Bioinformatics, 35, 2019, i183–i190
doi: 10.1093/bioinformatics/btz366
ISMB/ECCB 2019

pNovo 3: precise *de novo* peptide sequencing using a learning-to-rank framework

Hao Yang[1,2], Hao Chi[1,2,*], Wen-Feng Zeng[1,2], Wen-Jing Zhou[1,2] and Si-Min He[1,2,*]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing. Technology, Chinese Academy of Sciences, Beijing 100190, China and [2]University of Chinese Academy of Sciences, Beijing 100049, China

*To whom correspondence should be addressed.

Abstract

Motivation: *De novo* peptide sequencing based on tandem mass spectrometry data is the key technology of shotgun proteomics for identifying peptides without any database and assembling unknown proteins. However, owing to the low ion coverage in tandem mass spectra, the order of certain consecutive amino acids cannot be determined if all of their supporting fragment ions are missing, which results in the low precision of *de novo* sequencing.

Results: In order to solve this problem, we developed pNovo 3, which used a learning-to-rank framework to distinguish similar peptide candidates for each spectrum. Three metrics for measuring the similarity between each experimental spectrum and its corresponding theoretical spectrum were used as important features, in which the theoretical spectra can be precisely predicted by the pDeep algorithm using deep learning. On seven benchmark datasets from six diverse species, pNovo 3 recalled 29–102% more correct spectra, and the precision was 11–89% higher than three other state-of-the-art *de novo* sequencing algorithms. Furthermore, compared with the newly developed DeepNovo, which also used the deep learning approach, pNovo 3 still identified 21–50% more spectra on the nine datasets used in the study of DeepNovo. In summary, the deep learning and learning-to-rank techniques implemented in pNovo 3 significantly improve the precision of *de novo* sequencing, and such machine learning framework is worth extending to other related research fields to distinguish the similar sequences.

Availability and implementation: pNovo 3 can be freely downloaded from http://pfind.ict.ac.cn/software/pNovo/index.html.

Contact: smhe@ict.ac.cn or chihao@ict.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Shotgun proteomics research based on mass spectrometry data focuses on high-throughput peptide and protein identification. The main method is using dedicated sequence databases to identify peptides and proteins, such as SEQUEST (Eng *et al.*, 1994), Mascot (Perkins *et al.*, 1999), MaxQuant/Andromeda (Cox and Mann, 2008), PEAKS DB (Zhang *et al.*, 2012) and pFind (Chi *et al.*, 2015, 2018). Despite its indisputable popularity, database search still needs reference databases to retrieve peptide candidates, hence it cannot search for species without any proteome databases such as microbial communities (Hettich *et al.*, 2013) or unknown proteins such as monoclonal antibodies (Reichert *et al.*, 2005). Even for searching against known sequences, amino acid mutations, post-translational modifications (Chick *et al.*, 2015) and splice variants (Zhu *et al.*, 2014) are still hard to be identified by the existing database search strategies.

An alternative method for peptide and protein identification is *de novo* sequencing, which infers amino acid sequences directly from tandem mass spectra. *De novo* sequencing does not need any reference databases, so it has an irreplaceable advantage for identifying novel protein sequences. For example, many studies have used *de novo* sequencing methods to assemble monoclonal antibodies (Bogdanoff *et al.*, 2016; Guthals *et al.*, 2017; Tran *et al.*, 2016). Over the past decades, many *de novo* sequencing algorithms for shotgun proteomics have been proposed, such as PEAKS (Ma *et al.*, 2003), PepNovo (Frank and Pevzner, 2005), pNovo (Chi *et al.*, 2013, 2010; Yang *et al.*, 2017) and Novor (Ma, 2015).

Although many *de novo* sequencing tools have been proposed, the precision of *de novo* sequencing is still questionable. Muth and Renard (2017) stated that only ~40% of *de novo* sequencing results were consistent with the database search results, in which the analyses on simulation datasets showed that the low precision of *de novo* sequencing was mainly caused by the abundant noise peaks and the low fragment ion coverage in tandem mass spectra, especially for the latter. When the fragment ion coverage decreased from 100% to 50%, the proportion of correctly sequenced peptides dropped from 80% to only 20%, suggesting that the precision of *de novo* sequencing is very sensitive to the fragment ion coverage, whose fundamental cause is that the lack of fragment ions makes the order of consecutive amino acids indistinguishable, e.g. if no supporting fragment ions are detected between the first two amino acids of the peptide *AE*HDK in the tandem mass spectrum, then *EA*HDK may be wrongly regarded as the *de novo* sequencing result of this spectrum without any addition information.

In order to discriminate the similar peptide candidates, there needs a more powerful scoring method to better rerank *de novo* sequencing results of each single spectrum and, in particular, the difference among spectra does not need to be considered. Learning-to-rank models (Bartell *et al.*, 1995; Liu, 2010) are suitable for solving this problem, which are also useful for many applications in information retrieval. Given one query, all webpages should be ranked by the relevance between the query and each webpage, which is quite similar to ranking peptides (webpages) for each given spectrum (query), regardless of the diversity among different spectra. In addition, deep learning has a continuous upward trend in many research fields, even in the hard decision problems such as the game of Go (Silver *et al.*, 2016). Also, a few studies based on deep learning in proteomics have been proposed recently. For example, DeepNovo (Tran *et al.*, 2017) uses convolutional neural networks and recurrent neural networks (Hochreiter and Schmidhuber, 1997) to learn features of tandem mass spectra for *de novo* sequencing, and pDeep (Zhou *et al.*, 2017) uses the bidirectional long short-term memory (Graves *et al.*, 2013) to predict the theoretical spectrum for one given peptide with a median Pearson similarity of over 0.9. Generally, researches based on deep learning are still not very common in proteomics community. In fact, deep learning can automatically learn high-levels of representation of complex data without predesigned features based on domain-specific knowledge, so this character can be used to learn the fragmentation pattern and other important features in tandem mass spectra and construct a universal learning-to-rank model to discriminate very similar results of *de novo* sequencing.

In this article, we developed a novel *de novo* sequencing algorithm, pNovo 3. Unlike the way of using deep learning directly in DeepNovo (Tran *et al.*, 2017), peptide candidates were generated firstly using the traditional dynamic programming approach (Yang *et al.*, 2017) in pNovo 3, and then a few features were extracted based on the prediction results of pDeep (Zhou *et al.*, 2017) by deep learning, as well as other information related to the fragmentation patterns. Finally, a learning-to-rank model, trained by SVM-rank (Joachims, 2002; Joachims *et al.*, 2009), was built to rerank the peptide candidates generated previously.

In addition, a spectrum merging method was proposed to merge the results of spectra with similar precursor ion masses to further improve the performance of pNovo 3. Compared with three other state-of-the-art *de novo* peptide sequencing tools, the recall of pNovo 3 was increased by 29.4–96.1% at the full-length peptide level and 2.0–20.1% at the amino acid level on seven test datasets with different species. In addition, the recall of pNovo 3 was 20.6–49.8% higher than that of DeepNovo on nine other datasets, proving the significant improvement on the precision of *de novo* sequencing by using deep learning and learning-to-rank.

pNovo 3 can now be freely downloaded from the following website: http://pfind.ict.ac.cn/software/pNovo/index.html.

2 Materials and methods

pNovo 3 uses the same approach as pNovo+ and Open-pNovo (Chi *et al.*, 2013; Yang *et al.*, 2017) to get top-ranked peptide candidates for each spectrum, and then it has four steps to rerank the preliminary results. First, the theoretical spectrum for each candidate is predicted by pDeep (Zhou *et al.*, 2017) based on the deep learning approach. Second, features are extracted based on the results of pDeep and other statistics. Third, peptide candidates are reranked by the model trained by learning-to-rank (Joachims, 2002; Joachims *et al.*, 2009). Last, the results of the whole dataset are updated using the spectrum merging method. The workflow of pNovo 3 is shown in Figure 1. Before the introduction of the pNovo 3 workflow, we will first introduce the seven benchmark datasets, one of which was used in the following steps of model training.

2.1 Generating the benchmark datasets

Seven high-resolution datasets are used in this study. The first five datasets are acquired from the Thermo Scientific Q Exactive with the HCD activation mode (Cassidy *et al.*, 2016; Hu *et al.*, 2016; Nevo *et al.*, 2017; Paiva *et al.*, 2016; Seidel *et al.*, 2017) and the last two datasets are acquired from the Thermo Scientific Q Exactive HF-X (the latest MS instrument in the benchtop Orbitrap series) with the HCD activation mode (Kelstrup *et al.*, 2018). These datasets are from a wide variety of species to ensure an unbiased evaluation on different samples. All datasets can be downloaded from the ProteomeXchange website and the details are shown in Supplementary Table S1. The first one (*Vigna mungo*, *V.mungo*) is used for training the learning-to-rank model while the other six ones are used for the performance evaluation.

pFind (Chi *et al.*, 2015, 2018) and PEAKS DB (Zhang *et al.*, 2012) are used to search the seven datasets mentioned above against the proteins of the corresponding sample downloaded from the UniProt database in 2017.9. The search results of pFind and PEAKS DB are filtered with the false discovery rate (FDR) of 1% at the peptide level and the peptide-spectrum match (PSM) level, respectively. The detailed database search parameters of pFind and PEAKS DB are shown in Supplementary Table S2. To build the benchmark datasets, the inconsistent PSMs reported by pFind and PEAKS DB are removed. In addition, as the current version of pDeep cannot predict the theoretical spectra of modified peptides, PSMs with modified peptides except those with carbamidomethylation of cysteines only are removed. The retained PSMs consistently reported by pFind and PEAKS DB are used as the ground truth, and the corresponding MS/MS data are extracted from the original datasets for evaluating the performances of different *de novo* sequencing algorithms in the Section 3.

2.2 Generating theoretical spectra by pDeep

The intensity information is essential for calculating the similarity between a theoretical spectrum and a real spectrum. However, in most scoring methods for PSMs, the intensities of all peaks in theoretical spectra are set to equal values or only by a few simple rules, which makes it difficult to distinguish among different orders of consecutive amino acids if no fragment ions are observed among them. In order to

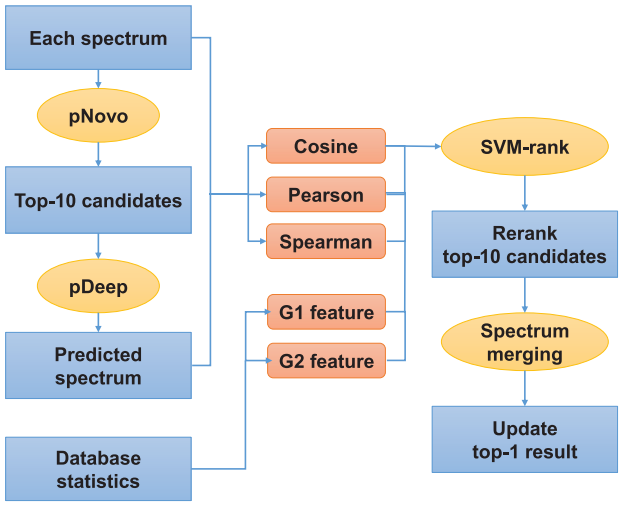

Fig. 1. The workflow of pNovo 3

show the importance of the intensity information, we take two peptide candidates, P_1: GTFSGLESSS*PE*VK and P_2: GTFSGLESSS*EP*VK, for one spectrum, as a running example. If there are no fragment ions observed between PE and between EP, then the scores of the two corresponding PSMs should be identical to each other. However, the fragmentation patterns of the two peptides are quite different that the intensity of the *y*-ion between E and P should be much higher than that between P and E, which is in general in CID or HCD activation mode (Snyder, 2000). If such pattern can be used in generating theoretical spectra, then P_1 will be more probable to be the result of the spectrum because the absence of the fragment ions between P and E is more consistent with its real fragmentation pattern.

The training datasets of pDeep are from several published datasets (Chick *et al.*, 2015; Kulak *et al.*, 2014; Sharma *et al.*, 2015) of a wide variety of species produced by Q Exactive or Q Exactive HF. The theoretical spectrum predicted by pDeep is composed of the masses and intensities of all backbone theoretical ions, including *b* and *y* ion series with 1+ and 2+ charge states. Assuming that $r_1, \ldots, r_n$ (n is the number of all ions) are the real intensities of all ions (b_1+, b_2+, …, b_1++, b_2++, …, y_1+, y_2+, …, y_1++, y_2++,…), $p_1, \ldots, p_n$ are the predicted intensities of the corresponding ions. $\bar{r}$ is the mean of $r_1, \ldots, r_n$, $\bar{p}$ is the mean of $p_1, \ldots, p_n$, $r'_1, \ldots, r'_n$ are the indexes of $r_1, \ldots, r_n$ if they are sorted in descending order, and $p'_1, \ldots, p'_n$ are the indexes of $p_1, \ldots, p_n$ if they are sorted in descending order, three measures of similarities, i.e. cosine, Pearson and Spearman between the theoretical and real spectra, are computed by formulas 1 to 3, respectively. The value of cosine similarity is from 0 to 1 and the values of the other two similarities are from −1 to 1.

$$SIM_{\cos} = \frac{\sum_{i=1}^{n} r_i p_i}{\sqrt{\sum_{i=1}^{n} r_i^2}\sqrt{\sum_{i=1}^{n} p_i^2}} \quad (1)$$

$$SIM_{pear} = \frac{\sum_{i=1}^{n} (r_i - \bar{r})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n} (r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n} (p_i - \bar{p})^2}} \quad (2)$$

$$SIM_{spear} = 1 - \frac{6\sum_{i=1}^{n} (r'_i - p'_i)^2}{n(n^2 - 1)} \quad (3)$$

2.3 Extracting gap features

The information of fragmentation gaps in PSMs is used to design features independently to the theoretical spectrum prediction of pDeep. In the running example, when the *b* and *y* ions fragmented between two consecutive amino acids (e.g. PE or EP) are not observed in the spectrum, we cannot distinguish between the two peptides without other information. However, we can compute the probabilities of losing fragment ions between PE and EP based on the statistics by using a large amount of the existing high-resolution MS/MS data. The probability of losing fragment ions between two consecutive amino acids *XZ* is defined by the number of *XZ* between which the ions are missing divided by the total number of *XZ* in the dataset used in the statistics. This probability is referred to as g_1, which is from 0 to 1. Specifically, considering that the order of the two N-terminal amino acids reported by *de novo* sequencing is often more error-prone (Fu and Li, 2005), we also compute the probability of losing fragment ions between the two N-terminal amino acids, which is referred to as g_2. Its value is also from 0 to 1. In the running example, the probabilities of losing fragment ions between EP and PE are 3.4% and 20.1%, respectively, indicating that the peptide with PE should be more confident than the peptide with EP.

Then, given one PSM, two features are generated based on g_1 and g_2, and then used in the learning-to-rank model. One is called G_1, which is the arithmetic mean value of all gaps g_1 found in the PSM. G_1 is set as 1.0 if no gaps are found, indicating that the PSM has no gaps. The other one is called G_2, which is equal to g_2 if there is an N-terminal gap, otherwise it is also set as 1.0. Take one peptide *P*: *GT*FSG*LE*SSS*PE*VK as an example, where three gaps, *GT*, *LE* and *PE* are detected and the first gap *GT* is the N-terminal gap. Then, two gap features are computed: $G_1(P) = \big(g_1(GT) + g_1(LE) + g_1(PE)\big)/3$ and $G_2(P) = g_2(GT)$. It is worth mentioning again that *GT* is involved in the feature extraction twice based on two different ways of data statistics, one is from all possible amino acid pairs and the other is only from the N-terminal ones. Therefore, the values of $g_1(GT)$ and $g_2(GT)$ may not be identical to each other.

2.4 Training the learning-to-rank model

Six features are finally extracted before model training, i.e. the original PSM score, the three similarities between the theoretical and real spectrum described in Section 2.2, and the two features, G_1 and G_2, related to gap information described in Section 2.3. Then, SVM-rank (Joachims *et al.*, 2009) is used to train the model for reranking top-ranked peptide candidates for each spectrum. All feature values are normalized to [0, 1] according to the value range of the corresponding feature of top-ranked peptide candidates for each spectrum. As mentioned in Section 2.1, the first benchmark dataset of *V.mungo* is used for training the model. For each spectrum in this benchmark dataset, pNovo+ (Yang *et al.*, 2017) is used to report *de novo* sequencing results and top-10 candidate sequences are retained. If the correct peptide, annotated by the database search results, is not contained in the top-10 candidate sequences for one spectrum, then this spectrum cannot be used in the model training; otherwise, the PSM with the correct peptide sequence is regarded as one positive sample, and the other nine PSMs with the incorrect peptide sequences are regarded as nine negative samples. SVM-rank is then trained on all of the positive and negative samples using the regularization parameter of 1000 based on a linear classifier rather than a kernel classifier, owing to the higher speed of the former.

2.5 Refining the top-1 results by spectrum merging

After reranking the top-10 candidate sequences for each spectrum by the output scores of SVM-rank, different spectra with similar

precursor ion masses within a pre-set tolerance (e.g. ±20 ppm) are further checked to see whether they are generated by the same peptide. In this step, the only top-1 sequence in each spectrum is retained. To avoid merging spectra incorrectly by the random match of similar precursor ions from different peptides, a few additional measures should be involved for evaluating the match quality between the current spectrum and each peptide from other spectra.

For example, for one spectrum s_1, assume that there is another spectrum s_2 with a similar precursor ion mass. The top-1 sequences for s_1 and s_2 are p_1 and p_2, respectively. Then the match quality of s_1 and p_2 is to be tested. To be more specific, for the PSM between s_1 and p_2, if the maximum matched tag length is greater than 3, and the summed intensities of matched peaks account for 5% of the total in s_1, then p_2 should be considered as a peptide candidate for updating the result of s_1. The match quality of s_2 and p_1 needs to be tested in the same way.

Supplementary Figure S1 shows an example to further explain the process of spectrum merging. For spectrum s_1, its top-1 result reranked by SVM-rank is A*D*CEFK with the score of 19. After spectrum merging, another five spectra are found and there are three different sequence candidates in total for the six spectra from s_1 to s_6.

Then, another SVM-rank model is trained, in which two features are considered: one is the number of PSMs supporting each peptide candidate, and the other is the mean value of SVM-scores of the supporting PSMs for each peptide candidate. The two features are based on the fact that a peptide candidate is more confident if it is supported with more PSMs and with a better score. Finally, in this example, the incorrect sequence A*D*CEFK of spectrum s_1 is corrected to the true one A*CD*EFK.

3 Results

3.1 The effect of different features

First, we have investigated the three similarity distributions based on the correct identified PSMs on the seven real datasets (Supplementary Fig. S2). The details about these datasets are shown in Supplementary Table S1. For *V.mungo* dataset, there are as high as 82–87% of results, whose similarities are larger than 0.9, and the median values of cosine, Pearson and Spearman similarities are as high as 0.97, 0.97 and 0.94, respectively (Table 1), suggesting the excellent performance of pDeep that the theoretically predicted spectra are very similar to the real ones.

Then, we also tested the performances of reranking separately using each of the six features and merging all features in one learning-to-rank framework (Supplementary Table S3). If the values of the five features (i.e. cosine, Pearson, Spearman, G_1 and G_2) of two peptide candidates are the same, the original PSM score is used to rerank these two candidates. Although the performances of separately using the last five features except the original score are all inferior to that of the original score, the performance of considering all features in the same learning-to-rank model is significantly better, suggesting the good effect and complementary of the features considered in our model. In order to further investigate the validity of the features, we have compared the distributions between the correct and incorrect results considering each feature (Supplementary Fig. S3). The two distributions shown in each subfigure are with a large K-S distance and the corresponding *P*-value is less than 0.01 based on the two-sample Kolmogorov–Smirnov test, suggesting that the features can effectively discriminate between the correct and incorrect results.

3.2 Performance of different *de novo* sequencing algorithms at the peptide level

pNovo 3 was compared with three other state-of-the-art *de novo* peptide sequencing tools (Supplementary Table S4), specifically PEAKS (Ma *et al.*, 2003) (v8.5), Novor (Ma, 2015) (v1.1) and pNovo+ (Yang *et al.*, 2017) (referred to as pNovo). The seven benchmark datasets described in Section 2.1 were used to measure the result accuracy. A PSM was regarded as correct if its peptide was the same as that annotated by database search in the benchmark datasets (regardless of the difference between Ile and Leu).

The recalls of the top-1 peptide for each spectrum reported by all of the four *de novo* sequencing algorithms were calculated. As shown in Table 2, on *V.mungo* dataset, which was also used for the model training, the recall of pNovo 3 was 64.6%, which was 50.6% higher than that of pNovo (42.9%) and 45.8% higher than that of PEAKS (44.3%). Novor recalled less PSMs, which might be caused by not training with the high-resolution datasets in its test version. On all seven datasets, the recall of pNovo 3 was 35.6–96.1% higher than that of pNovo and 29.4–102.4% higher than that of PEAKS, which demonstrated the good extendibility of the machine learning model. Figure 2 and Supplementary Figure S4 show the consistency of the correct results reported by pNovo 3, pNovo and PEAKS, in which pNovo 3 covered 88.1–94.6% of pNovo results and 82.5–89.6% of PEAKS results. Also, pNovo 3 independently reported 20.6–43.3% more PSMs, which were reported by neither pNovo nor PEAKS.

The recalls considering from top-1 to top-10 peptide candidates for each spectrum are also demonstrated in Figure 3 for the first three datasets and Supplementary Figure S5 for the other four datasets. The recall considering top-k ($1 \leq k \leq 10$) candidates was calculated by the number of the spectra whose correct peptide results were in the top-k sequences divided by the number of total spectra. As Novor only reported the top-1 results, it was not considered in this analysis. As shown in Figure 3 and Supplementary Figure S5, the recall of top-10 results reported by pNovo 3 was ~20.8% higher than that of pNovo and ~25.7% higher than that of PEAKS on all datasets. Although the recall of top-10 results reported by pNovo was slightly higher than that of PEAKS, the recall of top-1 results reported by pNovo was even a little worse than that of PEAKS. This meant that the scoring method in pNovo was less effective to distinguish the similar candidates in one spectrum. However, the refined scoring method in pNovo 3 was shown to be much more powerful. As a result, pNovo 3 yielded a large difference of recall compared with pNovo and PEAKS, especially for the top-1 results which were more important for real biological discoveries.

The recall difference between pNovo 3 and pNovo considering top-10 candidates demonstrated the effect of the spectrum merging method, the last step of pNovo 3, because the original top-10 peptide candidates were the same for both pNovo 3 and pNovo. Furthermore, we compared the performance between pNovo 3, the same algorithm without spectrum merging (referred to as pNovo 3-NM) and the same algorithm without SVM-rank model mentioned in Section 2.4 (referred to as pNovo 3-NR). The recall of top-1 results reported by pNovo 3 was 15.0–35.2% higher than that of pNovo 3-NM and 15.6–35.8% higher than that of pNovo 3-NR (Supplementary Table S5). This demonstrated that both of the two SVM-rank models are useful for increasing the number of correct results. Furthermore, pNovo 3 stably covered ~96% of pNovo 3-NM results on all datasets and independently reported 17.2–30.0% of the total results (Supplementary Fig. S6), which was also proved that this strategy hardly replaced a correct PSM from the learning-to-rank model by an incorrect one.

Table 1. Median values of three similarities on all datasets

	V.mungo	*M.musculus*	*M.mazei*	*S.cerevisiae*	*A.mellifera*	QE_HF_X1	QE_HF_X2
Cosine	0.97	0.97	0.97	0.96	0.97	0.96	0.96
Pearson	0.97	0.97	0.96	0.96	0.96	0.95	0.95
Spearman	0.94	0.94	0.94	0.92	0.94	0.94	0.93
#PSMs[a]	41 721	12 538	67 452	55 163	126 966	104 052	83 313

[a]All spectra whose top-10 peptide candidates contain the correct results are considered. This part accounts for 60–76% of total spectra on all of the seven datasets Vigna mungo (V.mungo), Mus musculus (M.musculus), Methanosarcina mazei (M.mazei), Saccharomyces cerevisiae (S.cerevisiae), Apis mellifera (A.mellifera), QE_HF_X1 and QE_HF_X2.

Table 2. Recall of top-1 peptides identified by different *de novo* sequencing algorithms

	V.mungo	*M.musculus*	*M.mazei*	*S.cerevisiae*	*A.mellifera*	QE_HF_X1	QE_HF_X2
pNovo 3	64.6%	50.4%	66.0%	64.7%	62.5%	47.8%	38.3%
pNovo	42.9%	25.7%	42.4%	47.7%	36.7%	29.8%	21.4%
PEAKS	44.3%	24.9%	42.4%	50.0%	38.0%	32.2%	24.6%
Novor	17.4%	9.7%	19.1%	19.1%	13.7%	10.9%	9.3%
#Total PSMs	62 089	25 354	103 959	81 326	217 841	196 759	201 301

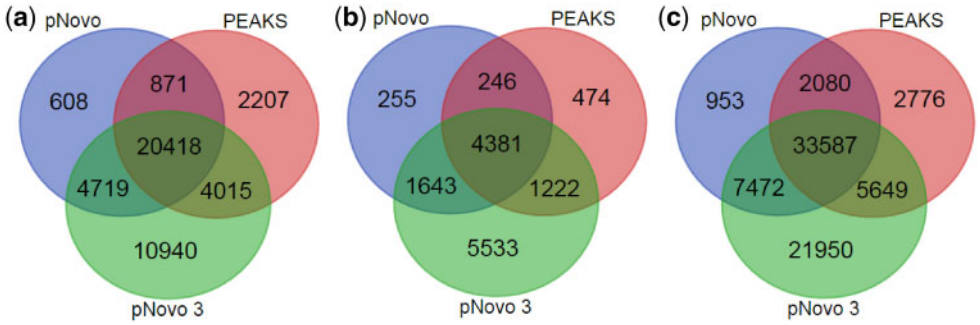

Fig. 2. Venn diagram of the correct results of pNovo 3, pNovo and PEAKS on the first three datasets: (**a**) *V.mungo*, (**b**) *M.musculus* and (**c**) *M.mazei*

3.3 Examples showing the effect of the features used in pNovo 3

Two examples were selected to explain why pNovo 3 could report more correct results than pNovo and PEAKS. The first one is shown in Figure 4. For this spectrum, only pNovo 3 reported the correct sequence (KYDEIDAA*P*EER) annotated by database search while pNovo and PEAKS reported the same incorrect sequence (KYDEIDAA*E*PER). If only considering the quality of the PSM, both two sequences matched the same backbone fragment ions, hence the match scores should be the same if no additional information was considered. However, according to the two theoretical spectra correspondingly predicted by pDeep, the fragmentation patterns of the two sequences were actually quite different, especially for the intensities of y_2, y_3 and y_4 ions (Supplementary Figs S7 and S8), which resulted in the different similarities. In addition, the probabilities of the existence of two gaps, PE and EP, were 0.2 and 0.03, respectively, which is also helpful in distinguishing between the two sequences. Another similar example is shown in Supplementary Figures S9–S11.

3.4 Recalls and precisions of different *de novo* sequencing algorithms at the amino acid level

PEAKS and Novor also reported scores for individual amino acids in the peptide results, which indicate the local confidence level of PSMs and is helpful in assembling entire protein sequences (Tran *et al.*, 2016). The same function of pNovo 3 was implemented by the newly developed software tool, pSite (Yang *et al.*, 2018). Considering the top-1 results reported by each algorithm, the recalls and precisions with different confidence score thresholds were computed (Fig. 5a–c for the first three datasets and Supplementary Fig. S12 for others), and the area under curve (AUC) (Davis, 2006) metric can be used to evaluate the overall accuracy of *de novo* sequencing at the amino acid level. On all datasets, the precision-recall (PR) curves of pNovo 3 were always higher than those of pNovo, PEAKS and Novor. The AUC of pNovo 3 was 12.1–34.4% higher than that of pNovo, 2.0–20.1% higher than that of PEAKS and 65.7–112.5% higher than that of Novor (Fig. 5d).

Supplementary Table S6 shows the total recall and precision of amino acids on the seven datasets regardless of the confidence level, i.e. all amino acids reported by each *de novo* sequencing tool were considered to compute the recall and precision. The recall and precision of pNovo 3 were always greater than 80% on the first five datasets in most cases. On the last two datasets produced by Q Exactive HF-X, the recall and precision decreased to 55–73%, which performed similarly to that for full-length peptides. Overall, the recall of pNovo 3 was 20.5%, 9.2% and 65.9% higher than those of pNovo, PEAKS and Novor, respectively; meanwhile, the precision of pNovo 3 was 18.4%, 17.5% and 83.8% higher than the above three algorithms, respectively.

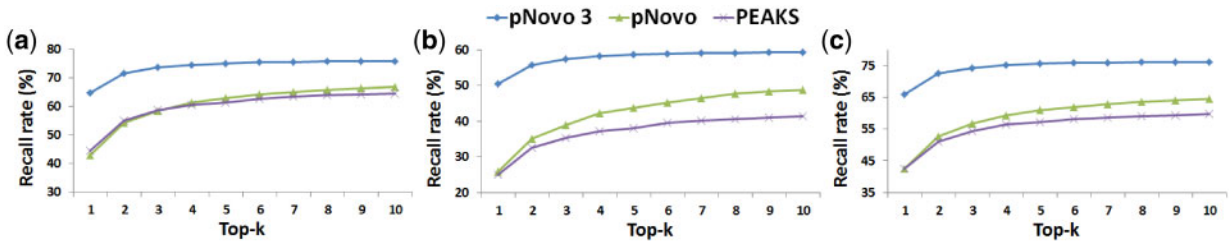

Fig. 3. The recalls of top-1 to top-10 on the first three datasets: (**a**) *V.mungo,* (**b**) *M.musculus* and (**c**) *M.mazei*

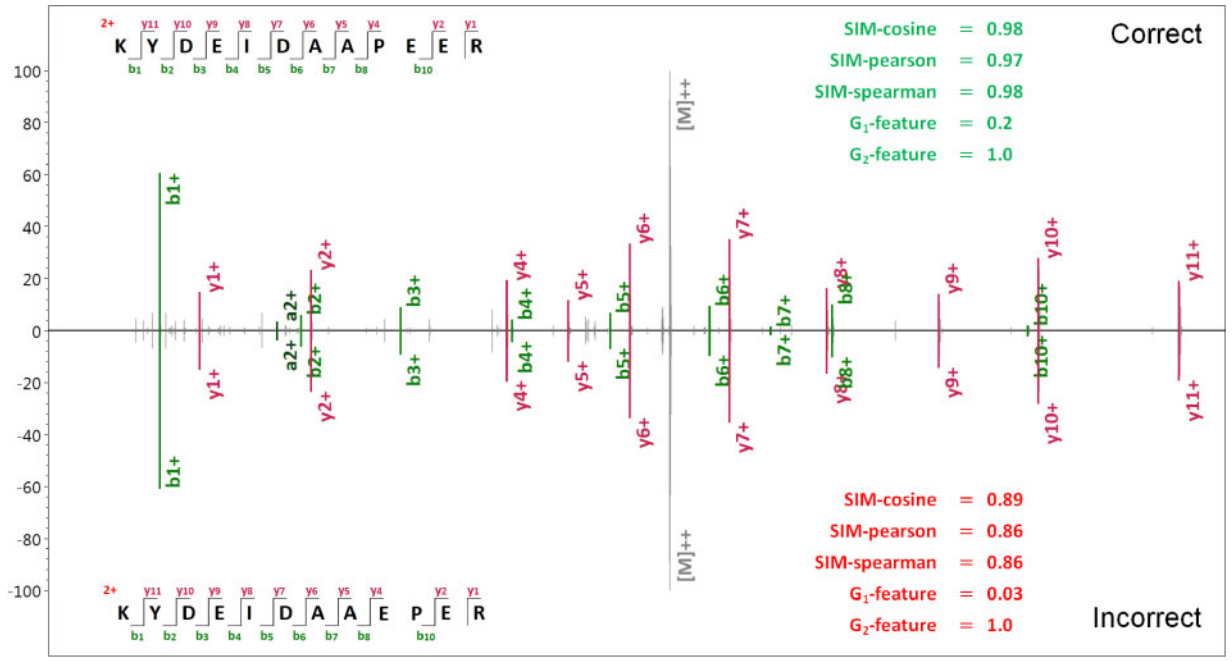

Fig. 4. One example shows that the features extracted by pNovo 3 can effectively discriminate between the correct and very similar incorrect results. The real spectrum is from *V.mungo* dataset and the title of this spectrum is 4723.8552.8552.2.dta. Both the correct (KYDEIDAA*PE*ER, the above subfigure) and incorrect (KYDEIDAA*EP*ER, the below subfigure) peptide sequences are matched to this real spectrum. Five features of the correct and incorrect peptide sequences are labeled with the green and red figures, respectively

3.5 Comparing the performance of pNovo 3 and DeepNovo

Both pNovo 3 and DeepNovo (Tran *et al.*, 2017) have used the deep learning model on the Google TensorFlow library, and the performances of them were also compared in this study. As DeepNovo used different models for the nine high-resolution test datasets (Cassidy *et al.*, 2016; Cypryk *et al.*, 2017; Hu *et al.*, 2016; Mata *et al.*, 2017; Nevo *et al.*, 2017; Paiva *et al.*, 2016; Petersen *et al.*, 2016; Reuß *et al.*, 2017; Seidel *et al.*, 2017), we downloaded the original results of DeepNovo rather than re-analysis the datasets using the unified model in the software, and then tested pNovo 3 with the same nine datasets to make a fair comparison. The benchmark strategy was the same as that shown in Section 2.1 that PSMs generated by the database search results of PEAKS DB at 1% FDR were used as the ground truth, while the PSMs whose corresponding spectra were not appeared in the results of DeepNovo were removed.

As shown in Supplementary Table S7, the recall of pNovo 3 was still 20.6–49.8% higher than that of DeepNovo. Furthermore, pNovo 3 also yields higher recall and precision at the amino acid level for the top-1 peptide sequences (Supplementary Table S8). This gap might be owing to the different ways of using deep learning between these two algorithms. DeepNovo combined deep learning and dynamic programming in a unified *de novo* sequencing workflow, while pNovo 3 divided this workflow into two steps: finding top-ranked candidates by the traditional algorithm, e.g. pNovo+, and then reranking candidates considering several different features extracted by deep learning, which was integrated into a learning-to-rank framework. The first step has been widely investigated in past decades, which might be more mature compared with the newly proposed deep learning approach. However, once the top-ranked peptide candidates were generated, the deep learning approach, which provided more accurate spectrum prediction for the following learning-to-rank model, played a more important role in distinguishing among the similar peptides in pNovo 3.

4 Discussion

In this study, we have used the deep learning approach to extract features, and built a learning-to-rank model to rerank the results of *de novo* sequencing. Until now, the problem of low precision on *de novo* sequencing has not been solved well because there are no effective methods to distinguish similar peptides if no pivotal peaks in

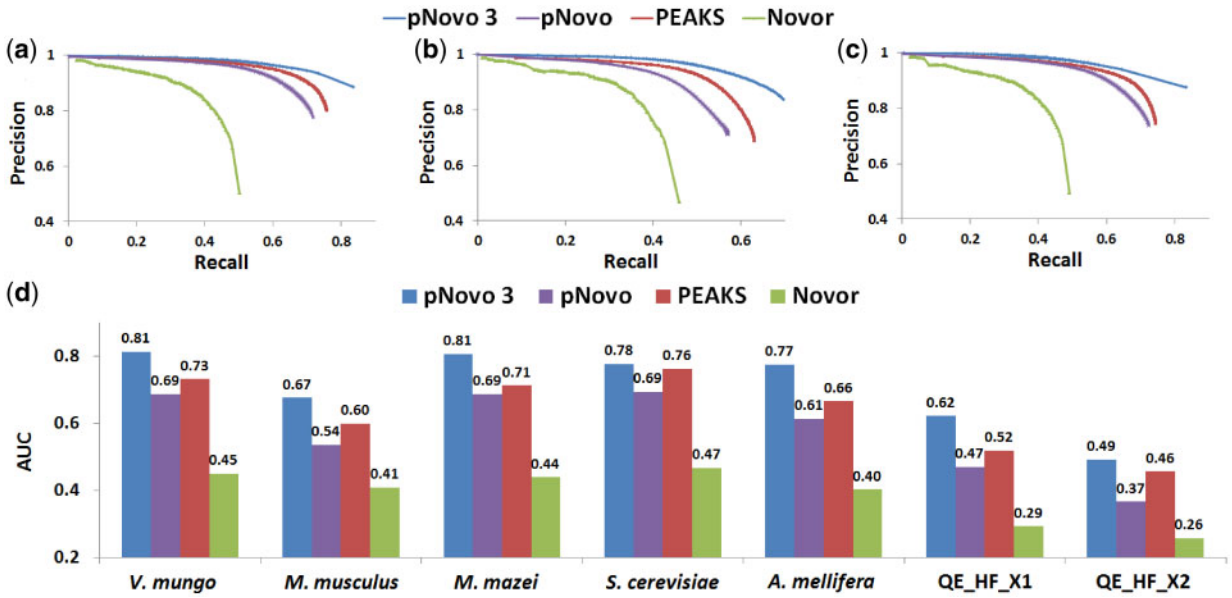

Fig. 5. The precision-recall (PR) curves of pNovo 3, pNovo, PEAKS and Novor on the first three datasets: (**a**) *V.mungo*, (**b**) *M.musculus*, (**c**) *M.mazei*. (**d**) The AUCs of the three algorithms on the seven datasets

one spectrum can be detected. As a result, a more powerful scoring function is needed, and using deep learning to learn the fragmentation pattern of peptides comes into the view of this study. However, deep learning models are often learned directly from raw data and do not rely on well-designed features, and we still need to find that which features are most useful for *de novo* sequencing to discriminate between correct and incorrect peptides. In this study, we found that the similarity between the experimental and theoretical spectra, which were measured by three types of metrics, was very important for reranking *de novo* sequencing results. As no model comprehensively considering modified peptides was yet trained by the current version of pDeep, only peptides without variable modifications were used in this study; however, the learning-to-rank model can be easily extended to modified peptides with the upgrading of pDeep.

We used learning-to-rank models [e.g. SVM-rank (Joachims, 2002; Joachims *et al.*, 2009) and RankBoost (Freund *et al.*, 2004)], rather than traditional machine learning models [e.g. SVM (Cortes and Vapnik, 1995; Vapnik, 1999) and decision tree (Quinlan, 1999)], in this study. In general, traditional machine learning models are more suitable to learn a global classification function to effectively discriminate between correct and incorrect PSMs from different spectra; however, as the comparison among different spectra is less important in *de novo* sequencing, learning-to-rank models are more applicable to solve the reranking problem, i.e. rerank the similar sequence candidates for each spectrum.

On all datasets, the recalls and precisions of pNovo 3 are always the highest compared with pNovo, PEAKS, Novor and DeepNovo. But the recalls of top-10 results from pNovo 3 are still only 60–76% on different datasets so that we are curious about the reason why the rest of the results cannot be sequenced even when as many as ten candidates are considered. For example, a total of 15 067 (24% of 62 089) PSMs in the *V.mungo* dataset are not recalled in the top-10 results, and 32% (4775/15 067) of which are difficult to be recalled by *de novo* sequencing because the maximum gap lengths are greater than 2. We further try to enumerate the similar peptide candidates based on the correct sequence from these low-quality PSMs, and then match them to the original spectra. For example, if the correct sequence is *A*SQEPK with a gapped subsequence ASQ, the similar candidates involve ASQEPK, AQSEPK, …, SQAEPK, and then the three similarity metrics used in this study are computed (Supplementary Fig. S13). We find that their similarities are too close to find which one is correct. This means that the *de novo* sequencing algorithms at the current stage may not be able to distinguish among the similar results with long gapped subsequences, even using the effective deep learning approach. In this case, the more effective way to improve the accuracy of *de novo* sequencing is to produce high-quality MS/MS spectra with higher fragment ion coverage.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFA0501300 to S.-M.H), the National Natural Science Foundation of China (31470805), the Youth Innovation Promotion Association CAS (2014091), the National High Technology Research and Development Program of China (863) (2014AA020902 to S.-M.H. and 2014AA020901 to H.C.).

Conflict of Interest: none declared.

References

Bartell,B. *et al.* (1995) Learning to retrieve information. In: *Current Trends in Connectionism*, Proceedings of the Swedish Conference on Connectionism, pp. 345–353.

Bogdanoff,W.A. *et al.* (2016) *De novo* sequencing and resurrection of a human astrovirus-neutralizing antibody. *ACS Infect. Dis.*, **2**, 313–321.

Cassidy,L. *et al.* (2016) Combination of bottom-up 2D-LC-MS and semi-top-down GelFree-LC-MS enhances coverage of proteome and low molecular weight short open reading frame encoded peptides of the archaeon *Methanosarcina mazei*. *J. Proteome Res.*, **15**, 3773–3783.

Chi,H. *et al.* (2013) pNovo+: *de novo* peptide sequencing using complementary HCD and ETD tandem mass spectra. *J. Proteome Res.*, **12**, 615–625.

Chi,H. *et al.* (2015) pFind-Alioth: a novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *J. Proteomics*, **125**, 89–97.

Chi,H. *et al.* (2018) Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine. *Nat. Biotechnol*, **36**, 1059–1061.

Chi,H. *et al*. (2010) pNovo: *de novo* peptide sequencing and identification using HCD spectra. *J. Proteome Res*., **9**, 2713–2724.

Chick,J.M. *et al*. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol*., **33**, 743–749.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn*., **20**, 273–297.

Cox,J. and Mann,M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol*., **26**, 1367–1372.

Cypryk,W. *et al*. (2017) Proteomic and bioinformatic characterization of extracellular vesicles released from human macrophages upon influenza A virus infection. *J. Proteome Res*., **16**, 217–227.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In: Cohen,W. and Moore,A. (eds.) *Proceedings of the 23rd International Conference on Machine Learning*. ACM, New York, pp. 233–240.

Eng,J.K. *et al*. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom*., **5**, 976–989.

Frank,A. and Pevzner,P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem*., **77**, 964–973.

Freund,Y. *et al*. (2004) An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res*., **4**, 933–969.

Fu,Q. and Li,L.J. (2005) *De novo* sequencing of neuropeptides using reductive isotopic methylation and investigation of ESI QTOF MS/MS fragmentation pattern of neuropeptides with N-terminal dimethylation. *Anal. Chem*., **77**, 7783–7795.

Graves,A. *et al*. (2013) Hybrid speech recognition with deep bidirectional LSTM. In: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* in Olomouc, Czech Republic, pp. 273–278.

Guthals,A. *et al*. (2017) *De novo* MS/MS sequencing of native human antibodies. *J. Proteome Res*., **16**, 45–54.

Hettich,R.L. *et al*. (2013) Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal. Chem*., **85**, 4203–4214.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput*., **9**, 1735–1780.

Hu,H. *et al*. (2016) Proteome analysis of the hemolymph, mushroom body, and antenna provides novel insight into honeybee resistance against varroa infestation. *J. Proteome Res*., **15**, 2841–2854.

Joachims,T. (2002) Optimizing search engines using clickthrough data. In: *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD) in Edmonton, Alberta, Canada*, ACM.

Joachims,T. *et al*. (2009) Cutting-plane training of structural SVMs. *Mach. Learn*., **77**, 27–59.

Kelstrup,C.D. *et al*. (2018) Performance evaluation of the Q Exactive HF-X for shotgun proteomics. *J. Proteome Res*., **17**, 727–738.

Kulak,N.A. *et al*. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods*, **11**, 319–324.

Liu,T.Y. (2010) Learning to rank for information retrieval. In: *Sigir 2010: Proceedings of the 33rd Annual International ACM Sigir Conference on Research Development in Information Retrieval in Geneva, Switzerland*, pp. 904–904.

Ma,B. (2015) Novor: real-time peptide *de novo* sequencing software. *J. Am. Soc. Mass Spectrom*., **26**, 1885–1894.

Ma,B. *et al*. (2003) PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*., **17**, 2337–2342.

Mata,C.I. *et al*. (2017) In-depth characterization of the tomato fruit pericarp proteome. *Proteomics*, **17**, 1–2.

Muth,T. and Renard,B.Y. (2017) Evaluating *de novo* sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Brief. Bioinform*.

Nevo,N. *et al*. (2017) Impact of cystinosin glycosylation on protein stability by differential dynamic stable isotope labeling by amino acids in cell culture (SILAC). *Mol. Cell. Proteomics*, **16**, 457–468.

Paiva,A.L.S. *et al*. (2016) Label-free proteomic reveals that cowpea severe mosaic virus transiently suppresses the host leaf protein accumulation during the compatible interaction with cowpea (*Vigna unguiculata* [L.] Walp.). *J. Proteome Res*., **15**, 4208–4220.

Perkins,D.N. *et al*. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Petersen,J.M. *et al*. (2016) Chemosynthetic symbionts of marine invertebrate animals are capable of nitrogen fixation. *Nat. Microbiol*., **2**, 16195.

Quinlan,J.R. (1999) Simplifying decision trees. *Int. J. Hum.-Comput. St*., **51**, 497–510.

Reichert,J.M. *et al*. (2005) Monoclonal antibody successes in the clinic. *Nat. Biotechnol*., **23**, 1073–1078.

Reuß,D.R. *et al*. (2017) Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism. *Genome Res*., **27**, 289–299.

Seidel,G. *et al*. (2017) Quantitative global proteomics of yeast PBP1 deletion mutants and their stress responses identifies glucose metabolism, mitochondrial, and stress granule changes. *J. Proteome Res*., **16**, 504–515.

Sharma,K. *et al*. (2015) Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci*., **18**, 1819–1831.

Silver,D. *et al*. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature*, **529**, 484.

Snyder,A.P. (2000) *Interpreting Protein Mass Spectra, A Comprehensive Resource*, pp. 265–299, Oxford University Press, Oxford.

Tran,N.H. *et al*. (2016) Complete *de novo* assembly of monoclonal antibody sequences. *Sci. Rep*., **6**, 31730.

Tran,N.H. *et al*. (2017) *De novo* peptide sequencing by deep learning. In: *Proceedings of the National Academy of Sciences of the United States of America*.

Vapnik,V.N. (1999) An overview of statistical learning theory. *IEEE Trans. Neural Netw*., **10**, 988–999.

Yang,H. *et al*. (2017) Open-pNovo: *de Novo* peptide sequencing with thousands of protein modifications. *J. Proteome Res*., **16**, 645–654.

Yang,H. *et al*. (2018) pSite: amino acid confidence evaluation for quality control of *de novo* peptide sequencing and modification site localization. *J. Proteome Res*., **17**, 119–128.

Zhang,J. *et al*. (2012) PEAKS DB: *de novo* sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell. Proteomics*, **11**, M111.010587.

Zhou,X.X. *et al*. (2017) pDeep: predicting MS/MS spectra of peptides with deep learning. *Anal. Chem*., **89**, 12690–12697.

Zhu,Y.F. *et al*. (2014) SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol. Cell. Proteomics*, **13**, 1552–1562.