Chemical Reactant Recommendation Using a Network of Organic Chemistry

John Savage*
Nuritas, Ireland

Akihiro Kishimoto
IBM Research, Ireland

Beat Buesser
IBM Research, Ireland

Ernesto Diaz-Aviles†
Libre AI Labs, Ireland

Carlos Alzate
IBM Research, Ireland

ABSTRACT

This paper focuses on the task of recommending to the chemist candidate molecules (reactants) necessary to synthesize a given target molecule (product), which is a novel application as well as an important step for the chemist to find a synthesis route to generate the product. We formulate this task as a link-prediction problem over a so-called Network of Organic Chemistry (NOC) that we have constructed from 8 million chemical reactions described in the US patent literature between 1976 and 2013. We leverage state-of-the-art factorization algorithms for recommender systems to solve this task. Our empirical evaluation demonstrates that Factorization Machines, trained with chemistry-specific knowledge, outperforms current methods based on similarity of chemical structures.

CCS CONCEPTS

• **Information systems → Recommender systems**;

KEYWORDS

Chemistry; reactant recommendation; link-prediction

1 INTRODUCTION

Chemistry is a scientific field where humans have to gain experience and intuition over many decades to be able to routinely discover new insights. The number of experienced chemists is small and a limiting factor in large industries like pharmaceuticals, nutrition, or materials that rely heavily on synthetic manufacturing of chemical molecules. Next to other factors this includes finding a feasible reaction pathway to synthesize a target molecule. Due to the significant human effort only a small number of feasible synthesis routes are usually discovered out of an incredibly large combinatorial space of possible chemical reactions, which limits the discovery of new and the optimization of exiting drugs and materials. Despite 40 years of efforts, e.g., [2, 9, 10, 15], computer aided design has not been successful to instigate widespread routine use by chemists. This can be attributed in part to the difficulty of building the generalized chemistry knowledge required for plausible analyses [15]. Currently, the main tools for designing synthesis routes are the Network of Organic Chemistry (NOC) [9] and rule-based synthesis planners [15]. The NOC is limited by the fact that it can only provide information on previously known reactions, while rule-based synthesis planners require rules designed by experienced humans and therefore can be overly rigid.

While collaborative filtering has been commonly used to recommend commodity items such as movies and music to a large number of users [12, 13], this paper is as far as we know the first attempt to employ such recommender systems to provide professional level decision support in chemistry. The main difficulty in using recommender systems in chemistry is the definition of recommendation accuracy. For movie recommendations or other similar applications, the cost of a bad recommendation is low. However, in chemistry the recommendation of a poor molecule could result in days or weeks of wasted time and money for a chemist.

This paper focuses on the task of recommending candidate reactants that could be used to synthesize a target product. We formulate this task as a link-prediction problem over a large-scale graph based on the NOC, a graph representation of the reactions in a database. New links predicted in the NOC contain the reactants recommended to a chemist for further consideration. Unlike the NOC, our approach can help the chemist find a new synthetic route. In addition, our approach can do so without using predefined over-rigid rules of the rule-based synthesis planners.

We apply to the NOC two factorization models that learn personalized ranking [21], which are closely related to collaborative filtering: one based on Matrix Factorization [12], and the second based on Factorization Machines (FMs) [20]. We enhance our models with chemical fingerprints representing chemistry-specific knowledge about the structure and properties of molecules.

We have constructed our NOC based on 8 million chemical reactions extracted from the US patent literature between 1976 and 2013 for performance evaluation. Our results indicate that FMs in combination with chemical fingerprints perform better than the similarity baseline which justifies further research to improve the accuracy based on collaborative filtering for applications in chemistry.

2 OUR APPROACH

Our approach regards chemical products and reactants as users and items of the recommender system, respectively, and attempts to recommend an item (i.e., reactant) to a user (i.e., product designed by a chemist). The product can be either an existing molecule or new molecule which does not exist yet in the world.

*This work was performed while Savage was an intern at IBM Research, Ireland.
†This work was performed while Diaz-Aviles was affiliated with IBM Research, Ireland.

RecSys '17, August 27-31, 2017, Como, Italy

ACM ISBN 978-1-4503-4652-8/17/08…$15.00
https://doi.org/10.1145/3109859.3109895

We first construct a large-scale NOC graph from many chemical reaction equations, then learn a model over NOC based on factorization models, and predict a link from a product to a reactant. We further add fingerprints of chemical molecules as attributes to the model, thus counteracting the cold-start problem of collaborative filtering, that is, for the case of which the target product has a new molecule structure is not therefore available in NOC.

Construction of NOC

Both products and reactants are represented as nodes in the large-scale NOC graph. We construct NOC as shown previously [1]. A directed link from node A to node B indicates that product B is synthesized from reactant A. For example, suppose that reaction $H_2 + O_2 \rightarrow H_2O$ is available. Then, NOC includes three nodes H_2O, H_2 and O_2 with two links $H_2 \rightarrow H_2O$ and $O_2 \rightarrow H_2O$.

Using all molecules that appear in all reactions results in constructing NOC with many connections to small, common, relatively uninteresting molecules, and leads to returning unmeaningful recommendations. For example, including H_2O in NOC and recommending it to a chemist is not insightful even if H_2O is used for a synthesis of the target.

As a simple yet practical choice, in each reaction we choose the heaviest reactant and heaviest product and connect these two molecules in NOC, since these heavy molecules tend to have complicated structures. An ideal way would be to construct a function that accurately evaluates the importance of the molecules in the reaction and include only important ones, which is future work.

NOC does not distinguish the reactants from the products, since a molecule can be a product as well as a reactant in practice. When a chemist selects a node in a graph, that node is marked as a product the chemist needs to synthesize. If the chemist designs a new product, it is added to NOC and marked as a product. The remaining nodes are the reactants that could be recommended to the chemist.

Chemical Fingerprints

A lot of work has been done in the chemoinformatics community on representing molecules *in silico* for the purposes of searching through chemical databases for similar molecules [26], since similar molecules often share common chemical characteristics. The molecular fingerprints used for this purpose represent molecules as bit vectors which indicate the presence or absence of molecular substructures. Our approach integrates the fingerprints with the factorization models to effectively capture the characteristics of the chemistry domain (see the next subsection).

Among many choices in the literature, we chose to use the MACCS fingerprint that is one of the most simple, standard versions [3], because it is readily available in open source chemoinformatics codes such as Open Babel [18]. The MACCS fingerprint consists a list of 166 chemical structures, such as a carbon atom, a carbon-oxygen bond or a benzene ring, represented as a bit vector of length 166, with bits set to 1 if the structure is present in the molecule.

Factorization Models

Our approach regards NOC as a training set. With this training set, our approach learns a model based on factorization models. We assume that the same reactant would be used if two products have similar sets of the reactants used to synthesize them in the past.

In the reactant recommendation task, positive training examples (i.e., successful reactions) are available such as in scientific reporting [5] and patent literature. However, no negative examples (i.e., failed reactions) are usually publicly available as global knowledge. In addition, considering the feature vector size and the number of available examples, our approach needs to perform machine learning under sparsity. We employ and evaluate the following two approaches that attempt to resolve these issues.

Matrix Factorization based on BPR. Bayesian Personalized Ranking (BPR) is a Matrix Factorization (MF) approach that deals with the case where only positive training examples are available [21]. In BPR, training examples are constructed as $D = \{(u, i, j)\}$, where i is an item already used by user u (i.e. known successful reaction) and j is an item which has not been used by u yet (i.e., unobserved whether a reaction will be successful or failed). That is, BPR only assumes that for each user an observed item is preferred to an unobserved item. Then, when two unobserved items are given to a user, BPR attempts to predict which item is preferred by that user. That is, by training a model with D, rather than calculating an accurate preference score of a single item, BPR attempts to accurately rank a pair of two unobserved items for each user by learning model parameters based on maximum a posteriori estimation.

We employ standard MF combined with BPR as Rendle et al. present [21]. Let q_i and p_u be two vectors representing the latent factors of item i and user u, respectively. The score of recommending item i to user u is approximated by the dot product: $\hat{r}_{ui} = q_i^\top p_u$.

The major issue in using this pure collaborative filtering technique in these scenarios is the cold-start problem, where there are not enough user-item interactions to provide good preference score estimates [6]. This is particularly important in the use case presented here, as the molecules chemists are interested in synthesizing have generally not been seen before. To solve the cold-start problem, we added attributes of users or items based on chemical fingerprints, bit vectors that capture characteristics of a molecules.

To incorporate fingerprints into MF, we used the Linear Mapping method from [6]. This method sets the number of factors to the number of attributes N_a and learns a mapping from the user (product) attributes to user factors, as well as the item (reactant) factors, and approximates the true user-item score matrix as before.

Factorization Machines. We also employ Factorization Machines (FMs) [19], since FMs are known to be able to address the problem with huge sparsity, which is the case of our recommendation task. We use FM combined with aforementioned BPR to perform the ranking task with only positive training examples [20].

FMs inherit advantages of both MF and Support Vector Machines (SVMs) [25]. Like SVMs, since FMs work as a general predictor with any real valued feature vector, we can easily add chemistry-specific features to FMs. Like MF, FMs model all interactions between variables using factorized parameters.

In the FM model, the users and items are represented in a feature vector $\mathbf{x}$ along with any other features deemed useful and the outcome y as the rating, or 1 in the case of positive training examples.

Table 1: Details of Dataset.

# Reactants	# Products	# Reactions
700,000	800,000	8,000,000

The model then becomes:

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left\langle \mathbf{v}_i, \mathbf{v}_j \right\rangle x_i x_j \quad (1)$$

with the model parameters to be estimated being: $w_0 \in \mathfrak{R}, \mathbf{w} \in \mathfrak{R}^n, \mathbf{V} \in \mathfrak{R}^{n \times k}$, and $\langle \cdot, \cdot \rangle$ is the dot product.

As shown in Figure 1, our feature vector consists of the user (product) elements, item (reactant) elements, and attributes. Each element in the feature vector is either 1 or 0. If a reaction in NOC indicates that reactant R is used to synthesize product P, the bits are set corresponding to P and R in the product and reactant elements, respectively. We use the fingerprints of a product and a reactant as attributes of FMs.

product	reactant	MACCS
one-hot encoding	one-hot encoding	fingerprint

Figure 1: Factorization Machines features.

3 EMPIRICAL EVALUATION

For our empirical evaluation we use a reaction database from the text mining of US patents from 1976–2013 provided by Lowe[1] [16].

The full reaction database was split in time to test whether knowledge of all previously patented reactions could predict reactions patented in the future. The split was made at January 1st 2013, and the numbers of reactions in the split are shown in Table 1.

With the training done on a random 100,000 subset of the full training data, the predictions were tested on the reactions in 2013. An ideal prediction would have a high recall@N for low N, and we set $N = 10$. We also show the results for the area under the receiver operating characteristic curve (AUC) as this is analogous to the metric that BPR optimizes on.

We do not show the precision@N, since it may not be a good measure of prediction quality. Note that the training and test set are not a "gold-standard", there may be many molecules that are good recommendations for a given molecule but are not included in the reactions in the training or test set.

We compared the performance of BPR-based MF and FMs. We also included an approach that recommends molecules that are similar to the target product, calculated by Jaccard similarity using the MACCS fingerprints.

We set the size of the latent dimensions in FMs to $k = 128$, and the latent vector size in MF to $k = 16$. We chose these numbers by cross validation in preliminary experiments. In case of MF with the attributes, the size of the latent vector was identical to the size of the MACCS fingerprint (i.e., $k = 166$). All algorithms were implemented by using libFM [20] and MyMediaLite [7]. The experiments were run on a Linux/GNU machine with 2.6GHz 12-core processor with 80GB of RAM.

[1] https://bitbucket.org/dan2097/patent-reaction-extraction/downloads

Results

We conducted a first round of experiments in a highly sparse NOC using all reactions and products in the dataset (either rare or common). Then, we conducted a set of experiments in a more realistic setting based on a denser NOC, which uses reactions with products that have been created more than 10 times in the patent dataset.

Highly Sparse NOC. We start with the results under sparsity. In this setting the MF approach underperformed the similarity-based recommendation, followed by the results of FMs that performed slightly better.

Recall@10 and AUC shown in Figure 2 clearly indicate that Matrix Factorization performed poorly. For pure MF the recall@10 score was zero, indicating that MF completely failed to recommend meaningful reactants due to the sparsity of the dataset. Since patents mainly file molecules that have new chemical structures, there are many examples of pairs of molecules appearing only once in the literature.

The similarity-based approach performs well since many reactions tend change only a small part of the chemical structure of a reactant to synthesize products. However, the low recall@10 score of 0.26 indicates that similarity-based approaches are not sufficient for reactant prediction. The AUC score indicates that MF had a problem training its parameters well enough to perform a good prediction. MF could not overcome the cold-start problem due to the fact the attribute interactions are learned on a per-reactant basis. As we can see from the performance of the similarity-based approach, the attributes of chemicals should interact consistently in reactions, and so we should be able to learn the parameters for these interactions and apply them globally, which we can do using Factorization Machines.

Figure 3 compares FMs against the similarity-based approach. "FM with attributes" indicates that FMs include the MACCS fingerprints as attributes, as in Figure 1. We also included an FM model that uses only fingerprints in its feature vector, denoted as "FM with ONLY attributes". With this FM model, we intended to evaluate the performance without the collaborative filtering characteristics.

FMs performed much better than MF. The high AUC scores indicate the success of model training. The full FM model returned a slightly better recall@10 score than the similarity-based approach (0.27 versus 0.26). When only the fingerprints were used for the FMs, the FM model worsened the recall@10 score. This shows that FMs exploiting the connectivity of NOC is important for improving their predictive power.

The reason why the full FM model achieved only small improvement to the similarity-based recommendation can be explained again by the data sparsity of the patent dataset caused by many reactions that appear only once.

Dense NOC. In practice, NOC would be more connected as the more common reactions would be included. Chemists would repeatedly use these reactions, and such common reactions would help connect the peripheral reactions to construct the core of the network.

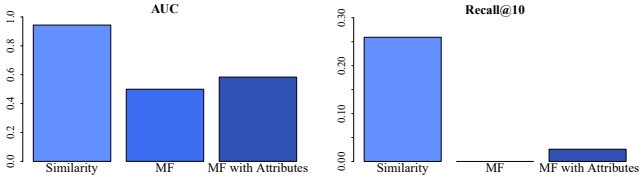

Figure 2: MF results on a sparse NOC.

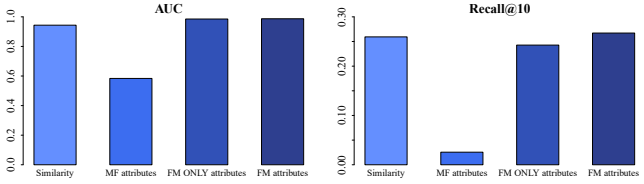

Figure 3: FMs results on a sparse NOC.

To approximate this scenario, we constructed NOC by only using reactions with products that have been created more than 10 times in the patent dataset.

Figure 4 shows the results of training the FMs on this denser NOC. The FM model with no attribute features shows that simple collaborative filtering performs better than similarity-based recommendation (recall@10 0.20 vs 0.13) in this denser network. FM with ONLY attributes performed closely to the similarity-based recommender (recall@10 = 0.14).

The full FM model successfully yielded a much better recall@10 score (0.36), showing the additive performance of the collaborative filtering and chemical attributes modeling. The full FM model inherits advantages of both MF and chemistry-specific knowledge. The former advantage of FMs accounts for the network connectivity of NOC by performing collaborative filtering, while the latter means it can learn about interactions between chemical structures in each pair of molecules. In this general dataset, it will learn parameters so it performs comparably to the similarity-based approach. However, in datasets containing limited numbers of reaction types, it should learn interactions specific to those reaction types.

4 RELATED WORK

Collaborative filtering has attracted great interest in recommender systems for movie rental and streaming platforms, internet-based retailers, etc. See [12, 13] for the literature review about the Matrix Factorization algorithm and its variants.

Computer aided design has been investigated for several tasks in chemistry since many decades with first rules of organic chemistry by Corey in 1969 [2]. The work closely related to our approach is graph-based analysis of existing reactions, reaction pathway discovery based on rule-based synthesis planners, and chemical reaction prediction, which we review below.

Grzybowski et al. [9] have generated the Network of Organic Chemistry (NOC), a graph representation of the chemical reactions for analyzing existing reactions [8, 14]. However, the NOC can only provide information on previously known reactions.

Several researchers have developed rule-based synthesis planners that return a reaction pathway to synthesize a target molecule designed by the chemist e.g., [10, 15]. But rule-based synthesis

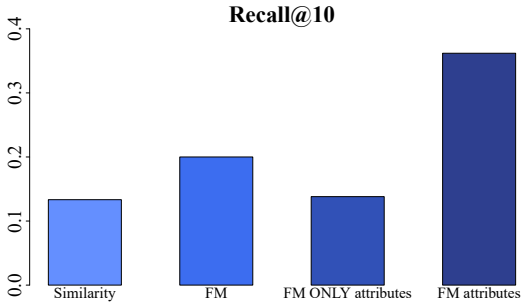

Figure 4: FMs results on a dense NOC.

planners rely heavily on reaction templates that describe the substructures of reactants and products of known reactions. While parameters of the reaction templates can be tuned by machine learning [24], rule-based synthesis planners are not able to find pathways that would require previously unseen reaction templates.

There are approaches to predict reactions to synthesize a new chemical product from given reactants, e.g., [11, 23], including the development of novel fingerprints [22]. However, these approaches do not enable the chemists to specify the target products they are interested in synthesizing. In contrast, our task is to predict for a given target molecule a ranked list of candidate reactants to synthesize the target molecule. In other worlds, our approach solves a so-called inverse problem to reaction prediction, while reaction problem is a forward problem. The results on reaction template parameter tuning [24] imply that our inverse problem would be more difficult to solve than their forward problem.

Other related work includes approaches to discover new enzyme reactions in the biology network [17, 27]. These approaches predict new links only between existing enzymes by addressing only the forward problem in the biology network (i.e., predicting an enzyme reaction in a normal direction).

5 CONCLUSIONS AND FUTURE WORK

We presented an approach to adapt factorization-model-based link-prediction techniques to the task of recommending chemical reactants to the chemists. Our approach based on FMs with chemical fingerprints showed promising results, especially when performing reactant predictions in highly-connected NOC, where the chemists would work on more commonly than the very sparse set of patent documents.

Our recall@10 scores indicate that the accuracy of the prediction still needs to be improved. On the other hand, since this is the first attempt to apply factorization models to the reactant recommendation domain in chemistry, followup research will appear which would achieve significant improvements. Ideas yet to explore as future work include:

- an incorporation of recently developed chemical fingerprints learned by Deep Learning (e.g., [4]),
- a combination with rule-based planners and reaction prediction technologies, and
- a better approach to extracting important molecules from reactions to include them in NOC.

REFERENCES

[1] Kyle J. M. Bishop, Rafal Klajn, and Bartosz A. Grzybowski. 2006. The core and most useful molecules in organic chemistry. *Angewandte Chemie - International Edition* 45, 32 (2006), 5348–5354. https://doi.org/10.1002/anie.200600881

[2] E. J. Corey and W. T. Wipke. 1969. Computer-assisted design of complex organic syntheses. *Science* 166, 3902 (October 1969), 178–92. https://doi.org/10.1126/science.166.3902.178

[3] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Modeling* 42, 6 (November 2002), 1273–1280. https://doi.org/10.1021/ci010132r

[4] David K. Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. 2015. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Proceeedings of the 28th Annual Conference on Neural Information Processing Systems*. 2224–2232.

[5] Kerry Dwan, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, Carrol Gamble, Davina Ghersi, John P. A. Ioannidis, John Simes, and Paula R. Williamson. 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PloS one* 3, 8 (January 2008), e3081. https://doi.org/10.1371/journal.pone.0003081

[6] Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, and Lars Schmidt-Thieme. 2010. Learning attribute-to-feature mappings for cold-start recommendations. In *Proceedings of the 10th IEEE International Conference on Data Mining*. 176–185. https://doi.org/10.1109/ICDM.2010.129

[7] Zeno Gantner, Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2011. MyMediaLite: A Free Recommender System Library. In *5th ACM International Conference on Recommender Systems (RecSys 2011)*.

[8] Chris M. Gothard, Siowling Soh, Nosheen A. Gothard, Bartlomiej Kowalczyk, Yanhu Wei, Bilge Baytekin, and Bartosz A. Grzybowski. 2012. Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry. *Angewandte Chemie International Edition* 51, 32 (2012), 7922–7927. https://doi.org/10.1002/anie.201202155

[9] Bartosz A. Grzybowski, Kyle J. M. Bishop, Bartlomiej Kowalczyk, and Christopher E. Wilmer. 2009. The 'wired' universe of organic chemistry. *Nature Chemistry* 1, 1 (April 2009), 31–6. https://doi.org/10.1038/nchem.136

[10] Abraham Heifets and Igor Jurisica. 2012. Construction of New Medicines via Game Proof Search. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, Vol. 2. 1564–1570. http://www.scopus.com/inward/record.url?eid=2-s2.0-84868287208

[11] Matthew A. Kayala and Pierre Baldi. 2011. A Machine Learning Approach to Predict Chemical Reactions. In *Proceeding of the 25th Annual Conference on Advances in Neural Information Processing Systems*. 747–755. http://papers.nips.cc/paper/4356-a-machine-learning-approach-to-predict-chemical-reactions.pdf

[12] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37.

[13] Yehuda Koren and Robert M. Bell. 2015. *Recommender Systems Handbook*. Springer, Chapter Advances in Collaborative Filtering, 77–118.

[14] Mikolaj Kowalik, Chris M. Gothard, Aaron M. Drews, Nosheen A. Gothard, Alex Weckiewicz, Patrick E. Fuller, Bartosz A. Grzybowski, and Kyle J. M. Bishop. 2012. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angewandte Chemie (International ed. in English)* 51, 32 (August 2012), 7928–32. https://doi.org/10.1002/anie.201202209

[15] James Law, Zsolt Zsoldos, Aniko Simon, Darryl Reid, Yang Liu, Sing Yoong Khew, A. Peter Johnson, Sarah Major, Robert A. Wade, and Howard Y. Ando. 2009. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *Journal of Chemical Information and Modeling* 49, 3 (2009), 593–602. https://doi.org/10.1021/ci800228y

[16] Daniel M. Lowe. 2012. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. Dissertation. University of Cambridge.

[17] Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. *Machine Learning and Knowledge Discovery in Databases* 6912 (2011), 437–452. https://doi.org/10.1007/978-3-642-23783-6_28

[18] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. 2011. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 3, 1 (January 2011), 33. https://doi.org/10.1186/1758-2946-3-33

[19] Steffen Rendle. 2010. Factorization Machines. In *Proceedings of the 10th IEEE International Conference on Data Mining*. 995–1000. https://doi.org/10.1109/ICDM.2010.127

[20] Steffen Rendle. 2012. Factorization Machine with libFM. *ACM Transactions on Intelligent Systems and Technology* 3, 3 (2012), 1–22. https://doi.org/10.1145/2168752.2168771

[21] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR : Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Vol. cs.LG. 452–461. arXiv:1205.2618 http://dl.acm.org/citation.cfm?id=1795167

[22] Nadine Schneider, Daniel M. Lowe, Roger A. Sayle, and Gregory A. Landrum. 2015. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of Chemical Information and Modeling* 55, 1 (2015), 39–53. https://doi.org/10.1021/ci5006614

[23] Marwin H. S. Segler and Mark P. Waller. 2017. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chemistry – A European Journal* 1521, 3765 (2017).

[24] Marwin H. S. Segler and Mark P. Waller. 2017. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chemistry – A European Journal* 1521, 3765 (2017).

[25] Vladimir N. Vapnik. 1998. *Statisical Learning Theory*. Wiley.

[26] Peter Willett, John M. Barnard, and Geoffrey M. Downs. 1998. Chemical Similarity Searching. *Journal of Chemical Information and Modeling* 38, 6 (November 1998), 983–996. https://doi.org/10.1021/ci9800211

[27] Yoshihiro Yamanishi, Jean-Philippe Vert, and Minoru Kanehisa. 2005. Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinromatics* 21 (2005), 468–477.