O^{-2}$, $Ga^{+3}As^{-3}$, etc). Unfortunately, this leads to a disappointingly low total of 55 possible materials and shows that vaster spaces need to be spanned to possibly identify new promising materials in the screening. This can be achieved by expanding our search space up to quaternary compositions and considering all combinatorial possibilities, which leads to a much larger library of over 600,000 candidates. However, these mostly correspond to unlikely (electronically unbalanced) compositions such as $Na^{+}O_3^{-2}$ or $Al^{+3}Ga^{+3}In^{+3}F^{-1}$.

To obtain a set of candidates that is less restrictive than the simple binaries and more physically plausible than the random combination of elements, we now define a grammar that allows the systematic composition of strings that correspond to quaternary compositions with balanced oxidation states (see Fig. 4 for a simplified version of the grammar). The production rules of this grammar ensure that non-terminal symbols can only be replaced by the corresponding elements or combinations of other non-terminal symbols which conserve the oxidation state (e.g. a halogen can be replaced by combining an alkali metal and a chalkogen). Furthermore, the grammar by construction only generates compositions with up to four elements. In this way, we end up with a significant screening space of ca. 1,500 systems that exclusively consist of chemically reasonable materials like $Ca^{+2}Sr^{+2}Ge^{-4}$ or $Li_3^{+}P^{-3}$.

A quantitative comparison of these approaches shows that about 30% of the grammatically generated compositions can be found on the Materials Project (MP) database,[6] whereas the same is true for only 1% of the random compositions (see Fig. 5). Moreover, for those structures found in the MP database, the mean energies above the convex hull (indicating thermodynamic (meta-)stability)[11] are 45 and 210 meV/atom for the most stable structure corresponding to each grammatical and random composition, respectively.

The 'known' structures proposed by the grammar are therefore significantly more likely to be (meta-)stable, compared to the ones found through random search. Overall, the grammar thus produces many systems that are known to be stable, but about 70% of the generated systems are unknown. In this sense it nicely balances between the overly restrictive 'simple-rule' approach and the chemically unreasonable random approach.

It might be argued that the benefit of using the grammar in the above example could also be achieved by simply enumerating all possible compositions and filtering charge-balanced ones out after the fact. Indeed, this was the strategy used by Davies *et al.* in their paper 'Computational Screening of All Stoichiometric Inorganic Materials'.[12,13] However, any such brute-force approach will eventually run into a combinatorical wall, with the number of quarternary compounds in that paper already exceeding 10^{12}. This number would further explode if the multitude of possible crystal structures for each composition were taken into account. Meanwhile, using a generative grammar ensures that only more interesting compositions are produced in the first place.

Additional physical understanding may or may not be included flexibly as additional production rules, this way further tailoring the generated chemical search space. In this respect, the use of a grammar in computational screening also blurs the boundary between mere discovery and purposeful design. While (exhaustive) searching in any enumerated space is a discovery process, production rules in grammars offer the prospect to introduce partial understanding of design rules (in the present example the understanding that balanced oxidation states favor stability). This partial understanding is likely not sufficient for a targeted atom-by-atom design. However, formulated as production rules within a grammar is allows to focus the search space to those materials that are consistent with the present understanding and unbiased regarding the rest.

While this illustrates the potential advantage of working with a catalyst grammar, the presented example is obviously only of a toy nature. Clearly, the elemental composition is an overly simplistic representation of a real catalyst. In the rest of this perspective, we therefore want to discuss some of the challenges and requirements for the development of a more general grammar of heterogeneous catalysts.

String representations: Since formal grammars are intimately connected with strings and languages, one way forward would be to develop more useful string-based representation of catalyst materials. The difficulty therein is that (unlike the elemental composition) the three dimensional arrangement of atoms in a solid does not naturally map onto a one-dimensional string. While this is also true for organic molecules, the SMILES language uses powerful, domain-specific abstractions like chemical bonds, implicit hydrogen atoms and atom-typing to achieve this mapping.[9] Defining such abstractions is strongly simplified by the small number of elements that are relevant in organic chemistry. Even with these advantages, non-local features like rings still cause problems with SMILES, e.g. for machine-learning applications.[10] Unfortunately, such features are ubiquitous in

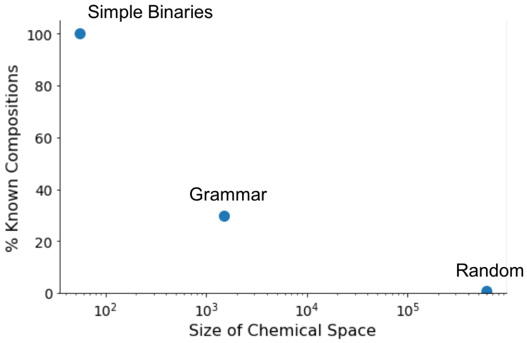

Figure 5. Percentage of compositions found in the Materials Project database vs. size of chemical space for three types of ionic composition databases, generated by considering only simple binaries, using a grammatical construction and randomly combining elements, respectively (see text).

solids, due to the presence of highly coordinated atoms (e.g. transition metals).

A powerful catalyst string representation must therefore be able to handle non-locality and an enormous variety of elemental compositions. A recently proposed approach to overcome these challenges is to use a coordinate-free representation based on crystallographic Wyckoff positions.[14] By avoiding the definition of bonds between atoms altogether, this is potentially a viable route towards powerful string representations of catalysts. In the field of Zeolites and Metal-Organic Frameworks, the classification of network topologies offers similar advantages.[15,16] If a generally useful representation could be defined along these lines, this would give access to the wealth of techniques developed in natural language processing, both in terms of grammatical inference and machine learning (e.g. recurrent neural networks and transformers). Indeed, this type of interdisciplinary approach has recently led to significant advances in organic synthesis planning.[17,18]

Graph representations: An alternative route towards a catalyst grammar would be to use graphs instead of strings to represent the catalysts. Graphs have a long tradition for representing chemical structures in terms of atoms and their connectivity. Unlike strings, they can easily represent cycles, branches and other non-local features of arbitrary complexity. Moreover, graphs and nets (their periodic equivalent) are already used to characterize inorganic solids such as Zeolites, Metal-Organic-Frameworks and carbon allotropes.[19,20] A further advantage of graphs over strings is that they can in principle directly encode the relative positions of atoms in three dimensional space and are thus overall more expressive. Furthermore, it is easier to define meaningful measures of similarity for graphs than for strings, which can be important in ML applications.[21]

An analogous concept to formal grammars also exists for graphs. Such *graph grammars* use production rules that define how subgraphs can be modified and replaced.[22] As a downside, developing and using graph grammars is significantly more complicated, however, because they operate on a more complex type of object. This is particularly true for periodic graphs. We also note that defining bonds in inorganic solids is not always unambiguously possible, so that a straightforward graph representation based on valence rules is not necessarily equally well suited for all types of materials. However, it has been shown that graph neural networks are able to learn powerful graph representations in very diverse settings, without prior definition of chemical bonds.[23] A combination of generative grammars with graph-based ML may therefore be a promising route.

Validity: As noted above, the central advantage of using a grammar in the context of catalyst discovery is that it allows the exclusive generation of syntactically 'valid' candidates, thus avoiding the unnecessary consideration of 'invalid' ones. We have so far been fairly vague about what is meant by valid structures, however. Indeed, this is not clear and depends on the context. In the case of SMILES, validity simply means that the valence rules of organic chemistry are not violated. This implies that the corresponding molecules will also be reasonably stable in most (but not all) cases. Relying on valence rules alone is unlikely an adequate concept of validity for the full periodic table though, not least due to the ambiguous nature of chemical bonds. Similarly, the charge balance condition used in the toy example above is not sufficient to guarantee stability and only applies to ionic materials. Yet another type of validity criterion can be defined based on atomic or ionic radii, as e.g. used in the Goldschmidt tolerance factor for perovskites.[24] In principle, a combination of these different validity measures could be encoded in a formal grammar, while in general and as noted above 'valid' could simply mean 'consistent' with available partial understanding.

An alternative approach would be to infer validity from data instead of defining it *a priori*. This could be achieved by treating a database of known 'valid' compounds (i.e. stable compounds or active catalysts) and view these as a corpus of examples generated from an unknown underlying grammar. The corresponding grammar could then be learned using the methods of *grammar induction* (also known as grammatical inference).[25] In this setting, a broader concept of validity (beyond e.g. mere stability) could in principle be obtained. For example, one could construct the database of examples to only include systems with certain conditions (such as adequate band-gaps for photocatalysts). This approach could also incorporate a notion of synthesizability into the grammar, which has recently been demonstrated to be a learnable property.[26]

Relation to Generative Deep Learning: The above already implies a close relationship between formal grammars and generative ML models. In particular, deep learning approaches like Generative Adversarial Networks or Variational Autoencoders have recently been the focus of intense study in materials design.[27] While such models can be extremely powerful tools for exploring chemical space they tend to require large amounts of data for training. Furthermore, the generation of unphysical or invalid structures and so-called 'mode collapses' (i.e. models which do not cover the full space of relevant structures but only generate highly similar outputs) are frequently observed issues that can be difficult to debug.

The grammatical approach outlined herein is in many ways complementary to such deep generative models. A simple generative grammar can be defined with very little reference data or derived from physical concepts (partial understanding) like charge neutrality. Furthermore, 'mode-collapse' is not an issue, as the grammar can simply be sampled uniformly. On the flipside, deep generative models are currently a much more powerful and mature technology. In this context we can again take cues from the related field of molecular design, where it has been shown that the robust SELFIES grammar can be used both to enhance the quality of deep generative models and as a competitive generative model in its own right.[10,28]

Conclusions: In this perspective, we have discussed the potential benefits of using formal grammars to discover new heterogeneous catalysts. This approach is intellectually stimulating, though it may seem slightly frivolous at first glance. Considering the leading role that string representations and grammars play in molecular design, we firmly believe that this can lead to real advances in catalyst discovery, however. To achieve this goal, we have sketched several promising research directions. These include the development of powerful string representations for solids and surfaces, the use of graph-based grammars and the combination of grammars with deep generative models.

References

(1) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37–46.

(2) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-Principles-Based Multiscale Modelling of Heterogeneous Catalysis. *Nat. Catal.* **2019**, *2*, 659–670.

(3) Bligaard, T.; Nørskov, J. K.; Dahl, S.; Matthiesen, J.; Christensen, C. H.; Sehested, J. The Brønsted-Evans-Polanyi Relation and the Volcano Curve in Heterogeneous Catalysis. *J. Catal.* **2004**, *224*, 206–217.

(4) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To Address Surface Reaction Network Complexity Using Scaling Relations Machine Learning and DFT Calculations. *Nat. Commun.* **2017**, *8*, 14621–14621.

(5) Andersen, M.; Levchenko, S. V.; Scheffler, M.; Reuter, K. Beyond Scaling Relations for the Description of Catalytic Materials. *ACS Catal.* **2019**, *9*, 2752–2759.

(6) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(7) Chomsky, N. Three Models for the Description of Language. *IRE Trans. Inf. Theory* **1956**, *2*, 113–124.

(8) Chomsky, N. On Certain Formal Properties of Grammars. *Information and Control* **1959**, *2*, 137–167.

(9) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

(10) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

(11) Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W. D.; Gamst, A. C.; Persson, K. A.; Ceder, G. The Thermodynamic Scale of Inorganic Crystalline Metastability. *Sci. Adv.* **2016**, *2*, e1600225.

(12) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; Walsh, A. Computational Screening of All Stoichiometric Inorganic Materials. *Chem* **2016**, *1*, 617–627.

(13) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Skelton, J. M.; Morita, K.; Walsh, A. SMACT: Semiconducting Materials by Analogy and Chemical Theory. *J. Open Source Softw.* **2019**, *4*, 1361.

(14) Goodall, R. E. A.; Parackal, A. S.; Faber, F. A.; Armiento, R.; Lee, A. A. Rapid Discovery of Novel Materials by Coordinate-Free Coarse Graining. *ArXiv210611132 Cond-Mat Physicsphysics* **2021**,

(15) Ockwig, N. W.; Delgado-Friedrichs, O.; O'Keeffe, M.; Yaghi, O. M. Reticular Chemistry: Occurrence and Taxonomy of Nets and Grammar for the Design of Frameworks. *Acc. Chem. Res.* **2005**, *38*, 176–182.

(16) Shevchenko, V. Y.; Krivovichev, S. V. Where Are Genes in Paulingite? Mathematical Principles of Formation of Inorganic Materials on the Atomic Level. *Struct. Chem.* **2008**, *19*, 571.

(17) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583.

(18) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nat. Mach. Intell.* **2021**, *3*, 144–152.

(19) Delgado-Friedrichs, O.; O'Keeffe, M. Identification of and Symmetry Computation for Crystal Nets. *Act. Crys.* **2003**, *A59*, 351–360.

(20) Strong, R. T.; Pickard, C. J.; Milman, V.; Thimm, G.; Winkler, B. Systematic Prediction of Crystal Structures: An Application to sp^3-Hybridized Carbon Polymorphs. *Phys. Rev. B* **2004**, *70*, 045101.

(21) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *ArXiv180204364 Cs Stat* **2019**,

(22) Rozenberg, G. *Handbook of Graph Grammars and Computing by Graph Transformation: Volume 1: Foundations*; WORLD SCIENTIFIC, 1997.

(23) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 3185–3191.

(24) Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **1926**, *14*, 477–485.

(25) D'Ulizia, A.; Ferri, F.; Grifoni, P. A Survey of Grammatical Inference Methods for Natural Language Learning. *Artif. Intell. Rev.* **2011**, *36*, 1–27.

(26) Jang, J.; Gu, G. H.; Noh, J.; Kim, J.; Jung, Y. Structure-Based Synthesizability Prediction of Crystals Using Partially Supervised Learning. *J. Am. Chem. Soc.* **2020**, *142*, 18836–18843.

(27) Kim, B.; Lee, S.; Kim, J. Inverse Design of Porous Materials Using Artificial Neural Networks. *Sci. Adv.* **2020**, *6*, eaax9324.

(28) Nigam, A.; Pollice, R.; Krenn, M.; dos Passos Gomes, G.; Aspuru-Guzik, A. Beyond Generative Models: Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) Algorithm for Molecules Using SELFIES. *Chem. Sci.* **2021**, *12*, 7079–7090.