OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–10
https://doi.org/10.1093/bib/bbac461
Advance access publication date 17 October 2022
Problem Solving Protocol

Multi-modal chemical information reconstruction from images and texts for exploring the near-drug space

Jie Wang †, Zihao Shen†, Yichen Liao, Zhen Yuan, Shiliang Li, Gaoqi He, Man Lan, Xuhong Qian, Kai Zhang and Honglin Li

Corresponding authors: Kai Zhang, School of Computer Science and Technology, Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200062, China. E-mail: kzhang@cs.ecnu.edu.cn; Honglin Li, Shanghai Key Laboratory of New Drug Design, East China University of Science & Technology, Shanghai 200237, China. Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200062, China. E-mail: hlli@ecust.edu.cn
†These authors contributed equally.

Abstract

Identification of new chemical compounds with desired structural diversity and biological properties plays an essential role in drug discovery, yet the construction of such a potential space with elements of 'near-drug' properties is still a challenging task. In this work, we proposed a multimodal chemical information reconstruction system to automatically process, extract and align heterogeneous information from the text descriptions and structural images of chemical patents. Our key innovation lies in a heterogeneous data generator that produces cross-modality training data in the form of text descriptions and Markush structure images, from which a two-branch model with image- and text-processing units can then learn to both recognize heterogeneous chemical entities and simultaneously capture their correspondence. In particular, we have collected chemical structures from ChEMBL database and chemical patents from the European Patent Office and the US Patent and Trademark Office using keywords 'A61P, compound, structure' in the years from 2010 to 2020, and generated heterogeneous chemical information datasets with 210K structural images and 7818 annotated text snippets. Based on the reconstructed results and substituent replacement rules, structural libraries of a huge number of near-drug compounds can be generated automatically. In quantitative evaluations, our model can correctly reconstruct 97% of the molecular images into structured format and achieve an F1-score around 97–98% in the recognition of chemical entities, which demonstrated the effectiveness of our model in automatic information extraction from chemical patents, and hopefully transforming them to a user-friendly, structured molecular database enriching the near-drug space to realize the intelligent retrieval technology of chemical knowledge.

Keywords: near-drug space, multi-modal learning, name entity recognition, image recognition

Introduction

The identification of new chemical compounds with desired structural diversity and biological properties (e.g. drug metabolism and pharmacokinetics) plays an essential role in drug discovery, and so pharmaceutical chemists are committed to constantly exploring the chemical space to identify useful drug candidates [1, 2]. Considering the volume of the chemical space with a combinatorial nature (10^{60} compounds obeying Lipinski's rule-of-five [3–5], Figure 1A), and the specificity of the drug space with only 2712 small molecules approved so far [6], a natural tradeoff is to study the drug-like space that lies in between the two. A prominent example of the drug-like space is the generic GDB-17 database, with around 10^{12} structures by enumerating virtual molecules containing up to 17 atoms [7]. Unfortunately,

Jie Wang is a PhD candidate from the School of Pharmacy at East China University of Science and Technology. Her research interests are drug design, cheminformatics and deep learning.
Zihao Shen is a PhD candidate from the School of Pharmacy at East China University of Science and Technology. His research interests are artificial intelligence, computer vision and deep learning.
Yichen Liao is a graduate student from the School of Pharmacy at East China University of Science and Technology. Her research interests are drug design, cheminformatics and deep learning.
Zhen Yuan is a PhD candidate from the School of Pharmacy at East China University of Science and Technology. She works on drug design, cheminformatics and molecular simulation.
Shiliang Li is an associate professor of pharmacy at East China University of Science and Technology and a young investigator of the Innovation Center for AI and Drug Discovery at East China Normal University. Her research interests are bioinformatics, cheminformatics, target discovery and drug design.
Gaoqi He is an associate professor in the School of Computer Science and Technology at East China Normal University. His research interests are computer vision, computer graphics and big data.
Man Lan is a professor in the School of Computer Science and Technology at East China Normal University. Her research interests are natural language processing, artificial intelligence and data mining.
Xuhong Qian is a professor and president of East China Normal University. His research interests are pharmaceutical chemistry, pesticide chemistry and dye chemistry.
Kai Zhang is a professor from the School of Computer Science and Technology at East China Normal University. His research interests are large-scale machine learning, time series analysis, dynamic brain networks and psychiatric diagnosis.
Honglin Li is a professor of medicinal chemistry and computational chemistry. He is the director of the Shanghai Key Laboratory of New Drug Design at East China University of Science and Technology and the director of the Innovation Center for AI and Drug Discovery at East China Normal University. His group is focused on artificial intelligence, target discovery and drug design, computational biology and cheminformatics.
Received: July 31, 2022. **Revised:** September 21, 2022. **Accepted:** September 26, 2022

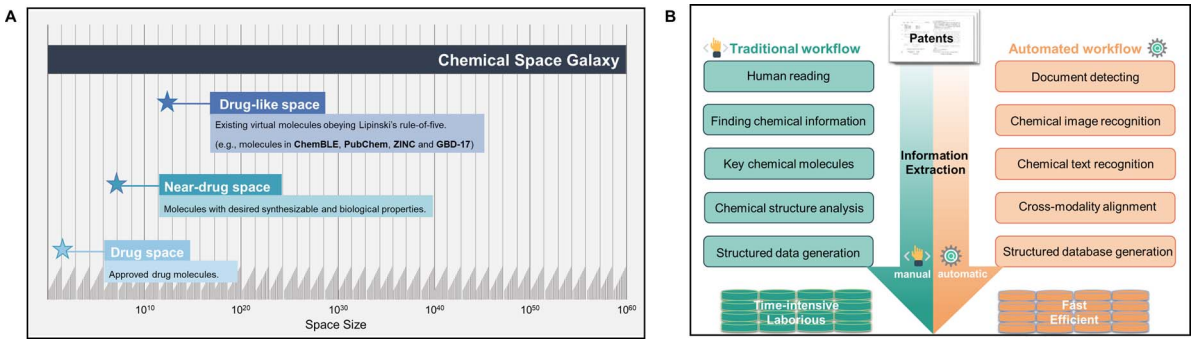

Figure 1. Chemical space composition and workflows of constructing the near-drug space. (**A**) Comparison of the drug space, known virtual drug-like space and novel near-drug space. (**B**) Workflows for information extraction from patents. Traditional workflow (left) involves significant manual operations, whereas our CIRS-system is more automated (right).

many molecules encompassed in the drug-like space could be difficult to synthesize or lacking desired drug-effects, and so how to build a 'near-drug' space composed of compounds with more synthesizable chemical structures and desired biological properties has become one of the most central goals toward improving the success rate and reducing the cost of the drug discovery process [8, 9].

The exponential growth of scientific publications and chemical patents has ushered in new opportunities in expanding and exploring the near-drug space [10]. In this work, we focus on chemical patents because the results disclosed in the patents could be more timely, reliable and comprehensive [11]. Moreover, patent documents cover a huge number of molecules with synthesizable structures and desired biological properties, which is particularly advantageous to finding useful chemical compounds [12].

The huge (and ever increasing) number of chemical patents and their complex, heterogeneous data organizations have made it a highly challenging endeavor to extract useful chemical structures from patents in an accurate and automatic way [13]. In particular, note that the major output of chemical patents are a mixture of text descriptions and image templates, a lot of efforts have been devoted to the development of scalable and accurate tools for recognizing named entities from the texts and chemical structures from the images.

Named entity recognition (NER) has been widely applied to the detection of chemical entities (compounds, proteins and diseases) or their relationship in text [14–16]. For instance, Leroy *et al*. developed an extendable chemical execution program to achieve an autonomous, 'paper in, chemical product out' workflow [17]. This subversive research has caused a surge of interest in extracting synthesis information from text, such as the paragraph2actions [18] proposed later, which can convert experimental procedures into action sequences. However, the outputs of these approaches are typically labeled texts, which cannot be utilized directly by pharmaceutical chemists but instead have to be further transformed into structural data (e.g. molecular structural database) [19–21]. Furthermore, annotated corpus for chemical structures in the patents is particularly limited, which can directly affect the training result of name entity recognition [22].

The molecular structure images in chemical patents, on the other hand, contain rich structured information (compounds and formulas) in visual forms, and image-processing algorithms have been widely applied to extract such information, also known as optical chemical structure recognition (OCSR). For example, OSRA is a chemical structure recognition software based on rules method with high accuracy [23]; DECIMER is dedicated to the recognition of structures in the images with more valid SMILES for conventional molecular structures [24]; Image2SMILES managed to identify unconventional atoms in chemical structures through data generator [25]. However, the accuracy of current approaches is still limited in the recognition of molecular structure images with special bonds and atoms. Furthermore, chemical patents are usually downloaded as XML, HTML or PDF, in which the images' quality is of low resolution and high noise, which makes it difficult to accurately extract the molecular structures [22, 26, 27].

The chemical information from different modalities of the patent, i.e. texts and images, should be exploited in coordination to deliver an accurate output. However, such a structural fusion remains an open challenge for researchers from both computational chemistry and artificial intelligence. Besides the difficulties arising from the recognition tasks in each individual domain, as noted above, the chemical information from language descriptions and graphical templates are of highly distinct format and statistical properties, and how to find the correspondence between the extremely huge number of chemical entities across the two domains in an accurate and automatic manner with minimal human intervention is the key challenge. In this regard, multimodal learning techniques have been used to build machine learning models that can process information from different modalities in such areas as image and voice recognition, and it currently has also been developed in the field of drug discovery [28, 29]. For example, ChemDataExtractor applied natural language processing and rule-based grammars to processing both chemical experiment texts and spectroscopic attributes tables [19]. KV-PLM, a unified pre-trained language model processing both molecule structures and biomedical text, assists drug discovery and documentation for biomedical research [30]. However, these methods are not suitable for identifying the implicit correspondence between chemical structure images and text descriptions for chemical information fusion.

The goal of this work is to build a multimodal chemical information reconstruction system (CIRS) to automatically process, extract and align heterogeneous information from patent texts and images, so as to facilitate the construction of chemical structure database with minimal human intervention. This would be a valuable tool from which pharmaceutical chemists could benefit significantly in the exploration and expansion of the near-drug space. Our key innovation lies in an advanced, heterogeneous data generator as the hub-module that produces cross-modality, yet tightly coupled chemical entities in the form of text descriptions and Markush structure images. On top of this data generator, a two-branch model will not only learn to recognize chemical entities accurately inside each domain, but will also

naturally capture the cross-modality correspondence embedded in the training data. By doing this, structural fusion of chemical entities becomes evident in the form of images and texts. We also make available a large structure database of chemical entity to convert chemical entities text into molecular structures, which can solve the difficulties arising from name entity recognition tasks mentioned above, whereas it also can provide the source of substituent structures for other researchers to furtherly explore combinatorial chemistry realm. Once this is achieved, the gap between the two domains can be filled to break the bottleneck of chemical information fusion, so that the vast amount of structural information in the form of text and image from patents can be effectively aligned with each other.

Materials and methods

Data collection and preprocessing

For Markush image recognition tasks, a set of chemical structures in an SMILES format were downloaded from ChEMBL database (version ChEMBL28). The RDKit software [31] was used to perform a washing procedure on the raw SMILES dataset containing 1 911 226 structures. Those structures that can't be retrieved by RDKit were removed; molecules with more than 50 heavy atoms were also removed because the molecule images would be too 'crowded' for processing. After the processing, the following random data split were used: (1) a training set of 150K images; (2) a validating set of 30K structure images and (3) a test set of 30K structure images.

In generating the Markush-like structure images, a number of commonly encountered functional group and R-group (see below) labels were used to replace the atoms in the molecule. RDKit was used to identify explicit hydrogen atoms and randomly replace them with the labels above. If one structure had one or more rings, R-groups and bonds crossing the ring bonds may be randomly added onto one or more rings to reproduce such a case in real chemical documents. Because RDKit can only generate aromatic rings in Kekulized style, the generated molecules were saved in the SVG format first. During the saving step, the image and atom label padding size, the bond line width and offset, the atom label font and the total rotation angles were randomly selected to generate highly diversified training images. The SVG strings were then parsed and aromatic rings in Kekulized style were randomly picked and converted into their aromatic styles (a ring with a circle in the middle). Finally, the SVG strings were rendered to generate the output PNG images. The coordinates and other additional basic primitive information of the atoms and bonds were also extracted by RDKit and used to create labels required by the semantic segmentation and object classification tasks.

The text data were collected from 2712 chemical patents in English, which were downloaded from the European Patent Office and the US Patent and Trademark Office under the keywords search of A61P, compound, structure and year from 2010 to 2020. The preprocessing mainly included intercepting the description text of the substituents and converting the intercepted text to editable text through the Optical Character Recognition (OCR) program [32]. A total of 2712 snippets were obtained, with 20 798 words. Among them, 2400 snippets were chosen as the training set and 312 snippets were chosen as the test set, which were annotated with YEDDA [33] in the BIOSE format [34, 35]. Then all snippets were annotated to two entities, including *example label (S-Entity)*, and *substituent type (S-component, B-component, M-component and E-component)* (Table 1). Finally, we manually created the substituent structure database, including the description, chemical name and SMILES string of 7781 substituents covered in the snippets, which is necessary to transform the text information to actual molecular structures.

Functional groups: Me, OMe, NHMe, Et, OEt, NHEt, Pr, OPr, NHPr, i-Pr, Bu, OBu, NHBu, i-Bu, s-Bu, t-Bu, Ph, OPh, NHPh, Tol, Ts, OTs, NHTs, Bz, NHBz, CF_3, CN, CHO, COOH, COOMe, COOEt, NHOH, NMe_2, NEt_2, N_3, NO_2, COCl, SOOMe, SOOEt, SOOPh, Bn, OBn, NHBn, Boc, OBoc, Cbz, OCbz, Tf, OTf, Piv, OPiv, Vin, All, TMS, OTMS, TBS, OTBS, THP, OTHP, TBDPS, OTBDPS, OMOM, TES, OTES, IPDMS, OIPDMS, DEIPS, ODEIPS, CIIS, OCIIS, TIPDS, TFA, OTFA, Fmoc, OFmoc, Alloc, OAlloc, Troc, OTroc, Teoc, OTeoc, Tr, OTr, DMTC, ODMTC, BPin, OLev, PMP, OPMP, PMB, OPMB, Bt, OMPA, Mes.

R-groups: R, $R^1 \sim R^{10}$, $R^a \sim R^e$, R', R", A, M, W, X, Y, Z, Ar, Hal, *, #.

Image-processing unit

The image-processing unit is composed of the semantic segmentation network and the classification network. The segmentation network is used to categorize each pixel into one of the following: the background, the atom or the bond, and store them in a segmentation map with pixel locations. We have used the UNet 3+ [36] for the semantic segmentation. It takes images of size 512 × 512 and can compute feature maps with the same size as the input image. The number of epochs is set to 15 with a batch of four images. Considering the imbalanced foreground (molecule) and background (empty pixel), the focal loss [37] was chosen as the loss function. The parameter of the UNet3+ network (space complexity) is 26.97M, and the time complexity is 798.68G in terms of FLOPs (floating-point operations per forward-evaluation). The classification network is chosen as the YOLO Object Detection Network (https://github.com/Okery/YOLOv5-PyTorch), in which atoms and bonds are detected and classified separately. The atoms were first located by performing non-maximum suppression based on the atom feature map, and then the geometrical centers of each atom point are calculated and recorded. The YOLO network takes the raw image and the center coordinates as inputs and predicts the types and charges of each primitive. The bond primitives are processed in a similar way. The number of epochs is chosen as 50 and the batch size is 16. The parameter and the FLOPs values of the YOLO network are 47.05M and 55.41G, respectively. Having identified the primitives in the input image, we will then integrate all the information (primitive types, charges, location and connectivity patterns) and transform the image into a molecule with structured format (such as SMILES) using RDKit.

Text-processing unit

The text-processing unit is composed of a sequence labeling network in order to recognize the chemical entities in patent. Here we have used BiLSTM combined CRF model, which not only captures bi-directional correlations as in BiLSTM but also inherits the capacity of CRF in extracting highly contextualized features, which has attracted extensive attention in this filed [38]. At the beginning of model input, each word token w_i in an input text sequence $w_1, w_2, \ldots, w_n$ is represented by a word-vector v_i using Word2Vec-based word embedding [39] that captures the semantic information of the input text and then fed into a BiLSTM encoder to convert it into a latent feature vectors h_i; the latent feature vector h_i is then transformed to a new representation p_i before being fed into a linear-chain CRF layer for NER label prediction [40], which is a task for detecting mentions of real-world entities from text and classifying them into predefined types. A cross-entropy loss is used and 10-fold cross validation is applied during training while the Viterbi algorithm is used for decoding. We have used a batch size of 64 sequences each with 256 tokens. Overall,

Table 1. Entity types for information extraction in the text-processing unit

Type	Description	Examples
S-Entity	Single R-group name	R1, R^a, X
S-component	Single substituent name	Methyl, benzyl, carbonyl
B-component	The beginning word of multi-word substituent name	'Branched' in term 'branched C1–C6 alkyl'
M-component	The middle word of multi-word substituent name	'C1–C6' in term 'branched C1–C6 alkyl'
E-component	The end word of multi-word substituent name	'alkyl' in term 'branched C1–C6 alkyl'

the number of parameters is 0.51M, and the time complexity is 3.07M in terms of FLOPs. The Adam optimizer is employed to optimize network weights [41].

Evaluation metrics

The metrics used to evaluate the models are precision (the ratio between correctly predicted mentions over the total set of predicted mentions for a specific entity), recall (the ratio of correctly predicted mentions over the actual number of mentions) and F1-score (the harmonic mean between precision and recall), as defined in Equations (1)–(3):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1} = \frac{2 \bullet \text{precision} \bullet \text{recall}}{\text{precision} + \text{recal}} \tag{3}$$

Here TP is the true positive rate, FP is the false positive rate and FN is the false negative rate.

Results and discussion

Overview of CIRS

We propose CIRS, a multimodal chemical information reconstruction system for processing both chemical structure images and texts in patents to extract the structure of molecules. The whole architecture has three main branches, namely the image-processing unit (left), heterogeneous data generator (middle) and text-processing unit (right), as illustrated in Figure 2. The two branches on the left and right sides are models taking in images and texts from chemical patents, respectively; these two branches are implicitly connected through the heterogeneous data generator as the hub module in the middle, whose role is to generate paired training data across domains. As a result, during the training process, the two models will automatically learn to coordinate with each other in terms of both recognizing the chemical entities across domains, and aligning them together.

The training process proceeds as follows. First, the heterogeneous data generator will generate tightly coupled chemical entity pairs in the form of a Markush structure images and the (pixel-wise) atom/bond labels. This then serves as the training data to feed into the image-processing unit, where we have used a segmentation module (U-Net+++) in conjunction with a classification module (YOLO) in order to segment the pixels in into atoms and bonds and assign the correct label to them. In the right branch, the text-processing unit adopted a BiLSTM-CRF model is to perform name entity recognition to identify chemical entities (R-group and substituent) in texts. Then the outputs of the two branches, in particular the atom/bond labels from the left and the chemical entities from the right, are aligned with each other to reconstruct their chemical information. Finally, in order to practically convert chemical entities identified through the left model into visible chemical structures, we build a structure database including the substituent descriptions (chemical names) and SMILES strings of 7781 substituent structures.

It is worth mentioning that heterogeneous data generator as the hub-module produces the key correspondence between the chemical entities across modalities. A particular advantage of the generator is that there is not a strict limit on the amount and diversity of training samples, as it can modify the numerator randomly according to user requirements. This can then be translated to the good generalization performance of the image-processing and text-processing unit, which is the key to the applicability of our model in extracting chemical information from the huge amount of chemical patents. Therefore, the extracted structure from two-branch model can be automatically aligned and generalize to diverse molecular structures and their combinations in chemical patents.

To comprehensively investigate the performance of the proposed framework, we conduct a number of experiments to evaluate the recognition accuracies of the chemical entities by the text- and image-processing units in our model. Additionally, we perform a case study to demonstrate the potential of our framework in assisting automatic information extraction in real-world scenarios.

Markush chemical image recognition

Recognition of chemical structure images is crucial to information extraction from chemical patents. When considering the construction and exploration of the near-drug space, Markush objects with R-groups, placeholders or dummy atoms and labels that can be added to ordinary backbone structures are particularly useful and have been commonly used in chemical documents. Unfortunately, public datasets and related methods mostly target common structure images (complete molecules with no uncertain labels), which limits their applications. Raw molecule structure data were collected from the ChEMBL database and used to generate molecule images of Markush-typed structures (see the 'Materials and Methods' section). The datasets were then used to train the image-processing unit to convert the given images into their machine-processable molecule formats and validate its performance. Figure 3 shows several examples of the generated molecule images, the images mainly contain R-groups, functional groups, ring R bonds and random salt and pepper noise.

The image-processing unit is composed of a semantic segmentation module that groups the pixels in molecular images into meaningful primitives such as atoms and bonds, and a classification module that identifies the necessary information of the atoms and bonds. In the classification module, the targeted prediction can include the auxiliary information (such as atom

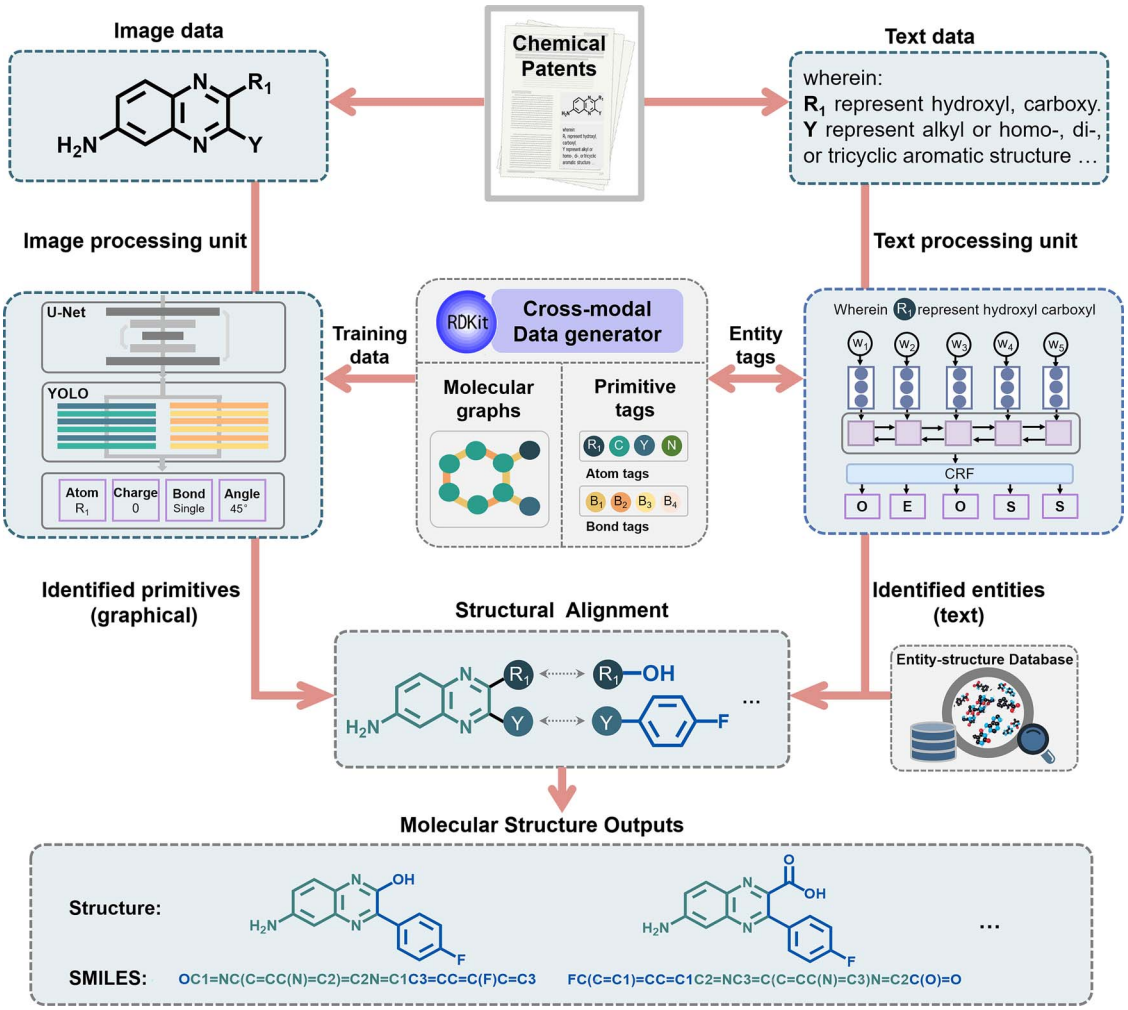

Figure 2. Workflow of CIRS. CIRS includes the image-processing unit (left branch), the cross-modal data generator (middle) and the text-processing unit (right branch). The two processing units take in the image and the text, respectively, from chemical patents; chemical entities are recognized in each modality and then aligned automatically to extract highly integrated information from patents, so as to build a highly expandable, structured molecular database to enrich the near-drug space.

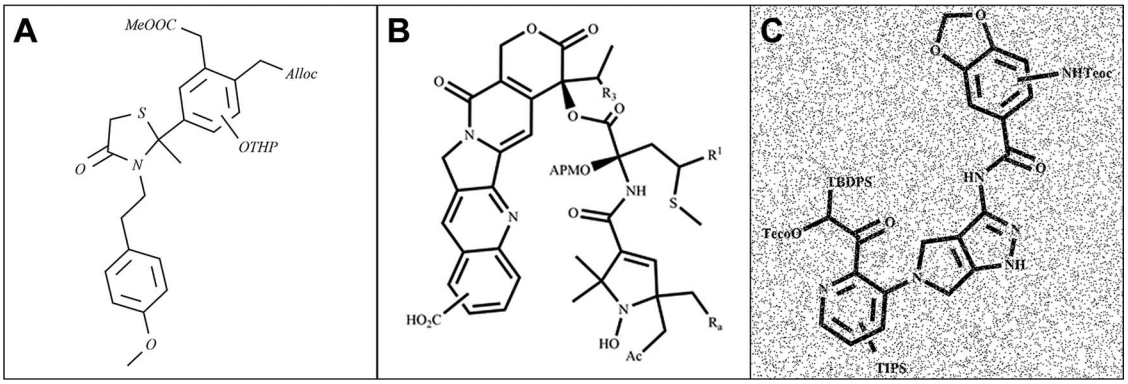

Figure 3. Examples of the Markush molecule image data. (**A**) Image containing several different functional-groups. (**B**) Ring R-bonds and common R-groups were added. (**C**) Salt-and-pepper noise applied in the training process.

types, atom charges, bond types, etc.) of the basic primitives. The coordinates of each estimated primitive are calculated and saved so they can be used to match the positions of the atoms with the positions of the bonds' endpoints. Then the connections between the atoms can be built, and the data are used to reconstruct the molecule structure and make the final output. The workflow of the model is shown in Figure 4A, and more details can be found in the 'Materials and Methods' section.

The performance of the image-processing unit is evaluated using two sources of datasets: (1) the artificial molecular images generated from our cross-modal data generator (by replacing implicit hydrogen atoms in the molecules to functional groups

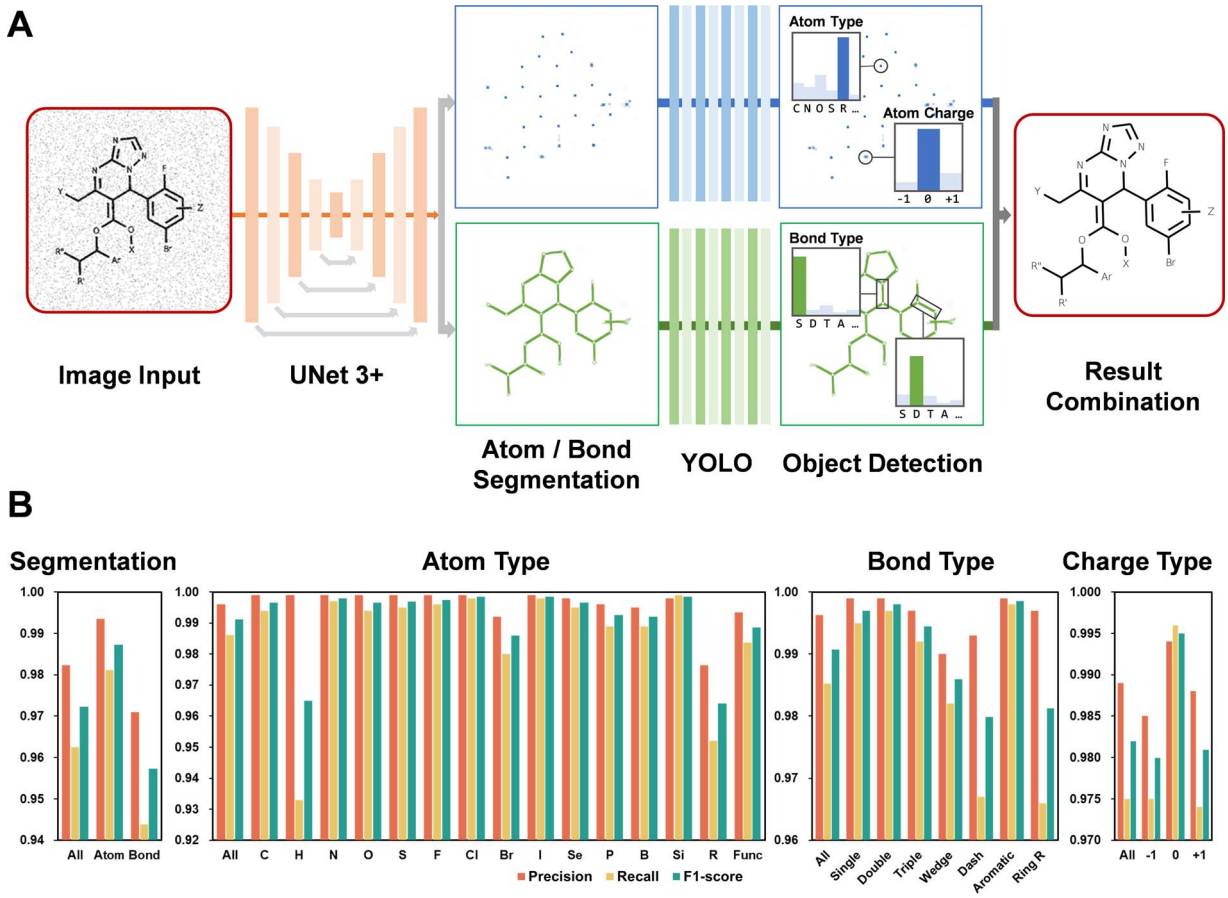

Figure 4. The image-processing unit of CIRS for semantic segmentation and classification. (**A**) Illustration of the image-processing unit architecture. (**B**) Performance of the image-processing unit in terms of precision, recall and F1-score.

and R-groups), with overall 30 000 images and (2) an external, MolrecUOB dataset [42] with 5740 real-world (noisy) images from real chemical documents with functional groups, and R-groups included. In both cases, the goal is to identify chemical primitives and predict their labels (atom/bond types, charges, etc.), and reconstruct the structure of the molecule in SMILES format based on the connectivity patterns of identified primitives.

The performance on the artificial dataset is reported in Figure 4B, where the semantic segmentation module can accurately identify the atoms and bonds, even with the presence of salt-and-pepper noises. The module achieves a pixel-wise precision of 0.982, which indicates that it can efficiently detect the positions of the atoms and bonds in the images. For the atom classification module, the precision of finding the correct atom types exceeds 0.996 on average, whereas the precision of the R-group detections is slightly lower (0.976), as the R-group styles and formats are commonly changeable: the corner marks attached to the 'R' labels (such as number 5 in label R^5) can be numbers, symbols or characters. As for the chemical bonds, the performance of the classification module is also high (0.996). The most frequent failure is the confusion of the wedge, dash and ring R-bonds, as the wedge and ring R-bonds may look similar to single bonds. The dash bonds are sometimes ignored by the model because of its lower visibility than common bonds, especially in case of a high level of background noise. The prediction accuracy of the atom charges is about 0.989. Charge symbols with small font sizes may get slightly lower precisions. The good performance of the model was mainly attributed to the high-quality training image generation. Our data generator can provide high-quality images containing various functional groups, R-groups and other structure styles, which enables the model to learn from highly diversified training images to perform segmentation and object detection. Besides, we also perturbed the training data to enhance the robustness of the model against noisy images. These two factors, in combination with state-of-the-art model UNet3+ and YOLO networks, have finally generated promising results in chemical information extraction and reconstruction from molecular images.

For the artificially generated image dataset and MolrecUOB dataset, the 'Ratio of Properly Reconstructed Images' is defined as the portion of molecular images for which our model can accurately reconstruct the SMILES representation; we also adopted the Tanimoto similarity, a commonly used metric to estimate the similarity between molecule structures. The performance is reported in Table 2. As can be seen, our model can correctly reconstruct the structures of 79% of the input MolrecUOB images, with a nice Tanimoto similarity score of 0.90, indicating that the model has acceptable generalization ability on real data. Note that the accuracy and the Tanimoto similarity metric were much higher on our generated dataset, which was 0.972 and 0.982, respectively. This is because the MolrecUOB dataset included some style of representations that were not covered in our data generator, examples including (1) the number of atoms in the carbon chains or on the rings might be unknown (e.g. '–$(CH_2)_n$–' indicating n chain-linked carbons); (2) nested super-atom labels with numbers (e.g. '$(CH_3CH_2)_2N$–' indicating two ethyl groups on atom N) and (3)

Table 2. Performance of the image-processing unit on the generated Markush images and a real (external) molecular image dataset

	Ratio of properly reconstructed images[a]	Tanimoto[b]
Generated Images (30,000 samples)	0.972	0.982
MolrecUOB Images (5740 samples)	0.791	0.904

[a]The ratio of correctly reconstructed images (all the atoms, bonds and charges in the image are correctly recognized). [b]The average Tanimoto similarity metric.

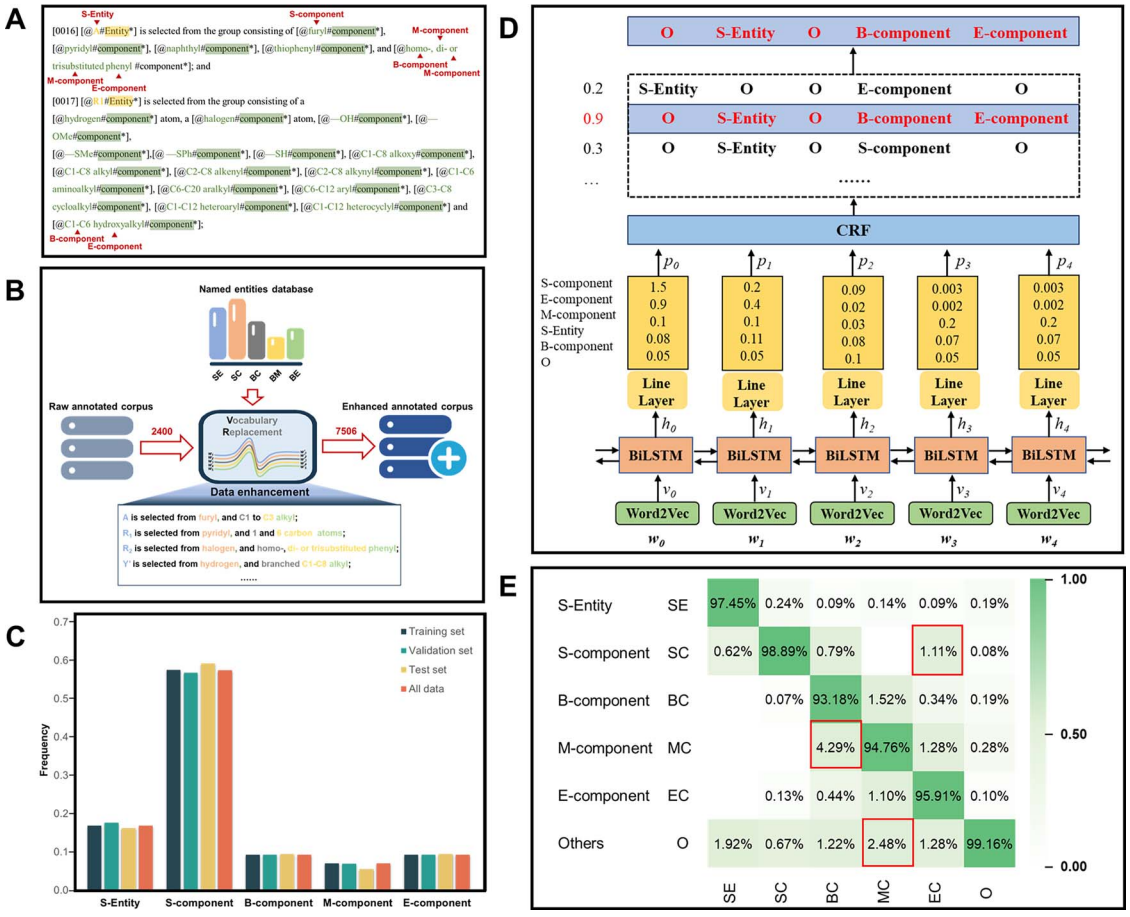

Figure 5. The text-processing unit and model evaluation in chemical entity recognition. (**A**) Annotated snippets for NER with some *substituent types (S-Entity, S-component, B-component, M-component and E-component)*. (**B**) Data enhancement protocol, which converted 2400 raw snippets into 7506 snippets. (**C**) Distribution of the chemical entities in the training, validation and test set. (**D**) Illustration of the test-processing unit, the BiLSTM-CRF architecture. (**E**) The confusion matrix of entity prediction; the *(ij)*th entry of the matrix signifies the portion of ith-type entities that are predicted as the *j*th-type entity. Dominant diagonal entities indicate an accurate prediction.

wavy bonds. It is worthwhile to note that those uncovered type of representations can be easily incorporated in our data generator, which would lead to a training dataset with wider coverage and so an improved reconstruction can be expected.

Chemical entity recognition

Recognition of meaningful chemical entities in the patent text and transforming them to predefined labels is the key to alignment of the chemical entities across the text and image modalities. As shown in Figure 5D, the BiLSTM-CRF model was chosen to detect substituent entities, which can learn the feature of entity processed by word embedding through LSTM and consider the correlation between the front and back of the sequence by CRF, more details can be found in the 'Materials and Methods' section. We note that the annotated corpus for chemical structure texts in the patents is particularly limited, which can directly affect the generalization capacity of the trained model. To address this question, we have downloaded a wide spectrum of 2712 chemical patents from the European Patent Office and the US Patent and Trademark Office, and collected entity texts and annotated training (2400 snippets) and test (312 snippets) sets for a total of 20 798 words (Figure 5A, see the 'Materials and Methods' section). During the training phase, we expanded the training set from 2400 to 7506 snippets by replacing the substituent entity with other substituent entity through the same label (Figure 5B), which can better tune hyper parameters of our models, respectively. The data set was split into train and test sets, and as a result of this new setting, 6755 snippets were available in train set, 751 in the validation set and 312 in test set. Figure 5C shows the entity distribution in appears similar for different datasets. The majority of the annotations are rom *S-component*, covering 57% of entities in the development phase. In contrast, *M-component* represent 7.0% of entities in the development phase.

We compared our BiLSTM-CRF model with other baseline methods for chemical name entity recognition, including LSTM

Table 3. Entity recognition performance of our model and two competing methods in terms of precision, recall and F1-score

Model	Type	Precision	Recall	F1-score
LSTM	Entity	0.79	0.90	0.84
	Component	0.90	0.94	0.92
LIME	Entity	0.90	0.91	0.91
	Component	0.85	0.90	0.88
BiLSTM-CRF	Entity	**0.97**	**0.96**	**0.97**
	Component	**0.98**	**0.98**	**0.98**

The metrics are evaluated on the test set, and the best values are indicated in bold.

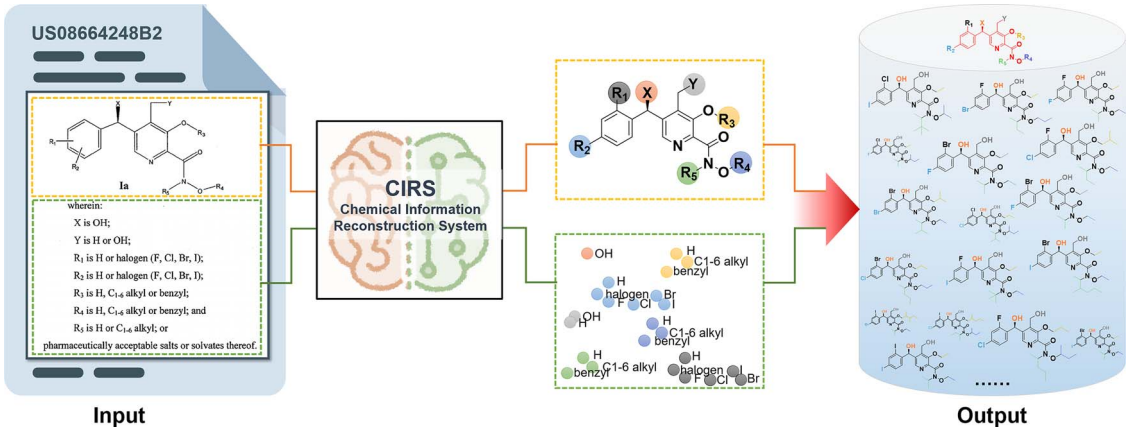

Figure 6. Automatic structure extraction from a chemical patent using CIRS. Our system generated a structural library with 2 082 500 molecules by finding the chemical entities in the patent and enforcing the replacement/combination rules as stated in its formula; in comparison, there were only 11 compound examples in the original patent.

[43], LIME [44]. Table 3 shows the performance of different approaches in terms of the precision, recall and F1-score. Note that the BiLSTM-CRF model trained on the enhanced dataset (through vocabulary replacement) achieves 97% of F1-score in the 'entity' type, and 98% of F1-score in the 'component' type, yielding more than six-point improvement over LIME and LSTM models in term of the F1-score. The superior performance of our model should be attributed its capacity of combing the bi-directional correlations captured in BiLSTM, as well as the highly contextualized temporal features extracted by the CRF model. Besides, the data augmentation or enhancement can also generally improve the performance of learning-based algorithms. See 'Materials and Methods' for details of the enhancement protocols.

Figure 5E shows the confusion matrix to illustrate the percentage of entities predicted by the BiLSTM-CRF model against the ground truth. As can be observed, the confusion matrix has a dominant diagonal, indicating that the correctly predicted entity types dominate. The top 2 best-performing entities identified by our models are *S-compound* and *S-Entity*. Most of the misclassification is associated with the *B-component* and *M-component*, which may be attributed to the insufficient training instances of such entities in the dataset (Figure 5C). Other than that, *E-component* is sometimes classified as *S-component*. Moreover, many predicted entities were shorter in length, for example, *aromatic structure* instead of *homo-, di, or tricyclic aromatic structure* and *alkyl* instead of *linear alkyl*. Indeed, the multi-word entities are a major challenge for NER [45]. We believe it could be also related to subword tokenization. Lastly, our model can be further improved by introducing more features, such as part-of-speech, lemma, Roman numerals, names of the Greek letters [46], which could better characterize chemical entities in more complex patent texts.

Case study of chemical information reconstruction

In this subsection, we demonstrate the practicability of CIRS through a case study, in which we have chosen a specific patent [47], extract the chemical entities from its images and text descriptions, align the entities together and finally transform the reconstructed information into a structured molecular database. The patent contains around 4 formulas and 11 molecular images to present their compounds invention. We selected a formula (Ia) for chemical information extraction to demonstrate the practicality of CIRS. As shown in Figure 6, formula (Ia) consists of two parts: Markush molecule image and substituent entity text. Through CIRS, a Markush structure and eight chemical entities with 123 substituent structures were extracted from images and texts, separately, and, as a result, 2 082 500 molecules were obtained by the aligning the chemical entities across the text and image modalities, and enforcing the replacement/combination rules as stated in the patent formula. This is a significant enrichment as compared to the 11 molecular examples reported in the original patent. As can be seen, our system can extract the chemical findings in the patent and transform them into a highly comprehensive collection of molecules with desired replacement rules to reconstruct their chemical information. This can serve as a useful molecular database for drug screening. As can be anticipated, by applying our system to the vast number of chemical patents, we can then obtain a significant number of structures to facilitate the generation of near-drug molecules, and hopefully construct a useful near-drug space for pharmaceutical

chemists to work with. Furthermore, automatic information extraction from chemical patents can make it much easier to define the scope of patent coverage, so as to better avoid chemical intellectual property conflicts in drug discovery in the future.

Conclusions

In summary, we explored a multi-modal chemical information reconstruction system, named CIRS, by identifying chemical entities from the texts and images of chemical patents, aligning them automatically, so as to facilitate the exploration and construction of the near-drug space. This is achieved through the parallel image and text-processing units that explore the two modalities, respectively, and in the meantime being connected with each other via the use of a cross-modal data generator, so that their predicted primitives can be naturally aligned for chemical information reconstruction. In quantitative evaluations, our accuracy (F1-score) in terms of correctly identifying chemical entities from Markush structure images and texts approaches 96% and 97%, respectively, meaning that the accuracy of aligning chemical entities from the two modalities will be around 96–97%, under various distributions of the chemical entities. This demonstrates the value of the proposed model in automatic information processing, extraction and reconstruction. Furthermore, our work showed that CIRS is a promising system in facilitating automatic information extraction, which could generate structures from images and texts in scientific literature to enrich near-drug space and boost drug discovery. Also, our automatic information extraction system may help to establishing a knowledge base (e.g. knowledge atlas) to realize the intelligent retrieval technology of chemical knowledge.

In our future research, we will consider more diversified features in chemical documents, such as part-of-speech, lemma, Roman numerals, names of the Greek letters; pursue more elastic solutions-based OCSR to process various types of R and functional groups, instead of using pre-defined fixed class/type lists; in addition, we are studying how to effectively incorporate chemistry or bioinformatics information from knowledge resources with more flexible formats and organization than patents (such as scientific papers) in the information reconstruction process. Also, we may explore how to directly obtain SMILES strings based on natural language descriptions, and further improve the generalization ability of the model by considering more types of chemical data to enrich the near-drug space.

Key Points

- We proposed a multimodal chemical information reconstruction system (CIRS) to automatically process, extract and align heterogeneous information from text descriptions and structural images of chemical patents, so that useful and expandable molecular structures toward drug discovery can be constructed efficiently to populate the near-drug space.
- A heterogeneous data generator can produce cross-modality training data, from which parallel processing units can learn to both recognize chemical entities from different modalities and simultaneously capture their correspondence; such an automatic information fusion framework and data-generative mechanism can be valuable for a great variety of chemical information mining and reconstruction applications.
- Comprehensive experiments demonstrate the effectiveness of our model in automatic information extraction from chemical patents and enriched structural library with a significantly larger number of candidate compounds can be generated from the patents.

Data availability

The chemical structures in SMILES format from ChEMBL database are available at https://www.ebi.ac.uk/chembl/. The chemical entity recognition datasets were collected from the European Patent Office (EPO) and United States Patent and Trademark Office (USPTO) under the keywords search of A61P, compound and structure.

Acknowledgements

We would like to thank the graduates Chonghui Wang, Luoyi Zhuo, Kaiyue Lian, Fei Xia, and Li Xu from East China University of Science and Technology for their help in the data processing stage.

Funding

The National Natural Science Foundation of China (grants 81825020, 82150208 to H.L.); Lingang Laboratory (grant LG-QS-202206-02 to S.L.); the National Program for Special Supports of Eminent Professionals (to H.L.); the National Program for Support of Top-notch Young Professionals (to H.L.).

References

1. Drews JJ. Drug discovery: a historical perspective. *Science* 2000;**287**(5460):1960–4.
2. Caron G, Digiesi V, Solaro S, *et al.* Flexibility in early drug discovery: focus on the beyond-Rule-of-5 chemical space. *Drug Discov Today* 2020;**25**:621–7.
3. Dobson CM. Chemical space and biology. *Nature* 2005;**432**:824–8.
4. Bohacek RSMC, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* 1996;**16**:3–50.
5. Coley CW. Defining and exploring chemical spaces. *Trends Chem* 2021;**3**:133–45.
6. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2017;**46**:D1074–82.
7. Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des* 2013;**27**:675–9.
8. Hert J, Irwin JJ, Laggner C, *et al.* Quantifying biogenic bias in screening libraries. *Nat Chem Biol* 2009;**5**:479–83.
9. Gromski PS, Henson AB, Granda JM, *et al.* How to explore chemical space using algorithms and automation. *Nat Rev Chem* 2019;**3**:119–228.
10. Hoffmann T, Gastreich M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discov Today* 2019;**24**:1148–56.

11. Seeber F. Patent searches as a complement to literature searches in the life sciences—a how-to' tutorial. *Nat Protoc* 2007;**2**: 2418–28.
12. González-Medina M, Naveja JJ, Sánchez-Cruz N, *et al.* Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv* 2017;**7**: 54153–63.
13. Saber AA, Hinnerk R, Markus S, *et al.* Automatic identification of relevant chemical compounds from patents. *Database (Oxford)* 2019;**2019**:baz001.
14. Jessop DM, Adams SE, Willighagen EL, *et al.* OSCAR4: a flexible architecture for chemical text-mining. *J Chem* 2011;**3**:1–12.
15. Rocktäschel T, Weidlich M, Leser U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* 2012;**28**: 1633–40.
16. Lowe DM, Sayle RA. LeadMine: a grammar and dictionary driven approach to entity recognition. *J Chem* 2015;**7**:S1–5.
17. Leroy C, Craven M, Leonov AI, *et al.* A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* 2020;**370**:101–8.
18. Vaucher AC, Zipoli F, Geluykens J, *et al.* Automated extraction of chemical synthesis actions from experimental procedures. *Nat Commun* 2020;**11**:3601–12.
19. Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J Chem Inf Model* 2016;**56**:1894–904.
20. Steiner S, Wolf J, Glatzel S, *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* 2019;**363**(6423):eaav2211.
21. Segler MH, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 2018;**555**: 604–10.
22. Akhondi SA, Klenner AG, Tyrchan C, *et al.* Annotated chemical patent corpus: a gold standard for text mining. *PLoS One* 2014;**9**:e107477.
23. Filippov IV, Nicklaus MC. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J Chem Inf Model* 2009;**49**:740–3.
24. Rajan K, Zielesny A, Steinbeck C. DECIMER: towards deep learning for chemical image recognition. *J Chem* 2020;**12**:65–74.
25. Khokhlov I, Krasnov L, Fedorov M, *et al.* Image2SMILES: transformer-based molecular optical recognition engine. *Chem Methods* 2022;**2**:e202100069.
26. Rajan K, Zielesny A, Steinbeck C. DECIMER 1.0: deep learning for chemical image recognition using transformers. *J Chem* 2021;**13**: 61–77.
27. Weir H, Thompson K, Woodward A, *et al.* ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning. *Chem Sci* 2021;**12**:10622–33.
28. Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. *J Mach Learn Res* 2012;**15**:2949–80.
29. Guy S, Lior R, Bracha S, *et al.* Explainable multimodal machine learning model for classifying pregnancy drug safety. *Bioinformatics* 2021;**38**:1102–9.
30. Zeng Z, Yao Y, Liu Z, *et al.* A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nat Commun* 2022;**13**: 862–73.
31. Landrum. RDKit: open-source cheminformatics. *Release* 2022; **03**:5.
32. Smith R.. An overview of the Tesseract OCR engine. In: *Proceedings of the Ninth International Conference on Document Analysis & Recognition (ICDAR 2007)* IEEE, 2007, pp. 629–33. https://dl.acm.org/doi/10.5555/1304596.1304846.
33. Yang J, Zhang Y, Li L *et al. YEDDA: A Lightweight Collaborative Text Span Annotation Tool*. 2018. http://arXiv:1711.03759v03753.
34. Reimers N, Gurevych I. *Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks*. 2017. http://arXiv:1707.06799.
35. Carpenter B. Coding Chunkers as Taggers: IO, BIO, BMEWO, and BMEWO +. LingPipe Blog 2009. http://lingpipe-blog.com/2009/10/14/coding-chunkersas-taggers-io-bio-bmewo-and-bmewo/.
36. Huang H, Lin L, Tong R *et al. UNET 3+: A Full-Scale Connected UNet for Medical Image Segmentation*. 2020. http://arxiv:2004.08790.
37. Lin T, Goyal P, Girshick R, *et al.* Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**:318–27.
38. Ma X, Hovy E. *End-to-End Sequence Labeling via Bi-directional LSTM-CNNS-CRF*. 2016. http://arXiv:1603.01354.
39. Naili M, HabachaChaibi A, Hajjami H, *et al.* Comparative study of word embedding methods in topic segmentation. *Proc Comput Sci* 2017;**112**:340–9.
40. Lafferty J, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning, Proceedings of Machine Learning Research*, ICML, 2001, pp. 282–9. http://repository.upenn.edu/cis_papers/159.
41. Kingma DP, Ba J. *Adam: A Method for Stochastic Optimization*. 2014. http://arXiv:1412.6980.
42. Sadawi NM, Sexton AP, Sorge V. Chemical structure recognition: a rule-based approach. *Proc SPIE* 2012;**8297**:32–41.
43. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
44. Ribeiro MT, Singh S, Guestrin C. "Why should I Trust You?": explaining the predictions of any classifier. In: *The 22nd ACM SIGKDD International Conference, Proceeding of Knowledge Discovery and Data Mining*, ACM, 2016, pp. 1135–44. https://dl.acm.org/doi/10.1145/2939672.2939778.
45. Copara J, Naderi N, Knafou J *et al. Named Entity Recognition in Chemical Patents Using Ensemble of Contextual Language Models*. 2020. http://arXiv:2007.12569v12562.
46. Kim S, Kim W, Comeau D *et al.* Classifying gene sentences in biomedical literature by combining high-precision gene identifiers. In: *The 2012 Workshop on BioNLP, Proceeding of Biomedical Natural Language Processing*, BioNLP, 2012, pp. 185–92. https://dl.acm.org/doi/10.5555/2391123.2391148.
47. Stranix BR, Milot G, Bouchard JE. Derivatives of pyridoxine for inhibiting HIV integrase. *US08664248B2* 2014.