SIGBio Record

Newsletter of the SIGBio
ACM Special Interest Group

Volume 8, Issue 1, May 2018 ISSN 2159-1210

Notice to Contributing Authors to SIG Newsletters

SIGBIO Record - Submission Guidelines

Submission categories

Submissions to the newsletter can be either on a special issue topic or on topics of general interest to the SIGBIO community.

These can be in any one of the following categories:

- Survey/tutorial articles (short) on important topics.
- Topical articles on problems and challenges
- Well-articulated position papers.
- Review articles of technical books, products and .
- Reviews/summaries from conferences, panels and special meetings within 1 to 4 pages [1500-2500 words]
- Book reviews and reports on relevant published technical books
- PhD dissertation abstracts not exceeding 10 pages
- Calls and announcements for conferences and journals not exceeding 1 page
- News items on the order of 1-3 paragraphs

Brief announcements Announcements not exceeding 5 lines in length can simply be sent as ASCII text to the
editors by e-mail. SIGBIO Record publishes announcements that are submitted as is without review.

Announcements cannot be advertisements and should be of general interest to the wider community. The Editor reserves the right to reject any requests for announcements at his discretion.

Authors are invited to submit original research papers or review papers in all areas of bioinformatics and computational biology. The papers published in SIGBioinformatics Record will be archived in ACM Digital Library. Papers should follow the ACM format, and there is no page limitation.

http://www.acm.org/sigs/publications/proceedings-templates

Submissions should be made via email to one of the two editors Pierangelo Veltri (veltri@unicz.it) *(University Magna Graecia of Catanzaro, Italy)- Pietro Hiram Guzzi (hguzzi@unicz.it),or to ACM SIGBio email acmsigbiorecord@gmail.com*

Index

Seq3seq Fingerprint: Towards End-to-end Semi-supervised Deep Drug Discovery

Xiaoyu Zhang
The University of Texas at Arlington
Arlington, Texas
xiaoyu.zhang2@mavs.uta.edu

Sheng Wang
The University of Texas at Arlington
Arlington, Texas
sheng.wang@mavs.uta.edu

Feiyun Zhu
The University of Texas at Arlington
Arlington, Texas
feiyun.zhu@uta.edu

Zheng Xu
The University of Texas at Arlington
Arlington, Texas
zheng.xu@mavs.uta.edu

Yuhong Wang
National Center for Advancing
Translating Sciences, NIH
Rockville, Maryland
yuhong.wang@nih.gov

Junzhou Huang*
The University of Texas at Arlington
Tencent AI Lab
jzhuang@uta.edu

ABSTRACT

Observing the recent progress in Deep Learning, the employment of AI is surging to accelerate drug discovery and cut R&D costs in the last few years. However, the success of deep learning is attributed to large-scale clean high-quality labeled data, which is generally unavailable in drug discovery practices.

In this paper, we address this issue by proposing an end-to-end deep learning framework in a semi-supervised learning fashion. That is said, the proposed deep learning approach can utilize both labeled and unlabeled data. While labeled data is of very limited availability, the amount of available unlabeled data is generally huge. The proposed framework, named as **seq3seq fingerprint**, automatically learns a strong representation of each molecule in an unsupervised way from a huge training data pool containing a mixture of both unlabeled and labeled molecules. In the meantime, the representation is also adjusted to further help predictive tasks, e.g., acidity, alkalinity or solubility classification. The entire framework is trained end-to-end and simultaneously learn the representation and inference results. Extensive experiments support the superiority of the proposed framework.

CCS CONCEPTS

• **Theory of computation → Unsupervised learning and clustering**; **Structured prediction**; • **Applied computing → Molecular sequence analysis**; **Sequencing and genotyping technologies**; *Bioinformatics*; Imaging;

KEYWORDS

Semi-Supervised Learning; Unsupervised Learning; Structured Prediction; Learning Representation; Sequence to Sequence Learning; Deep Learning; Drug Discovery; Virtual Screening; Molecular Representation; Imaging; Computational Biology

ACM Reference Format:
Xiaoyu Zhang, Sheng Wang, Feiyun Zhu, Zheng Xu, Yuhong Wang, and Junzhou Huang. 2018. Seq3seq Fingerprint: Towards End-to-end Semi-supervised Deep Drug Discovery. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233548

1 INTRODUCTION

In the past few years, the application of Artificial Intelligence (AI) technologies in drug discovery has become significant and increasingly popular. Observing the most recent rapid growth of a key technology in AI, namely **deep learning** (or **deep neural network**), the whole industry and academia are looking towards AI to speed up the drug discovery, cut R&D cost and decrease the failure rate in potential drug screening trials [6].

However, the previous success of deep learning in multiple applications, e.g., image understanding [8, 34], medical imaging [16, 23, 36, 40], video understanding [2, 46], bioinformatics [43–45], and machine translation [19], etc., has implied a reliance on large-scale high-quality labeled data-sets. The training procedure of those deep-learning-based state-of-the-art models generally involve millions of labeled samples. In the meantime, however, for the drug discovery tasks, the scale of labeled data-set stays around only thousands of examples due to the insanely high cost of obtaining the clean labeled data through the biological experiments. The available amount of the labeled training data is absolutely insufficient to secure the success of the application of deep learning in the drug discovery [27]. This huge gap between the requirement and availability of the labeled data in drug discovery has become a bottleneck of applying deep learning techniques into drug discovery.

Given the high cost of obtaining sufficient labeled data points, it seems impractical to increase the labeled data-set scale to a satisfactory level. To address this issue, we propose a semi-supervised deep learning modeling strategy. In simple terms, the proposed deep

*This work was partially supported by U.S. NSF IIS-1423056, CMMI-1434401, CNS-1405985, IIS-1718853, and NSF CAREER grant IIS-1553687.

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233548

learning framework can learn from both labeled and unlabeled data, while the unlabeled data is almost infinitely available. For instance, the ZINC data-set [17] is publicly available and contains over 35 million unlabeled molecule data. With such scale of data being used, the deep learning model is expected to be trained with enough representation power to help the inference task.

In this paper, we propose a semi-supervised data-driven multi-task deep-learning-based drug discovery method, named as **seq3seq fingerprint**. The reasons behind this naming are two-fold: 1) this is the **next-generation seq2seq fingerprint** [43], whose major upgrade is that the original two-stage pipeline has been combined into an multi-task one-stage end-to-end pipeline to ensure much more decent inference performance; 2) the seq3seq fingerprint framework contains **three** ends with one input and two outputs while the seq2seq fingerprint contains **two** ends with one input and one output.

To briefly introduce the proposed seq3seq fingerprint framework, the seq3seq fingerprint network can be considered as a pipeline with one input and two outputs. The designed neural network can take the molecule inputs for training, **with or without labels**. The input is the raw sequence representation of a molecule, namely SMILE representation. Examples are referred in Figure 1. The two outputs will correspond to the two tasks inside this network. The first one is the **self-recovery**. The network is expected to be able to generate a vector representation which is able to be recovered back to original raw sequence representation. The second task is the **inference** whenever the label is available. For instance, it can be a task to predict the acidity, alkalinity or solubility of a single molecule. The two tasks are trained within the same network in an end-to-end fashion. As a result, in a specific inference task, the vector representation will be able to provide both good recovery performance and inference performance. Also, the network can be trained inside a mixture data pool with both labeled and unlabeled data, which is sufficient enough to ensure the fine training of the neural network.

The benefits of the seq3seq fingerprint are three folds: 1) the training phase of seq3seq fingerprint takes both labeled and unlabeled data into consideration, which is able to provide both strong vector representation and good inference performance. 2) it is data-driven, eliminating the reliance on expert's subjective knowledge. 3) since the unlabeled data is almost unlimited in practice, it will significantly complement the sole training with labeled data, ensuring a final good inference performance.

The technical contributions of this paper are summarized as: 1) the seq3seq fingerprint method is obviously the first attempt to utilize both labeled data and unlabeled data for sequence-based end-to-end deep learning in drug discovery. 2) several important features are enabled in the seq3seq fingerprint to help inference:

- this is the first **end-to-end** framework coupling both the recovery and inference task.
- the proposed framework is general enough to suit **different prediction tasks**, e.g., classification, regression, etc.
- it is feasible to use **different inference network structures**, e.g., Convolutional Neural Networks (CNNs), Multi-Layer Perceptrons (MLPs), etc.

3) extensive experiments demonstrate the superior performance on different tasks over both supervised and unsupervised state-of-the-art fingerprint methods.

The rest of the paper is organized as follows. We summarize several related work in drug discovery, in Section 2. In Section 3, we describe our entire pipeline in details. We show our experiment results in Section 4, demonstrating the superior performance of our method. We conclude and discuss the future direction of our paper in Section 5.

2 RELATED WORK

In this section, we briefly introduce several related works. First, we present the raw representation of molecules, namely SMILE representation, i.e., the persistence form of the molecular data in the cold data storage. Second, we list a few state-of-the-art fingerprint methods, including the ones using human-designed and hash-based features.. Finally, we briefly describe some most recent deep learning based methods, e.g., neural fingerprint [5], seq2seq fingerprint [43].

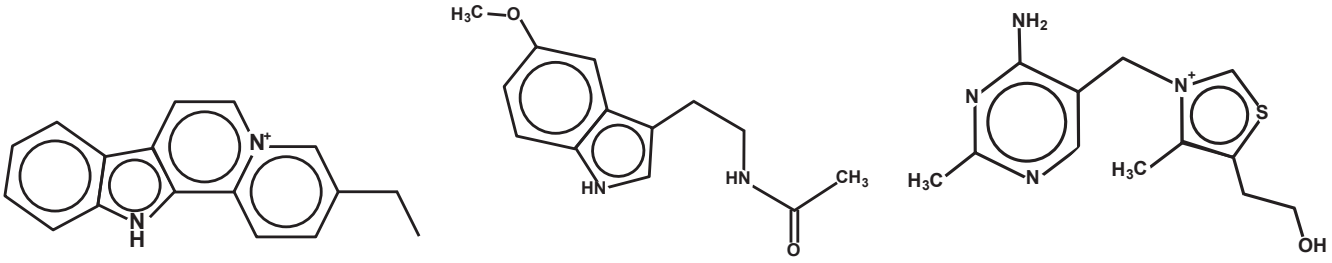
2.1 SMILE Representations of Molecules

Initially, the molecules are stored in the form of a sequence representation, namely the Simplified Molecular-Input Line-Entry system (SMILE) [37], which is a line notation for describing the structure of chemical species using text strings. The SMILE system represents the chemical structures in a graph-based definition, where the atoms, bonds and rings are encoded in a graph and represented in text sequences. Simple examples of SMILE representations are 1) dinitrogen with structure $N \equiv N$ (N#N), 2) methyl isocyanate with structure $CH_3 - N = C = O$ (CN=C=O), where corresponding SMILE representations are included in the brackets. Simply speaking, the letters, e.g., C, N, generally represent the atoms, while some symbols like $-, =, \#$ represent the bonds. We show some more complicated examples in Figure 1.

2.2 Fingerprint Methods

Hash-based Fingerprints. Many hash-based methods has been developed to generate unique molecular feature representation [12, 15, 24]. One important class is called **circular fingerprints**. Circular fingerprints generate each layer's features by applying a fixed hash function to the concatenated features of the neighborhood in the previous layer. One of the most famous ones is Extended-Connectivity FingerPrint (ECFP) [29]. However, due to the non-invertible nature of the hash function, the hash-bashed fingerprint methods usually do not encode enough information and hence result in lower performance in the further predictive tasks.

Biologist-guided Local-Feature Fingerprints. Another mainstream of traditional fingerprint methods is designed based on the biological experiments and the expertise knowledge and experience, e.g., [26, 30]. Biologists look for several important task-related sub-structures (fragments), e.g., $CC(OH)CC$ for pro-solubility prediction, and count those sub-structures as local features to produce fingerprints. This kind of fingerprint methods usually work well for specific tasks, but poorly generalize for other tasks.



(a) Flavopereirin: CCc(c1)ccc2[n+]1ccc3c2Nc4c3cccc4 (b) Melatonin: CC(=O)NCCC1=CNc2c1cc(OC)cc2 (c) Thiamine: OCCc1c(C)[n+](cs1)Cc2cnc(C)nc2N

Figure 1: The examples of SMILE representations.

2.3 Deep-learning-based Models

The growth of deep learning [20, 39] has provided the great flexibility and performance to create the molecular fingerprint from data samples, without explicit human guide, [3, 9, 18, 31, 35, 43]. In this subsection, we discuss two major classes, namely supervised and unsupervised learning models.

Supervised Models. Many of deep learning-based fingerprint methods are still trained in a supervised-learning fashion [31, 38], which is using only labeled molecular data samples as inputs and adjusting model weights according to their labels [21]. However, as mentioned earlier, the performance of the deep supervised learning models are generally limited by the availability of the labeled data. The state-of-the-art work is the neural fingerprint [9]. The neural fingerprint mimics the process of generating circular fingerprint but instead the hash function is replaced by a non-linear activated densely connected layer. This method is based on the deep graph convolutional neural network [13, 21, 22, 25]. There are also few attempts that address the insufficient label issue by using few-shot learning strategies, e.g., [4]. To secure a satisfactory performance and acquire enough labeled data, biologists need to perform a sufficiently large number of tests on chemical molecules, which is prohibitively expensive.

Unsupervised Models. Recently, few unsupervised fingerprint methods, e.g., seq2seq fingerprint [43], are proposed to alleviate the issue brought by the insufficient labeled data. These models generally train deep neural networks to provide strong vector representations using a big pool of unlabeled data. The vector representation model is thereafter used for supervised training with other models, e.g., Adaboost [10], GradientBoost [11], and RandomForest [14], etc. Since the deep models are trained with a sufficiently large data-set, the representation is expected to contain enough information to provide good inference performance. However, this type of methods are not trained end-to-end, meaning that the representation only adjusts to the recovery task of the original raw representation. It is robust to the specific labeled task, but might not provide optimal inference performance for each task.

3 METHODOLOGY

In this section, we describe the details of our semi-supervised seq3seq fingerprint model. First, an overview of the proposed seq3seq fingerprint model is given. The proposed semi-supervised model is trained in an end-to-end fashion by completing two tasks, a self-recovery task for molecule (without any label) and an inference task (with specific classification/regression label). After that, we describe the recovery task and the inference task in detail, their loss functions and how the two tasks are trained. Then the semi-supervised loss is described. In the end, we offer a multi-task scaffolding view from frame-semantic parsing [33] in natural language processing area to explain the proposed model.

3.1 Overview

Different from traditional models [5, 43], the proposed seq3seq fingerprint model works in a semi-supervised fashion. It means that our training data comes from two sources, the labeled data, for classification/regression, as well as the unlabeled data. The labeled data contains the SMILE strings for molecule data and their labels, such as acidity or other molecular activities. The unlabeled data contains just molecular SMILE strings and the unlabeled data is almost infinitely available. The proposed seq3seq fingerprint model takes the mixture of the labeled data and unlabeled data together as training inputs to the network. The work flow is depicted in Figure 2. The semi-supervised training is done by two tasks: the self-recovery task and the inference task. The whole pipeline is illustrated in Figure 3.

3.2 The Duo Tasks in Seq3seq Fingerprint Model

The Self-recovery Task The self-recovery task is to learn a vector representation (usually noted as **fingerprint** in the drug discovery literature) for each input molecular SMILE string. This task also requires the SMILE string of the molecule can be recovered from its fingerprint vector. It is an unsupervised learning problem since no label information is required in training. As shown in Figure 3, this task contains a perceiver network and an interpreter network. This structure is motivated by the seq2seq model [32, 43]. The original seq2seq model is used in machine translation [32]. It is to learn a vector representation from a sentence in a given language, e.g., English, then translate the learned representation into another language such as French. Seq2seq fingerprint [43] combines the idea from seq2seq learning and the idea of auto-encoder to learn the vector representation for molecule.

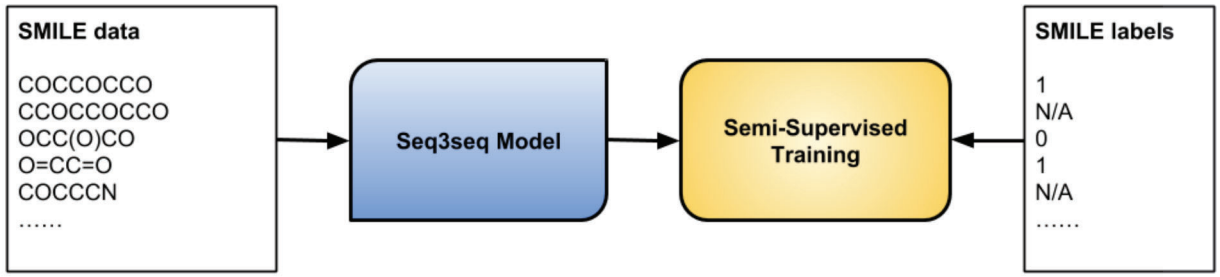


Figure 2: This figures shows how semi-supervised training is used for our proposed model. We mix the unlabeled data and labeled data together to train our proposed model. The SMILEs with label 0/1 come from labeled dataset and the SMILEs without labels (N/A in the figure) come from unlabeled dataset.

We generalize the idea of seq2seq [5, 43] in two views. First, the perceiver network and the interpreter network in the proposed seq3seq fingerprint model can be any recurrent deep neural networks such as LSTM, GRU neural networks. The only limitation is that the perceiver network could map the string tokens into a vector representation and the interpreter could map the vector back into string tokens. Second, we introduce unlabeled molecule data into our training process to learn better representations. Instead of using the SMILE strings of only the labeled molecule data, we take advantage of the **almost infinite** unlabeled data and use both unlabeled and labeled data for the self-recovery task to learn a more accurate vector presentation than those models which only use labeled data or unlabeled data separately. The loss function in our proposed model follows the one in [43]. It is the sum of multiple cross-entropy loss and we denote it as $\mathcal{L}_{unsup}$.

The Inference Task The inference task in the proposed seq3seq fingerprint model is to predict the activity of molecules. In the proposed model, the inference task includes the perceiver network and the inference network. The perceiver network is shared in both self-recovery and inference tasks. It is trained by both labeled and unlabeled data in an end-to-end fashion. The inference network maps the seq3seq fingerprint to a final inference result on a certain prediction task. The structure of the inference network can be any trainable network which maps the vector into a inference value. It allows huge flexibility for the choice of the inference network. For instance, it could be a Convolutional Neural Network (CNN), a Multi-Layer Perceptron (MLP) or even a single fully-connected layer. Depending on whether the inference task is classification or regression, the loss for the inference task $\mathcal{L}_{sup}$ could be either classification loss (usually a cross entropy loss) or regression loss (usually a ℓ_1 smooth/ℓ_2 distance loss). Since computing the $\mathcal{L}_{sup}$ needs labels, the inference task is only trained on labeled data.

3.3 End-to-end Semi-supervised Learning

As shown in Figure 3, the semi-supervised loss $\mathcal{L}_{semi}$ combines the unsupervised loss $\mathcal{L}_{unsup}$ and the supervised loss $\mathcal{L}_{sup}$ together as

$$\mathcal{L}_{semi} = \mathcal{L}_{unsup} + \lambda \mathcal{L}_{sup}. \tag{1}$$

where λ is a hyper-parameter of the proposed model to balance the two tasks. The proposed model is trained with both supervised data and unsupervised data. When the data is unlabeled, the supervised loss $\mathcal{L}_{sup}$ will be zero. Thus, in this case, only the part of the model in self-recovery task will be trained. While the data is labeled, both the part of the model in self-recovery and inference will be trained. The end-to-end training avoids the multi-stage training, i.e., pre-trained model training or separated classifier training [43]. As a result, the proposed end-to-end model is expected to provide an optimal inference performance as well as shorter training time for specific task than that in a multi-stage model from [43].

3.4 A Multi-task Scaffolding View of Seq3seq Fingerprint

In [43], the authors viewed seq2seq fingerprint as a machine translation problem in the Natural Language Processing (NLP) area, with both source and target language set to be the SMILE representation. Interestingly, the proposed seq3seq fingerprint model can be viewed, to some extent, as **a multi-task scaffolding framework** [33] in the NLP area as well. In [33], the authors focus on solving the frame-semantic parsing problem, which is basically finding the *action* (frame) with its associated objects from a sentence. For example, in sentence "Alice loves Bob.", the frame is "loves" with its associated objects being "Alice" and "Bob". However, a single sequence-to-frame network model generally performs poorly in this task. In [33], they proposed to use a multi-task framework to refine the predictions. Besides the frame parsing task, they also introduce the syntactic parsing task. The second task is basically predicting the word categories, e.g., nouns, adverbs, adjectives, etc. For the previous "Alice loves Bob." sentence, the result will be that "Alice" being noun, "loves" being verb and "Bob" being another noun. In [33], it is demonstrated that the second task significantly helps the success of the main (frame parsing) task. To sum up, the multi-task scaffolding frame parsing framework utilizes a second *syntactic parsing* task to reinforce the main task which is the *frame parsing*. Our seq3seq fingerprint can be viewed in a very similar fashion: the **self-recovery task** serves as the auxiliary task to augment the main **prediction task**. This modification is also further demonstrated superior in our experiments described in Section 4.

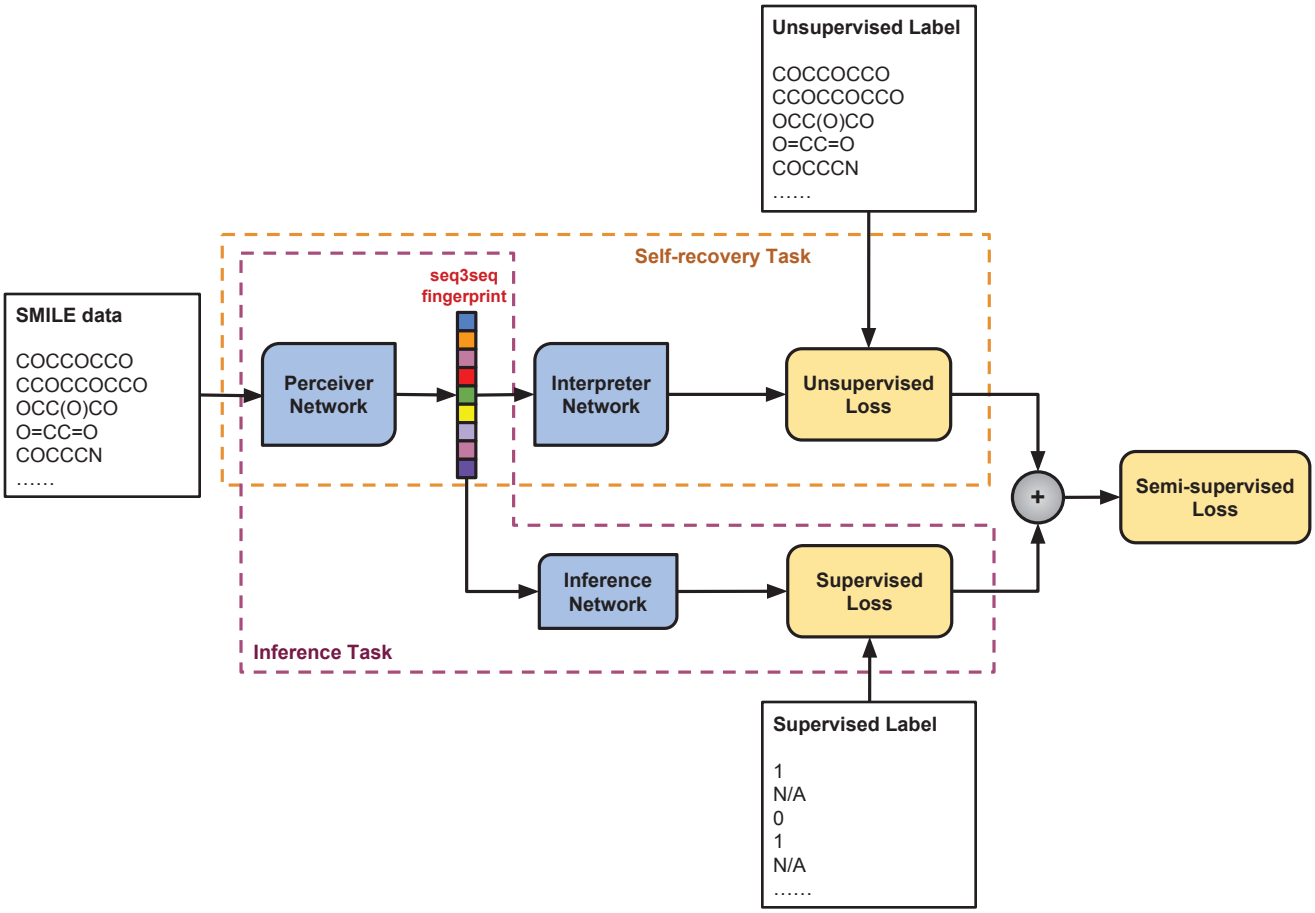


Figure 3: This figure shows the proposed seq3seq fingerprint model. The proposed model is trained through two tasks: a self-recovery task and an inference task. The self-recovery task contains a perceiver network and an interpreter network; the inference task shares the perceiver with self-recover task and has an inference network. The semi-supervised loss is the sum of supervised loss and unsupervised loss.

4 EXPERIMENTS

In this section, we first detail the experimental setup, e.g., the data set description, hardware and software settings, etc. Then we report the benchmark performance of the seq3seq fingerprint methods among state-of-the-art methods. Furthermore, to show the flexibility of our methods and complete our experiments, we offer ablation studies for the sensitivity of the hyper-parameters of our seq3seq fingerprint models, e.g., the multi-task balance weight λ, the Recurrent Neural Network (RNN) layer hidden size and layer number, etc.

4.1 Experiment Setup

Datasets As we mentioned in the introduction, the seq3seq fingerprint can be trained from a mixture of both unlabeled and labeled data. In practices, we usually use an unlabeled data set of a much larger size than that of a labeled dataset.

Unlabeled Dataset For (large) unlabeled dataset, we use ZINC drug-like datasets [17]. ZINC is a free database of commercially-available compounds for virtual screening. The drug-like dataset from ZINC contains 18,691,354 molecular SMILE representations.
Labeled Dataset Two additional datasets, LogP and PM2-10k, were used for semi-supervised training and test. They are obtained from National Center for Advancing Translational Sciences (NCATS) at National Institutes of Health (NIH). Each of them contains around 10,000 molecular SMILE representations with multiple scores, each score quantifies some chemical property. Classification was conducted on LogP and PM2-10k.

- **LogP**: Totally 10,850 samples were used from LogP, Each sample contains a pair of a SMILE string and a water-octanol partition coefficient (LogP) value. A threshold of 1.88 is used to label the data. For those samples with LogP value smaller than 1.88 were classified as negative samples, the rest were labeled as positive samples.

Table 1: The comparison of classification accuracy on the LogP data. We report the average classification accuracy (Mean) and the corresponding Standard Deviation (StDev) of 5-fold cross-validation result.

	Circular [29]	Neural [9]	seq2seq [43]	seq3seq (Ours)
Mean	36.74%	60.80%	76.64%	**89.72%**
StDev	0.74%	1.35%	0.43%	0.41%

Table 2: The comparison of classification accuracy on the PM2-10k data. We report the average classification accuracy (Mean) and the corresponding Standard Deviation (StDev) of 5-fold cross-validation result.

	Circular [29]	Neural [9]	seq2seq [43]	seq3seq (Ours)
Mean	39.38%	52.27%	62.06%	**68.45%**
StDev	1.14%	1.12%	1.98%	0.80%

- **PM2-10k**: PM2-10k dataset contains 10,000 samples of SMILE strings and binary promiscuous class labels. Similarly, a threshold of 0.024896 was used to classify each SMILE. Samples with value larger than the threshold were considered as positive 1; otherwise, labeled as 0.

We mix the ZINC drug-like dataset with the labeled dataset and train the recovery and inference task simultaneously on the mixed dataset.

Neural Network Structures As we mentioned earlier, the proposed seq3seq fingerprint framework is super flexible in the choice of the network structure. Theoretically, both perceiver and interpreter network can use any stacked Recurrent Neural Network (RNN) with different layers and layer hidden sizes. Also the RNN cell can be formed in different types, e.g., LSTM, GRU, etc. Due to the page limit of this paper, we hereby assume the perceiver and interpreter network always use the same type of RNN cells with the same number of layers and hidden sizes. In this section, we only discuss Gated Recurrent Unit (GRU) [7] as the RNN cell. Also, we limit the discussion of the inference network to a single densely connected layer with the output number equaling the number of the classification class number. For simplicity, we use $GRU - L - H$ to represent the network structure, where GRU is the RNN cell type, $L \in N^+$ is the stacked RNN layer number and $H \in N^+$ is the RNN cell hidden size. For instance, $GRU - 2 - 256$ represents a seq3seq model where both perceiver and interpreter network use 2-layer GRU cell with 256 hidden units.

Learning Hyper-parameters For optimization, we use the Stochastic Gradient Descent (SGD) with a heuristic learning rate decaying schedule. The initial learning rate is 0.5 for any training models. The learning rate will be decayed by a factor of 0.99 if the test loss does not decrease after 600 training steps. The training will automatically halt if the learning rate is smaller than $1e-7$. Under the above hyper-parameter sets, the training of each model in the semi-supervised setting can generally finish within a few hours.

Evaluation Metrics Given that we have two tasks of our semi-supervised learning framework, i.e., recovery and inference task, we report two evaluation metrics for each model we trained. For recovery task, we use an Exact Match Accuracy (EMA) for evaluation. This metric measure the portion of the exactly recovered sequence within the entire set of sequences. Furthermore, we report the classification accuracy (hereafter SSLA for Semi-Supervised Learning Accuracy) for our classification task.

Comparison Methods We compare our semi-supervised method with the unsupervised seq2seq fingerprint method [43] as well as several other state-of-the-art methods: the ECFP [29] (circular fingerprint) and the neural fingerprint method [9]. We download the official implementation of the seq2seq fingerprint [1] and carefully follow the experimental setting of the authors. The circular fingerprint is a hand-crafted hash-based feature that was generated through RDKit [2]. The neural fingerprint implementation is obtained from https://github.com/HIPS/neural-fingerprint, which we slightly modify to adapt our dataset file format.

Infrastructure and Software The seq3seq fingerprint method was implemented through Tensorflow package [1], and our semi-supervised model was trained in a self-hosted 16-GPU cluster platform with Intel i7 6700K @ 4.00 GHz CPU, 64 Gigabytes RAM and four Nvidia GTX 1080Ti GPUs on each workstation. The code will be released upon the acceptance of this paper.

4.2 Comparison with State-of-the-art Methods

In Table 1 and 2, we report the 5-fold cross validation average classification accuracy on LogP and PM2-10k datasets. The proposed methods are compared with ECFP (circular) fingerprint [12], neural fingerprint [5] and seq2seq fingerprint [43]. For seq2seq fingerprint, according to their paper, the seq2seq fingerprint with length 1024 + Gradient Boosting always provides best performance, so we only report those results on our paper.

It is shown that on both datasets, the seq3seq fingerprint always provides best inference performance. On LogP dataset, our seq3seq model performs significantly superior than the other state-of-the-art methods, up to 13% in terms of classification accuracy (SSLA in the tables). Compared with circular fingerprint, the seq3seq fingerprint is data-driven and contains enough information to be recovered. The performance of neural fingerprint is generally limited by the availability of the labeled data. Seq2seq fingerprint is the closest

[1] https://github.com/XericZephyr/seq2seq-fingerprint
[2] http://www.rdkit.org

work in terms of accuracy for now since it can be also trained on the huge pool of unlabeled data, extracting a good representation and train/infer with a sophisticated classification model. However, seq2seq fingerprint is, unfortunately, not an end-to-end framework, which means the recovery and inference training of seq2seq fingerprint are separate. The unsupervised recovery training can bring in considerable amount of noise in the representation which limits further improvements of the inference performance. The seq3seq fingerprint, which uses the inference task to correct the recovery task during training, can constantly provide the best performance among all of the comparison methods.

Table 3: The performance variations with λ and GRU model parameters for LogP data. Layer: the stacked layer number of RNN cells. LD: Latent Dimension (hidden size) of RNN cells. EMA: Exact Match Accuracy for self-recovery task. SSLA: classification accuracy for inference task.

Layer	LD	λ	EMA	SSLA
2	128	1	86.31%	89.46%
		0.1	91.80%	89.62%
		0.01	90.23%	81.05%
		0.001	91.42%	64.95%
2	256	1	93.59%	90.18%
		0.1	94.52%	89.35%
		0.01	95.77%	84.65%
		0.001	95.48%	69.16%

Table 4: The performance variations with λ and GRU model parameters for PM2-10k data. Layer: the stacked layer number of RNN cells. LD: Latent Dimension (hidden size) of RNN cells. EMA: Exact Match Accuracy for self-recovery task. SSLA: classification accuracy for inference task.

Layer	LD	λ	EMA	SSLA
2	256	1	87.48%	65.28%
		0.1	89.84%	64.85%
		0.01	91.73%	62.37%
		0.001	91.31%	50.66%
3	256	1	82.40%	64.90%
		0.1	87.61%	67.92%
		0.01	89.33%	68.24%
		0.001	90.25%	50.07%

4.3 Sensitivity Analysis of Multi-task Weight Balance Parameters

In multi-task machine learning practice, the weight balancing hyper-parameters among different tasks (in our case, λ in the loss function) are sometimes critical and sensitive to data. This might not be an intriguing feature in practices. However, our method is quite robust and tolerant with λ variations. In this subsection, we report our sensitivity studies of λ. We choose different scale of λ to see how the final model performance responds to the variance of λ , showing the robustness of our method with regard to different weight balancing hyper-parameters.

In Table 3, 4 as well as Figure 5, we vary λ in the logarithm scale with a base of 10. We tried $10^0, 10^{-1}, 10^{-2}, 10^{-3}$. On both datasets, it looks that within a quite wide range of λ, i.e., $10^{-2} - 10^0$, the performance is quite robust to the change of λ. The reason behind this robustness might be the huge unlabeled data pool used in the training process. Given the model has been trained with a sufficiently large (up to dozens of millions) molecular data pool, the resulting model will automatically adjust to a small task weight perturbation.

4.4 The Ablation Study of Neural Network Structures

In this section, we provide a comprehensive study of the impacts of different layers and layer hidden sizes of our seq3seq fingerprint models. We report the 5-fold cross validation Exact Match Accuracy (EMA) and the classification accuracy (SSLA) in Table 5 and 6 for each of the two datasets, respectively. Figure 4 (a) and (b) also illustrates the trends when varying the layer numbers and layer hidden sizes.

Inference Task It is super exciting to reveal the **robustness of classification accuracy to the change of network structures** on both datasets. In Figure 4, the classification accuracy (blue bars) almost stays at the same height when varying the layer numbers and layer hidden sizes. This implies the importance of the representation learning inside the seq3seq fingerprint. This further support the positive effects of the large-scale (up to dozens of millions) unlabeled data utilization.

When the inference is super robust to the network changes, for self-recovery task (in terms of EMA), we observe a decreasing trend when increasing the layer depth (numbers). Meanwhile, the increasing number of hidden units inside each layer generally yields better EMA. This suggests that the improvement of self-recovery task has higher reliance on the layer hidden sizes. Deeper network might not always be an elixir for a simple auxiliary task like self-recovery. This observation might help future network design. To simultaneously ensure high inference performance and reduce training time (deeper network generally takes longer to train.), it might be a good idea to use reasonably deep and wide RNN networks.

5 CONCLUSIONS

In this paper, we discuss a new semi-supervised deep learning based molecular prediction system, called **seq3seq fingerprint**. Our model is the first attempt in sequence-based deep learning method utilizing both unlabeled and labeled data for drug discovery. The reinforcement from the unlabeled data is demonstrated to significantly improve the inference performance by enhancing the representation power of the perceiver network. As a result, the superior inference performance over multiple state-of-the-art methods is revealed in our extensive experiments.

In the future, a potential direction might be improving the training algorithm [28, 41, 42]. Furthermore, our seq3seq fingerprint method still share some common aspects with Natural Language Processing (NLP) area as the seq2seq fingerprint does [43]. As described in Section 3, it looks that we have found a new direction

Table 5: The comparison of 5-fold cross validation classification accuracy among different seq3seq GRU models on the LogP data. Both average (Mean) and Standard Deviation (StDev) are reported for the 5-fold splits. FP Length: FingerPrint Length. SSLA: classification accuracy for inference task. EMA: Exact Match Accuracy for self-recovery task.

	GRU-2-128	GRU-3-128	GRU-4-128	GRU-5-128	GRU-2-256	GRU-3-256	GRU-4-256	GRU-5-256
FP Length	256	384	512	640	512	768	1024	1280
SSLA Mean	89.62%	89.12%	89.05%	**89.72%**	89.48%	89.64%	88.90%	88.11%
SSLA StDev	0.62%	0.22%	0.10%	0.41%	0.44%	0.42%	0.31%	0.40%
EMA Mean	91.39%	85.75%	77.13%	68.64%	96.13%	94.24%	87.99%	83.86%
EMA StDev	0.46%	0.53%	0.56%	0.80%	0.21%	0.31%	0.45%	0.41%

Table 6: The comparison of 5-fold cross validation classification accuracy among different seq3seq GRU models on the PM2-10k data. Both average (Mean) and Standard Deviation (StDev) are reported for the 5-fold splits. FP Length: FingerPrint Length. SSLA: classification accuracy for inference task. EMA: Exact Match Accuracy for self-recovery task.

	GRU-2-128	GRU-3-128	GRU-4-128	GRU-5-128	GRU-2-256	GRU-3-256	GRU-4-256	GRU-5-256
FP Length	256	384	512	640	512	768	1024	1280
SSLA Mean	65.65%	67.11%	65.80%	67.23%	66.74%	68.08%	**68.45%**	67.09%
SSLA StDev	0.19%	0.85%	0.61%	0.52%	0.57%	0.35%	0.80%	0.67%
EMA Mean	83.84%	81.24%	78.60%	74.38%	92.49%	91.72%	87.36%	82.64%
EMA StDev	0.45%	0.67%	0.88%	0.88%	0.37%	0.25%	0.29%	0.76%

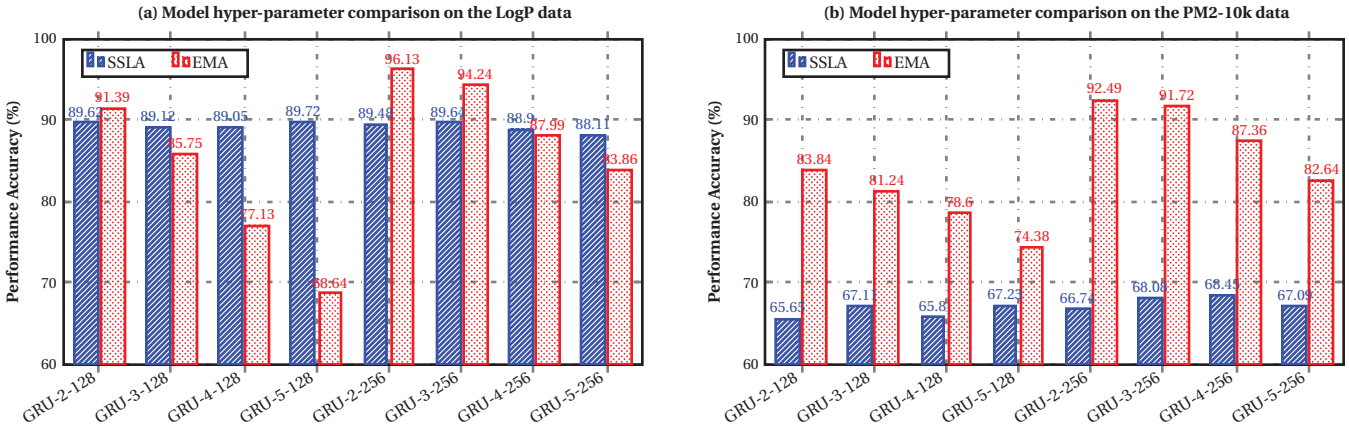


Figure 4: Impacts of the network structures on different metrics on both LogP and PM2-10k dataset. 1) The robustness of inference performance (SSLA, blue bars) is revealed. 2) The positive and negative correlations with regard to the self-recovery performance (EMA, red bars) are observed for RNN network depths and widths, respectively.

to invent new drug discovery methods. In the future, it might be interesting to further investigate bonds between drug discovery and NLP area, which might bring in many novel methods to further accelerate drug discovery research.

Acknowledgment The authors would like to thank NVIDIA for GPU donation and the NIH and UCSF for sharing the drug discovery datasets. The statements contained herein are solely of the authors and do not represent or imply concurrence or endorsement by NCI.

REFERENCES

[1] MartÃŋn Abadi and et.al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. http://download.tensorflow.org/paper/whitepaper2015.pdf

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[3] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. 2016. Low Data Drug Discovery with One-shot Learning. *arXiv preprint arXiv:1611.03199* (2016).

[4] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. 2017. Low data drug discovery with one-shot learning. *ACS central science* 3, 4 (2017), 283–293.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz,

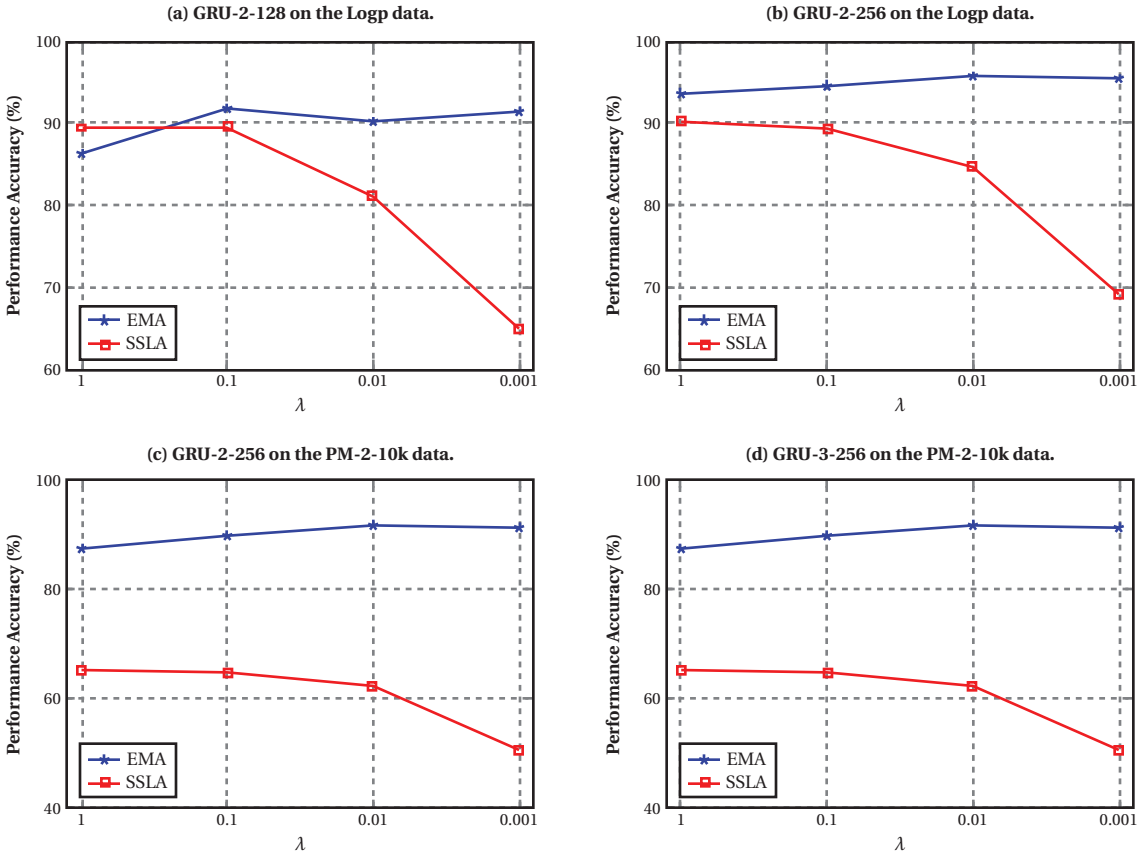


Figure 5: Impacts of the multi-task balance weights on different scales on both LogP and PM2-10k dataset. Within a very wide range (usually $10^{-2} - 10^{0}$), both self-recovery (EMA) and inference (SSLA) performance are quite robust to the change of λ.

Michael M Hoffman, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv* (2018), 142760.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[9] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*. 2224–2232.

[10] Yoav Freund and Robert E Schapire. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*. Springer, 23–37.

[11] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[12] Robert C Glen, Andreas Bender, Catrin H Arnby, Lars Carlsson, Scott Boyer, and James Smith. 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* 9, 3 (2006), 199.

[13] Joseph Gomes, Bharath Ramsundar, Evan N Feinberg, and Vijay S Pande. 2017. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv preprint arXiv:1703.10603* (2017).

[14] Tin Kam Ho. 1995. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, Vol. 1. IEEE, 278–282.

[15] Ye Hu, Eugen Lounkine, and Jürgen Bajorath. 2009. Improving the Search Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit-Density-Dependent Similarity Function. *ChemMedChem* 4, 4 (2009), 540–548.

[16] Junzhou Huang and Zheng Xu. 2017. Cell Detection with Deep Learning Accelerated by Sparse Kernel. In *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 137–157.

[17] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. 2012. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* 52, 7 (2012), 1757–1768.

[18] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* 30, 8 (2016), 595–608.

[19] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[21] Ruoyu Li and Junzhou Huang. 2017. Learning Graph While Training: An Evolving Graph Convolutional Neural Network. *arXiv preprint arXiv:1708.04675* (2017).

[22] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2018. Adaptive Graph Convolutional Neural Networks. *arXiv preprint arXiv:1801.03226* (2018).

[23] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.

[24] HL Morgan. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chemical Documentation* 5 (1965), 107–113.

[25] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*. 2014–2023.

[26] Noel M O'Boyle, Casey M Campbell, and Geoffrey R Hutchison. 2011. Computational design and selection of optimal organic photovoltaic materials. *The Journal of Physical Chemistry C* 115, 32 (2011), 16200–16210.

[27] Hao Pan, Zheng Xu, and Junzhou Huang. 2015. An effective approach for robust lung cancer cell detection. In *International Workshop on Patch-based Techniques in Medical Imaging*. Springer, 87–94.

[28] Zhongxing Peng, Zheng Xu, and Junzhou Huang. 2016. RSPIRIT: Robust self-consistent parallel imaging reconstruction based on generalized Lasso. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 318–321.

[29] David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling* 50, 5 (2010), 742–754.

[30] Chetan Rupakheti, Aaron Virshup, Weitao Yang, and David N Beratan. 2015. Strategy to discover diverse optimal molecules in the small molecule universe. *Journal of chemical information and modeling* 55, 3 (2015), 529–537.

[31] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. 2016. Computational Modeling of β-secretase 1 (BACE-1) Inhibitors using Ligand Based Approaches. *Journal of Chemical Information and Modeling* 56, 10 (2016), 1936–1949.

[32] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[33] Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A Smith. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528* (2017).

[34] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In *AAAI*, Vol. 4. 12.

[35] Izhar Wallach, Michael Dzamba, and Abraham Heifets. 2015. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* (2015).

[36] Sheng Wang, Jiawen Yao, Zheng Xu, and Junzhou Huang. 2016. Subtype cell detection with an accelerated deep convolution neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 640–648.

[37] David Weininger. 1970. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In *Proc. Edinburgh Math. SOC*, Vol. 17. 1–14.

[38] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 9, 2 (2018), 513–530.

[39] Zheng Xu and Junzhou Huang. 2015. Efficient lung cancer cell detection with deep convolution neural network. In *International Workshop on Patch-based Techniques in Medical Imaging*. Springer, 79–86.

[40] Zheng Xu and Junzhou Huang. 2016. Detecting 10,000 Cells in One Second. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 676–684.

[41] Zheng Xu and Junzhou Huang. 2017. A general efficient hyperparameter-free algorithm for convolutional sparse learning.. In *AAAI*. 2803–2809.

[42] Zheng Xu, Yeqing Li, Leon Axel, and Junzhou Huang. 2015. Efficient preconditioning in joint total variation regularized parallel MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 563–570.

[43] Zheng Xu, Sheng Wang, Feiyun Zhu, and Junzhou Huang. 2017. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In *BCB*.

[44] Jiawen Yao, Sheng Wang, Xinliang Zhu, and Junzhou Huang. 2016. Imaging biomarker discovery for lung cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 649–657.

[45] Feiyun Zhu, Jun Guo, Zheng Xu, Peng Liao, and Junzhou Huang. 2018. Group-driven Reinforcement Learning for Personalized mHealth Intervention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[46] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. 2017. The Kinetics Human Action Video Dataset.

Bioinformatic analysis of nucleotide cyclase functional centers and development of ACPred webserver

Nuo Xu
Department of Computer Science
Wenzhou-Kean University
88 Daxue Rd, Wenzhou, Ouhai, Zhejiang, China 325060
xunu@kean.edu

Changjiang Zhang
Department of Computer Science
Wenzhou-Kean University
88 Daxue Rd, Wenzhou, Ouhai, Zhejiang, China 325060
czhang@kean.edu

Leng Leng Lim
Department of Mathematics
Wenzhou-Kean University
88 Daxue Rd, Wenzhou, Ouhai, Zhejiang, China 325060
llim@kean.edu

Aloysius Wong
Department of Biology
Wenzhou-Kean University
88 Daxue Rd, Wenzhou, Ouhai, Zhejiang, China 325060
alwong@kean.edu

ABSTRACT

Cyclic mononucleotides, in particular 3',5'-cyclic guanosine monophosphate (cGMP) and 3',5'-cyclic adenosine monophosphate (cAMP), are molecular signals that mediate a myriad of biological responses in organisms across the tree of life. In plants, they transduce signals such as hormones and peptides perceived at receptors on the cell surface into the cytoplasm to orchestrate a cascade of biochemical reactions that enable them to grow and develop, and adapt to light, hormones, salt and drought stresses as well as pathogens. However, their generating enzymes (guanylyl cyclases, GCs and adenylyl cyclases, ACs) have just been recently discovered and are still poorly understood. Here, we employed a computational approach to probe the physicochemical properties of the catalytic centers of these enzymes and the knowledge of which, was used to create a web-based tool, ACPred (http://gcpred.com/acpred) for the prediction of AC functional centers from amino acid sequence. Understanding the nature of such catalytic centers have enabled the creation of predictive tools such as ACPred which will in turn, facilitate the discovery of novel cellular components across different systems.

CCS CONCEPTS

• **Computing methodologies** → *Data mining; Information extraction; Prediction algorithm;* • **Applied computing** → *Sequence analysis; Bioinformatics*

ACM-BCB'18, August 29–September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08… $15.00
DOI: https://doi.org/10.1145/3233547.3233549

KEYWORDS

Search Motif; Functional Centers; Sequence Analysis; ACPred; Adenylyl Cyclase Prediction; Webserver; Computational Biology

1 INTRODUCTION

Cells make use of molecular signals to relay information such as hormones and growth factors perceived at the cell surface, into the cell to trigger a cascade of biochemical reactions that will lead to responses at the physiological level [1, 2]. This transduction of signal from the external environment into the cell is crucial for the growth and development of an organism and, to allow the organism to efficiently response to changes in the environment [3-6]. Universal molecular signals such as small peptides, hormones, organic molecules, calcium ions and cyclic nucleotides (3',5'-cyclic guanosine monophosphate (cGMP) and 3',5'-cyclic adenosine monophosphate (cAMP)) have long been shown to exist in organisms across the tree of life and signaling a myriad of biological responses [7-13]. In animals including humans, cGMP is a vasodilating signal that is perhaps most famous for causing penile erection upon sexual stimulation [14]. In plants, cGMP also signals many biological responses including responses to light, hormones, salt and drought stresses as well as ozone and pathogens [for review, see [10, 13, 15]. Meanwhile, cAMP signals polarized pollen tube growth, stomatal opening, responses to light and temperature, and modulates ion transport [16-20]. These signaling molecules are particularly important for plants because unlike animals, they are sessile organisms and cannot run away from danger. They must therefore rely on a set of molecular signals such as cGMP and cAMP to provide efficient cellular signaling mechanisms in order for them to adapt and survive [1-3, 9, 21, 22]. However, the enzymes (guanylyl cyclases, GCs and adenylyl cyclases, ACs) that generate cGMP and cAMP in plants have just been recently discovered and are still poorly understood although they have been well-characterized in other systems such

as animals and bacteria. One reason for their apparent elusiveness is that plant cells have complex domain architecture of proteins that can perform multiple functions e.g., at the extracellular region, they can perceive and bind to ligands while at the cytosolic region, they can bind to proteins and/or organic compounds at modulatory sites or catalyze certain reactions [12, 13, 15, 23-26]. This is attributed to divergent evolution where plant signal perception and downstream cellular reactions are distinct from those of other eukaryotes. In a relatively crowded plant cell occupied by a large central vacuole, proteins assume multiple roles and GCs and ACs are well-placed in their microenvironments to perform highly localized signaling functions that include for e.g., switching from one signaling network to another [27-32]. Therefore, in plants, receptors and signaling molecules cannot be identified using standard homology-based searches querying with proteins from lower or higher eukaryotes because it is beyond detection limits, hence their apparent elusiveness [24, 33].

Recently, a motif-based approach has led to the discovery of a new class of GCs and ACs some of which, have been studied in greater detail [34]. This new class of enzymes are known as functional centers and are structurally different from canonical GCs and ACs found in other organisms. They are usually found embedded within larger primary domains in complex multi-functional proteins and while they possess the conserved key amino acids for catalysis, they do not resemble the overall structure of stand-alone canonical GCs and ACs [26, 33]. Their discovery has in recent years, prompted intriguing questions regarding their regulatory roles at both the molecular and biological levels. Emerging experimental data have shed light on some of these queries, but many more remain unanswered. Discovering new functional centers will help elucidate unknown functions and contribute to the understanding of the nature of these group of enzymes. As such, the discovery of these functional centers requires automation in the form of a web-server. We have recently created a web-tool for the prediction of GC functional center called GCPred (http://www.gcpred.com) and have tested this tool on both plant and animal proteins [33].

In the same manner, a predictive tool is required for the identification of candidate AC centers. Here, we probe the physicochemical properties of known GCs and ACs to understand the nature of these highly similar catalytic centers and use this knowledge to develop ACPred (http://gcpred.com/acpred).

2 METHODS

2.1 Bioinformatic analysis of AC and GC centers

Experimentally validated GC and AC centers from plant proteins were analyzed in terms of their overall domain organization and structural architecture. The domain organization of proteins were presented as 2D columns with lengths adjusted to approximately reflect their relative amino acid lengths and they are all aligned at their AC centers. Information about the annotated domains, transmembrane regions and other key components of the proteins were obtained from UniProt (http://www.uniprot.org) and presented in different colors accompanied by a legend. Representative 3D structure of each GC and an AC containing docked GTP or ATP substrates were prepared using UCSF Chimera visualization software available at https://www.cgl.ucsf.edu/chimera. The catalytic centers were colored, and substrate orientations clearly defined. The key amino acids in the motif that are involved in substrate interaction functions were colored according to surface charge.

Next, the physicochemical properties of the catalytic centers were analyzed based on known values of 3 categories: hydrophobicity, molecular weight and isoelectric point, of each amino acid. The data for each protein is presented individually as a heatmap with red representing highest and white representing lowest values in that category. Average values of GCs and ACs in each category were also represented in line graphs to show variations between GCs and ACs.

2.2 Development of the ACPred server

The website ACPred was written in HTML, PHP and CSS, running on an Ubuntu server that LNMP package (Nginx, MySQL, PHP) installed. The header of the home page with brown-white linear ingredient background was defined by a CSS library Bootstrap. The image set on the right of the title ACPred, is a representative model of an AC center docked with the ATP substrate, making it clear for the user to understand the purpose of our website, that is to identify candidate ACs. In the home page, we first explain the utility of this tool giving the background knowledge and other important details as well as providing detailed instructions written in large size Arial font to guide first-time users. Next, we include a ratio element following the instruction asking users if metal ion binding feature should be included in the prediction. We also include a FASTA format example that contains the name of the protein in the header and its full-length amino acid sequence. Below the example, there is a large text area that can automatically adjust its width according to the screen size of the web browser. Characters typed inside the text area will be automatically converted to small sized font with courier style and the user is also able to enter multiple sequences at one time. Although the height of text area is fixed at 400px, the user can still input as many as they want because the text area allows for over-flow. At the bottom of the text area, there is a button element that enables user to submit the input data as string to the server. This submit button is disabled until the text area is filled, otherwise a warning message will be shown. To quickly learn the feature of our webserver, the user only need to copy the example into the text area (full sequence required) and then click the submit button to view the result page. At the footer of the home page, we provide the prior works that have led to the development of this server as clickable reference links which will take the user directly to the primary source of the articles.

The main calculation is done in the result page and written in PHP. The header of the result page is almost the same as that in the home page, but we replaced the instruction with an interpretation guide that helps the user make good judgment on their retrieved hits. The AC report for each sequence can be

divided into 3 parts: a sequence panel, a hits table and 3 bar charts. If the users enter more than one sequence, these 3 parts will be shown as a loop, and the report for each sequence is independent. However, none of the 3 parts will be shown if no AC center was identified in any given sequence. The sequence panel contains a bold header with the protein name and a numbered body that shows 100 amino acids in each line. To do this, we created a PHP function "split" that could first split the string that was entered in the home page to several single sequences according to the special symbol ">" which is the symbol at the beginning of a FASTA format sequence and check if the full string contains more than one sequence. Then, position numbers will be generated for each sequence through the PHP function "position" if more than one hits (that contain 14 amino acids and satisfy the AC motif shown in Fig. 1) were identified within each sequence. Each hit with its own position number will be highlighted in the sequence panel and shown in the hits table. Next, to fill in the ACC hydrophobic value, ACC molecular weight, ACC isoelectric point and ACC mean value, we created another PHP function "input" that will call MySQL procedure "input check" with 2 parameters: the full sequence and the unique position number for each ACC hit. Thus, the number of invoke times of this PHP function depends on the number of ACC hits found. The procedure "input check" will first pick out the objective hit according to the 2 parameters that passed from the PHP function, then splits the hit into 14 amino acids. In MySQL database, we have a "hydrophobic table" that stores known amino acids' hydrophobic values, a "molecular weight" table that stores known amino acids' molecular weights, an "isoelectric point" table that stores known amino acids' isoelectric point values. The procedure will calculate those 4 values (termed ACC hydrophobicity, ACC molecular weight, ACC isoelectric point) for the objective AC hit according to the algorithm shown in Fig. 7B. Those 4 values will be scaled 0 to 1 with 3 decimal numbers and shown in 3 possible colors (green, red, black) based on specific cut off points determined from currently available experimental data. Color definition and labels are described under each table. Following each table, there are 3 bar charts created using the open source PHP graphic library "pchart". Because each hit contains 14 amino acids, the horizontal axis of the bar chart is fixed to represent these amino acids beginning from amino acid at position 1 to 14 respectively. Unlike the hits table where each row would only cover the values of one AC domain, the 3 types of the values (ACC hydrophobicity, ACC molecular weight, ACC isoelectric point) of the hits within the same sequence will be combined into the 3 bar charts named as ACC hydrophobicity charts, ACC molecular weight chart and ACC isoelectric point chart. All these 3 charts show the deviation from mean values for each hit. For instance, if the ACC hydrophobic value of one ACC domain is shown in green, then the average height of the bar in the hydrophobic chart will be closer to 0 than the black or red values. However, if there are more than one hit within the same sequence, each hit will be given a unique color, starting from green, then red, blue, etc. All the bar charts have the same height and weight, but they will automatically adjust the unit size of the vertical axis according the maximum and minimum heights of bars. At the bottom of the result page, there is a link that allow the user to download the retrieved hits as a *.csv file.

3 RESULTS AND DISCUSSION

3.1 Domain architecture of GCs and ACs

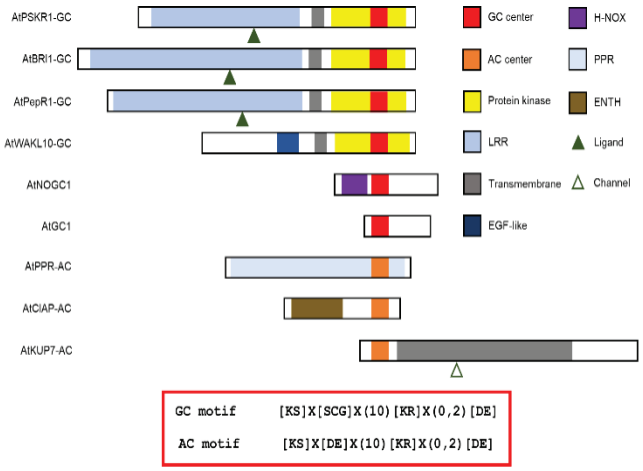


Figure 1: Domain organizations of GC and AC centers.

A motif-based approach has in the past, identified several GC and AC centers [13, 26]. The motif was constructed by including only conserved amino acids within the catalytic centers of canonical GCs and ACs from prokaryotes and eukaryotes. Specifically, the motifs consist of 14-amino acid long amino acids with amino acid in position 1 forming hydrogen bonds with guanine or adenosine of GTP or ATP respectively and amino acid in position 3 determining substrate specificity while amino acid in position 14 of the motif binds to the phosphate acyl group and stabilizes the transition of substrate to its cyclic form [35]. The motifs may also undergo rational modifications to include species-specific and metal-binding filters or amino acids of similar chemical properties and they have been particularly successful when used in tandem with structural modeling and docking simulations [23-25]. In higher plants, known GC centers are AtPSKR1, AtBRI1, AtWAKL10, AtPepR1, AtPNP-R1, AGNOGC1, AtGC1, PnGC1 and HpPepR1 [35-43] while AtClAP, AtPPR-AC, AtKUP7 remain the only experimentally confirmed AC centers to-date [44]. Here, we show the domain architecture of experimentally confirmed proteins with GC and AC activities from the model plant *Arabidopsis thaliana* in Fig. 1. It is obvious that these catalytic centers occupy complex proteins that have different primary functions. For instance, many GC centers are found embedded within a larger kinase domain of hormone/peptide receptor complexes (AtPSKR1, AtBRI1, AtPepR1, AtPNP-R1 and AtWAKl10) thus suggestive of a role for GCs in regulating reactions in the hormone/peptide-dependent pathways [36, 37, 39-41, 45]. Indeed, the GC activities of AtBRI1 and AtPSKR1 have been shown to be intricately linked to their kinase domains which they reside in [27, 32, 37, 45]. Furthermore, binding of the extracellular receptor domains to their natural ligands can elevate

cytosolic cGMP levels and in the case of AtPNP-R1, enables regulation of ion and water homeostasis [41]. Meanwhile, AC centers are found in proteins that have more varied primary functions. For example, AtKUP7 acts primarily as a potassium transporter while AtClAP which assembles clathrin during endocytosis, is implicated in plant defense [44, 46, 47].

In addition to the domain architectures, we also show the 3D structures of typical GC and AC centers using AtPSKR1 and AtKUP7 as representatives (Fig. 2). From a structural perspective, the GC and AC centers share similar secondary folds where amino acids from position 1 to 14 of the motif form an alpha-helix that is followed by a solvent-exposed loop harboring a positively charged [RK] amino acid. At the tertiary level, they form a clear cavity that spatially fits the GTP or ATP substrate albeit accommodating them only at specific substrate orientations. Previous structural works have ascertained that the substrate must assume a binding pose where the nucleotide region of the substrate points towards the residue at position 1 of the motif located deep into the pocket at the catalytic center while the hydrophilic negatively charged phosphate end points towards the positively charged amino acid at position 14 of the motif and protruding outward from the cavity orifice. This binding pose is deemed favorable for catalysis [23-25]. The structural similarity between GC and AC centers is not surprising given that their substrates share considerable physical and chemical properties. But, is there a difference in substrate affinity between such GC and AC centers? And if so, how do they discriminate the GTP and ATP substrates?

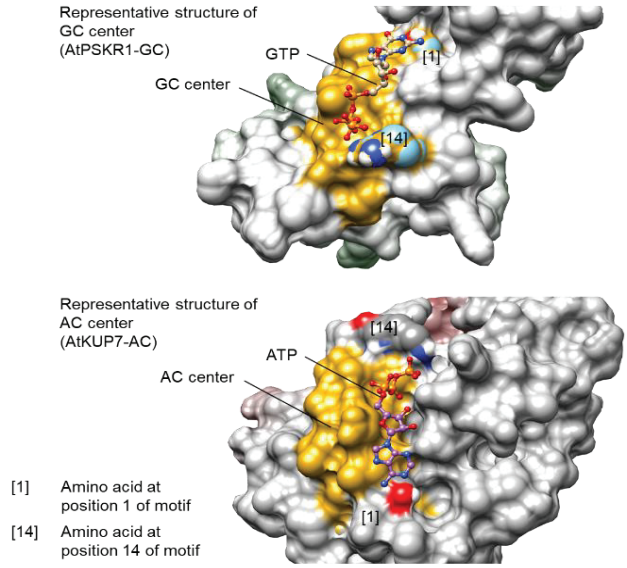


Figure 2: Representative structure of GC and AC centers.

3.2 Physicochemical signatures of GCs and ACs

Previous mutational works and computational simulations have showed hints of substrate preference at the catalytic center of GCs. For instance, in AtBRI1, mutations at position 3 ($G^{989}K$ and $G^{989}I$) of the motif reduced the catalytic activity of the GC center with G-I mutation being the most severe [45] while previous computational simulations on AtPSKR1 also indicate reduced substrate affinity when amino acid at the same position of the motif was mutated [25].

Since mutation of the amino acid at position 3 retains some catalytic activity [45] and it is also unlikely that these catalytic centers can spatially discriminate GTP and ATP, substrate preference must therefore be conferred by surface charges and/or other physical means such as hydrophobicity. To this end, we probe the physicochemical properties of known GC and AC centers. In particular, we analyzed the hydrophobicity, isoelectric point and molecular weight of amino acids in the GC and AC centers and expressed them as heatmap in Fig. 3. Among the GCs, the amino acids at each position of the motif show consistent physicochemical properties but among the ACs, these properties are more varied from one protein to another. For instance, the hydrophobicity of amino acids residing at positions 5 and 8 of the motif show considerable variation among the ACs while the same is also observed for the isoelectric point of ACs at position 10 (red boxes in Fig. 3). In general, the AC centers show greater variation in physicochemical properties across position 1 to 14 of the motif compared to the GC centers and we argue that this is necessary to compensate for the lack of 'stickiness' of the adenine nucleotide that lacked a =O group (red circle in Fig. 4).

We also expressed this analysis as average values for all three physicochemical properties in Fig. 5. When comparing ACs to GCs, there is significant difference in the hydrophobicity and isoelectric point of the amino acids at the positions indicated by red arrows in Fig. 5. We suspect that this, and the greater variation among amino acids within the centers might be an intrinsic nature of ACs that is necessary for optimal binding to the ATP substrate. Unlike GTP, ATP lacks a =O group (see blue and red circles in Fig. 4), thus may require greater difference in amino acid charges and hydrophobicity at the catalytic center for optimal binding and catalysis. It was previously suggested that the negatively charged amino acid [DE] at position 3 of AC motif confers substrate specificity to ATP [48] but it is likely that intermediary or flanking amino acid residues at the catalytic center play a role as well since mutations at position 3 of this motif did not completely abolish enzymatic activity of AtBRI1-GC [45].

Here, we showed that there is indeed considerable difference in the physicochemical properties of intermediary amino acids between GCs and ACs where in ACs, there is high variation in these properties which together with the negatively charged amino acid at position 3 [DE] of the motif, enable ACs to bind ATP. Other factors such as spatial and temporal abundance of GTP and ATP in microenvironments of the cell as well as the dependence on extracellular ligand binding and catalytic activity of primary domains such as kinases, can regulate cyclic mononucleotide generation by GCs and ACs. Further research is required to determine how these functional centers discriminate their substrates as computational methods including the bioinformatics analysis given here as well as docking simulations and biochemical evidence done elsewhere, have previously demonstrated that these catalytic centers can discriminate substrates despite being very similar in nature [15, 25, 33].

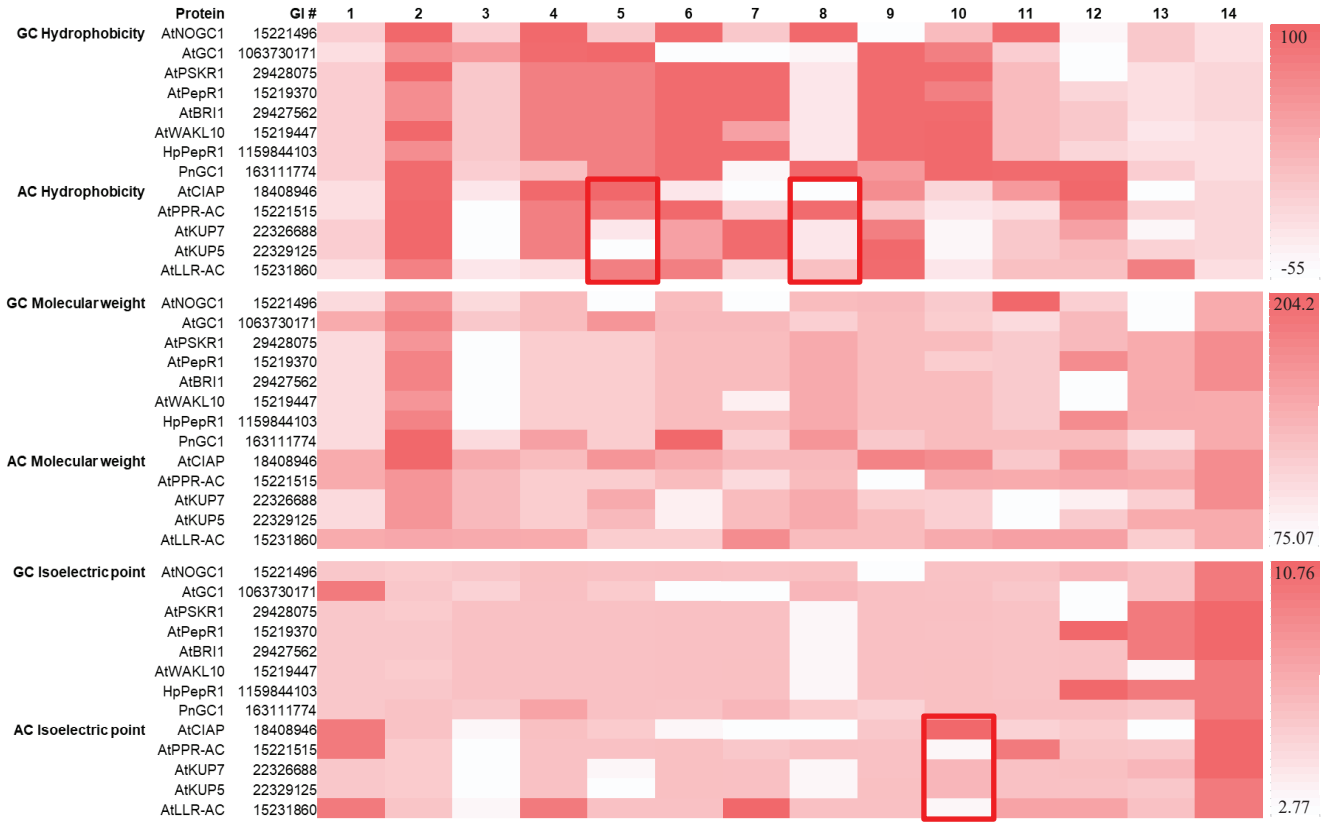


Figure 3: Heatmap illustrating the physicochemical properties of known GC and AC centers.

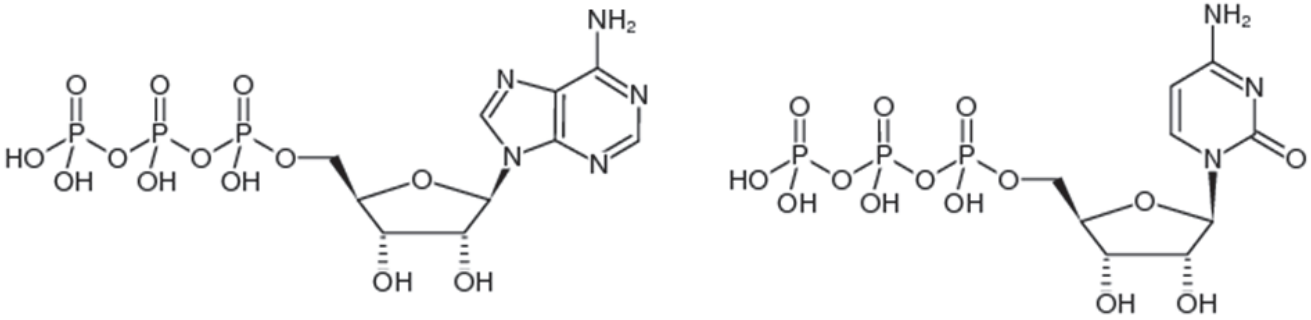


Figure 4: Structure of ATP (left) and GTP (right).

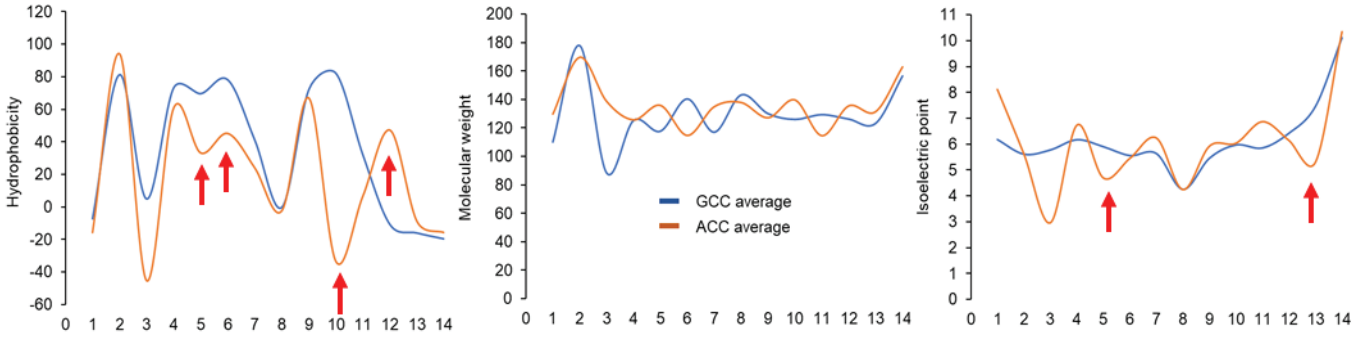


Figure 5: Average physicochemical values of known GC and AC centers.

3.3 The ACPred server

In order to rapidly predict candidate AC centers, we have developed ACPred. ACPred is a webserver built based on the algorithm and function of its parent server GCPred created previously for the prediction of GC centers [33]. The ACPred server allows user to enter single or multiple amino acid sequences in FASTA format and returns predicted hits ranked based on a set of numerical ACC values. The server uses algorithm that calculates predicted ACCs based on a set of physicochemical properties of amino acids in experimentally validated ACCs as presented in Fig. 6, where hits that contain the conserved amino acids at positions 1, 3 and 14 of a continuous string of amino acids are assigned statistical values 0-1, with 1 being the closest to the physicochemical properties of current ACC population. After submitting the queried sequence, the server returns a table of predicted ACCs that is accompanied by a set of ACC values normalized 0-1, which are color-coded to aid interpretations. In addition to the table, the result page also provides visual aids in the form of graphs that depict deviations of amino acids at individual positions of the ACC from mean values calculated from known ACC population. This comparative analysis at the single amino acid level may be useful for those interested in further probing of their candidates by e.g., guiding mutation and structural experiments. Previous works have established that this class of ACC resembles the GCCs where it typically contains 14 amino acids where the amino acids at positions 1, 3 and 14 have direct substrate binding and catalytic functions. We note that ACPred is only able to predict functional AC centers and not canonical AC domains or transmembrane regions.

The workflow and algorithm of ACPred is presented in Fig. 7. In step 1 of the ACPred workflow, the user enters single or multiple amino acid sequence in FASTA format and then click submit. The server screens the user input sequence for the presence of conserved amino acids at positions 1, 3 and 14 in continuous fashion in step 2 and if present, a data filtering process in step 3 removes other amino acids from the input sequence thus retaining only the identified ACC candidates. In step 4, a set of calculations are performed to determine the physicochemical properties of amino acids at each position of the ACC candidates and assigned values of 0-1 based on how close they resemble values of experimentally validated ACC population. The equations are this calculation is presented in the algorithm of Fig. 7. Specifically, if AC Domain "j" exists in sequence A, the AC algorithm calculates ACC hydrophobic (GH), molecular weight (GW), isoelectric point (GP) and ACC mean ($\bar{A}$) values (0-1) using the equations shown in the "perform calculations" box for k^{th} amino acid (where k = intermediary amino acids in the ACC) based on mean values of ACC population in the ACC database. The algorithm then generates a report that includes tables of ACC values and charts depicting variation from population mean and HTTP response sends the result page to the user in step 5 of the workflow. ACC database in Fig. 6 contains mean physicochemical values of amino acids at each position of known ACCs and from which calculations of input sequence were based on. Calculated values of each amino acid property were scaled 0-1 giving rise to "ACC values" where 1 represents closest to ACC population mean thus most probable and 0 is least probable.

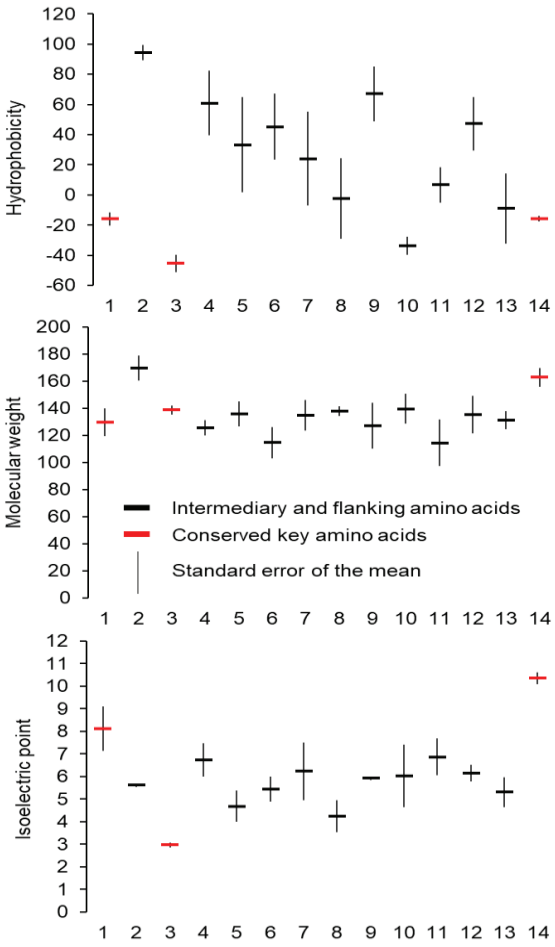


Figure 6: Mean physicochemical values of amino acids at each position of the 14-amino acid long AC motif calculated from experimentally validated ACC population.

An ACC mean value is also generated by averaging the three properties. The ACC values are color-coded where highly recommended hits should contain at least 2 green physicochemical values in addition to green ACC mean and contain no red values. Based on currently available experimental data, cut-off ACC values are determined where the values higher than the upper cut-off limit are colored green (high confidence) and those below the lower cut-off limit are colored red (low confidence). The ACC values are currently set at upper/lower limits of 0.700/0.500 for all physicochemical properties considered (i.e., hydrophobicity, Molecular Weight, Isoelectric point) as well as for the ACC Mean. ACPred also presents an option for metal ion binding which is typically afforded by negatively-charged amino acids at 0-2 positions downstream of the ACC [12]. ACPred is available at http://gcpred.com/acpred without registration or license.

The ACPred webserver is to our knowledge, currently the only tool that can rapidly identify ACs and importantly, it provides statistical values in the form of ACC values which allow user to order the retrieved hits. The latter feature is especially useful in high-throughput applications where ACC values can serve as a reliable indicator of confidence. As such, ACPred provides an added layer of confidence and a way of ranking retrieved hits in the form of scaled color-coded 0-1 ACC values. Previously, the AC motif used for the discovery of AC centers have identified several candidates from the proteome of *Arabidopsis thaliana* but there is no way of ranking them [12]. Using this server, we can now rank selected candidate ACs based on ACC values generated from algorithm that considers the physicochemical properties of intermediary amino acids (Table 1). We demonstrated the utility of this server on the AC candidates reported by [12] and they all contain hits of high confidence (Table 1). However, we note that the predictive strength of ACPred may be weaker in comparison to its parent webserver GCPred (used for the prediction of GC centers). This is due to the fact that experimentally validated AC centers are more varied in terms of their physicochemical properties which we have suggested to be an intrinsic nature of AC centers to enable more optimal binding to the ATP substrate. In addition, AC centers have only been recently identified and currently lack detailed characterization e.g., they are lacking mutational, structural and biochemical analyses that might reveal their substrate binding and inter-domain regulatory mechanisms. As such, we have decided to introduce a relatively low cut-off points to provide less stringent prediction conditions. We will continuously refine the ACC cut-off values to improve the predictive strength and reliability as more experimental data surface and also extend its service to predict other modulatory sites in the near future [49].

Table 1: Testing of Arabidopsis AC centers on ACPred

Name; TAIR ID	Position	ACC Hp	ACC MW	ACC Ip	ACC Mean
*AtClAP; At1g68110	329-342	0.646	0.780	0.724	0.716
*AtPPR; At1g62590	485-498	0.706	0.731	0.765	0.734
*AtKUP7; At5g09400	80-93	0.801	0.810	0.904	0.839
At1g25240	324-337	0.562	0.782	0.821	0.722
At2g34780	271-284	0.743	0.780	0.650	0.725
At3g02930	149-162	0.815	0.762	0.851	0.809
At3g04220	62-75	0.537	0.735	0.743	0.672
At3g18035	382-395	0.684	0.853	0.735	0.757
At3g28223	276-289	0.637	0.839	0.847	0.774
At4g39756	250-263	0.718	0.747	0.793	0.753

Note: These are candidate AC centers reported by [12] and * indicates experimentally confirmed ACCs. Hp: Hydrophobicity; MW: Molecular weight; Ip: Isoelectric point.

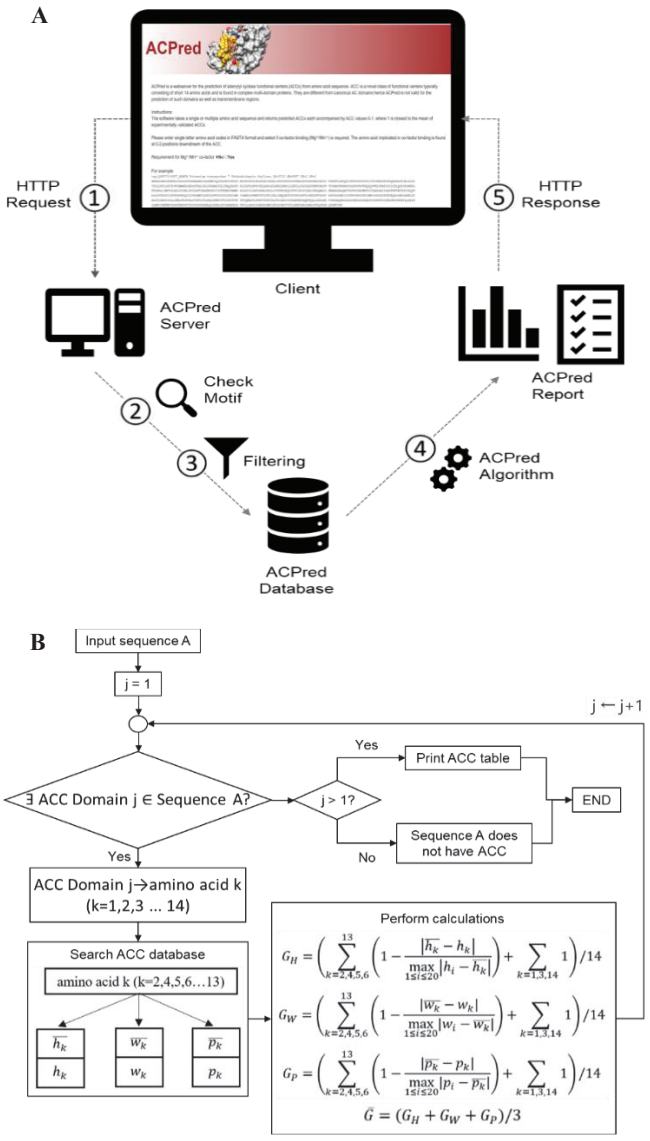


Figure 7: ACPred webserver workflow and algorithm.

4 CONCLUSION

In summary, we have conducted bioinformatic analysis on both GC and AC centers specifically probing their physicochemical properties to determine if there is any molecular basis for GTP and ATP substrate discrimination. Based on our analysis, we have linked the higher variation in charge and hydrophobicity within AC centers in addition to previously assigned [DE] amino acid at position 3 of the motif, for preferential ATP binding. We have also presented the development of a new webtool ACPred, that can rapidly predict candidate AC centers with novel ACC values that enable ranking of retrieved hits. We believe that the understanding of the nature of these new class of catalytic centers (GCs and ACs) have enabled the creation of predictive tools such as ACPred which will in turn, facilitate the discovery of novel cellular components across different biological systems.

ACKNOWLEDGMENT

This research is supported by the Student-partnering-with-Faculty (SpF) program (Grant Number: WKU201718009 awarded to AW) of Wenzhou-Kean University, Zhejiang China. The authors would like to thank Prof. Chris Gehring (University of Perugia, Italy) for testing the server and providing valuable feedback.

REFERENCES

[1] S. Gilroy, and A. Trewavas, 2001. Signal processing and transduction in plant cells: the end of the beginning? *Nature Reviews Molecular Cell Biology* 2, 4, 307-314.

[2] A. Trewavas, 2002. Plant intelligence: Mindless mastery. *Nature* 415, 6874, 841-841.

[3] C. M. Prasch, 2015. Signaling events in plants: Stress factors in combination change the picture. *Environmental and Experimental Botany* 114, 4-14.

[4] J. D. Jones, and J. L. Dangl, 2006. The plant immune system. *Nature* 444, 323-329.

[5] S. E. Clark, 2001. Cell signalling at the shoot meristem. *Nature Reviews Molecular Cell Biology* 2, 4, 276-284.

[6] A. Santner, and M. Estelle, 2009. Recent advances and emerging trends in plant hormone signalling. *Nature* 459, 7250, 1071-1078.

[7] D. R. McCarty, and J. Chory, 2000. Conservation and Innovation in Plant Signaling Pathways. *Cell* 103, 2, 201-209.

[8] N. Pauly, M. R. Knight, P. Thuleau, A. H. van der Luit, and others. 2000. Cell signalling: Control of free calcium in plant cell nuclei. *Nature* 405, 6788, 754-755.

[9] S. Meier, L. Madeo, L. Ederli, L. Donaldson, and others. 2009. Deciphering cGMP signatures and cGMP-dependent pathways in plant defence. *Plant Signaling & Behavior* 4, 4, 307-309.

[10] F. Lemtiri-Chlieh, L. Thomas, C. Marondedze, H. Irving, and others. 2011. *Cyclic Nucleotides and Nucleotide Cyclases in Plant Stress Responses*. InTech, Rijeka, Croatia.

[11] L. Ederli, S. Meier, A. Borgogni, L. Reale, and others. 2008. cGMP in ozone and NO dependent responses. *Plant Signaling & Behavior* 3, 1, 36-37.

[12] C. Gehring, 2010. Adenyl cyclases and cAMP in plant signaling - past and present. *Cell Communication and Signaling* 8, 15.

[13] C. Gehring, and I. S. Turek, 2017. Cyclic nucleotide monophosphates and their cyclases in plant signaling. *Frontiers in Plant Science* 8, 1074.

[14] D. I. Lee, G. Zhu, T. Sasaki, G.-S. Cho, and others. 2015. Phosphodiesterase 9A controls nitric-oxide-independent cGMP and hypertrophic heart disease. *Nature* 519, 472.

[15] C. Marondedze, A. Wong, L. Thomas, H. Irving, and others. 2017. *Cyclic Nucleotide Monophosphates in Plants and Plant Signaling.* Springer International Publishing, Cham (ZG), Switzerland.

[16] A. Moutinho, P. J. Hussey, A. J. Trewavas, and R. Malho. 2001. cAMP acts as a second messenger in pollen tube growth and reorientation. *Proceedings of the National Academy of Sciences of the United States of America* 98, 18, 10481-10486.

[17] L. Thomas, C. Marondedze, L. Ederli, S. Pasqualini, and others. 2013 Proteomic signatures implicate cAMP in light and temperature responses in Arabidopsis thaliana. *Journal of Proteomics* 83, 47-59.

[18] N. Curvetto. 1994. Effect of two cAMP analogs on stomatal opening in Vicia faba : possible relationship with cytosolic calcium concentration. *Plant Physiology and Biochemistry* 32, 365-372.

[19] F. Lemtiri-Chlieh, and G. A. Berkowitz. 2004. Cyclic adenosine monophosphate regulates calcium channels in the plasma membrane of Arabidopsis leaf guard and mesophyll cells. *Journal of Biological Chemistry* 279, 34, 35306-35312.

[20] A. J. Trewavas. 1997. Plant cyclic AMP comes in from the cold. *Nature* 390, 6661, 657-658.

[21] A. Trewavas. 2002. Plant cell signal transduction: the emerging phenotype. *Plant Cell* 14, Suppl, S3-S4.

[22] S. H. Spoel, and X. Dong. 2012. How do plants achieve immunity? Defence without specialized immune cells. *Nature Reviews Immunology* 12, 2, 89-100.

[23] A. Wong, and C. Gehring. 2013. The Arabidopsis thaliana proteome harbors undiscovered multi-domain molecules with functional guanylyl cyclase catalytic centers. *Cell Communication and Signaling* 11, 48.

[24] A. Wong, and C. Gehring. 2013. *Computational identification of candidate nucleotide cyclases in higher plants.* Methods in Molecular Biology 1016, 195-205.

[25] A. Wong, C. Gehring, and H. R. Irving. 2015. Conserved functional motifs and homology modelling to predict hidden moonlighting functional sites. *Frontiers in Bioengineering and Biotechnology* 3, 82.

[26] A. Wong, X. Tian, C. Gehring, and C. Marondedze. 2018. Discovery of novel functional centers with rationally designed amino acid motifs. *Computational and Structural Biotechnology Journal* 16, 70-76.

[27] L. Freihat, V. Muleya, D. T. Manallack, J. I. Wheeler, and others. 2014. Comparison of moonlighting guanylate cyclases: roles in signal direction? *Biochemical Society Transactions* 42, 6, 1773-1779.

[28] C. J. Jeffery. 1999. Moonlighting proteins. *Trends in Biochemical Sciences* 24, 1, 8-11.

[29] C. J. Jeffery. 2015. Why study moonlighting proteins? *Frontiers in Genetics* 6, 211.

[30] V. Muleya, J. I. Wheeler, O. Ruzvidzo, L. Freihat, and others. 2014. Calcium is the switch in the moonlighting dual function of the ligand-activated receptor kinase phytosulfokine receptor 1. *Cell Communication and Signaling* 12, 60.

[31] P. Tompa, C. Szász, and L. Buday. 2005. Structural disorder throws new light on moonlighting. *Trends in Biochemical Sciences* 30, 9, 484-489.

[32] H. R. Irving, D. M. Cahill, and C. Gehring. 2018. Moonlighting proteins and their role in the control of signaling microenvironments, as exemplified by cGMP and phytosulfokine receptor 1 (PSKR1). *Frontiers in Plant Science* 9, 415.

[33] N. Xu, D. Fu, S. Li, Y. Wang, and others. 2018. GCPred: a web tool for guanylyl cyclase functional centre prediction from amino acid sequence. *Bioinformatics* 34, 12, 2134-2135.

[34] S. Meier, C. Seoighe, L. Kwezi, H. Irving, and others. 2007. Plant nucleotide cyclases: an increasingly complex and growing family. *Plant Signaling & Behavior* 2, 6, 536-539.

[35] N. Ludidi, and C. Gehring. 2003. Identification of a novel protein with guanylyl cyclase activity in Arabidopsis thaliana. *Journal of Biological Chemistry* 278, 8, 6490-6494.

[36] L. Kwezi, S. Meier, L. Mungur, O. Ruzvidzo, and others. 2007. The Arabidopsis thaliana brassinosteroid receptor (AtBRI1) contains a domain that functions as a guanylyl cyclase in vitro. *PloS ONE* 2, 5.

[37] L. Kwezi, O. Ruzvidzo, J. I. Wheeler, K. Govender, and others. 2011. The phytosulfokine (PSK) receptor is capable of guanylate cyclase activity and enabling cyclic GMP-dependent signaling in plants. *Journal of Biological Chemistry* 286, 25, 22580-22588.

[38] T. Mulaudzi, N. Ludidi, O. Ruzvidzo, M. Morse, and others. 2011. Identification of a novel Arabidopsis thaliana nitric oxide-binding molecule with guanylate cyclase activity in vitro. *FEBS Letters* 585, 17, 2693-2697.

[39] S. Meier, O. Ruzvidzo, M. Morse, L. Donaldson, and others. 2010. The Arabidopsis wall associated kinase-like 10 gene encodes a functional guanylyl cyclase and is co-expressed with pathogen defense related genes. *PloS ONE* 5, 1, e8904.

[40] Z. Qi, R. Verma, C. Gehring, Y. Yamaguchi, and others. 2010. Ca2+ signaling by plant Arabidopsis thaliana Pep peptides depends on AtPepR1, a receptor with guanylyl cyclase activity, and cGMP-activated Ca2+ channels. *Proceedings of the National Academy of Sciences of the United States of America* 107, 49, 21193-21198.

[41] I. Turek, and C. Gehring. 2016. The plant natriuretic peptide receptor is a guanylyl cyclase and enables cGMP-dependent signaling. *Plant Molecular Biology* 91, 3, 275-286.

[42] A. Szmidt-Jaworska, K. Jaworski, A. Pawelek, and J. Kocewicz. 2009. Molecular cloning and characterization of a guanylyl cyclase, PNGC-1, involved in light signaling in Pharbitis nil. *Journal of Plant Growth Regulation* 28, 367-380.

[43] B. Swiezawska, K. Jaworski, M. Duszyn, A. Pawelek, and others. 2017. The Hippeastrum hybridum PepR1 gene (HpPepR1) encodes a functional guanylyl cyclase and is involved in early response to fungal infection. *Journal of Plant Physiology* 216, 100-107.

[44] I. Al-Younis, A. Wong, and C. Gehring. 2015. The Arabidopsis thaliana K(+)-uptake permease 7 (AtKUP7) contains a functional cytosolic adenylate cyclase catalytic centre. *FEBS Letters* 589, 24 Pt B, 3848-3852.

[45] J. I., Wheeler, A. Wong, C. Marondedze, A. J. Groen, and others. 2017. The brassinosteroid receptor BRI1 can generate cGMP enabling cGMP-dependent downstream signaling. *Plant Journal* 91, 4, 590-600.

[46] P. Chatukuta, T. Dikobe, D. Kawadza, K. Sehlabane, and others. 2018. An Arabidopsis clathrin assembly protein with a predicted role in plant defense can function as an adenylate cyclase. *Biomolecules* 8, 2, E15.

[47] O. Ruzvidzo, B. T. Dikobe, D. T. Kawadza, G. H. Mabadahanye, and others. 2013. *Recombinant Expression and Functional Testing of Candidate Adenylate Cyclase Domains.* Humana Press, New York City, USA.

[48] C. L. Tucker, J. H. Hurley, T. R. Miller, and Hurley, J. B. 1998. Two amino acid substitutions convert a guanylyl cyclase, RetGC-1, into an adenylyl cyclase. *Proceedings of the National Academy of Sciences of the United States of America* 95, 11, 5993-5997.

[49] A. Ooi, F. Lemtiri-Chlieh, A. Wong, and C. Gehring. 2017. Direct modulation of the guard cell outward-rectifying potassium channel (GORK) by abscisic acid. *Molecular Plant* 10, 11, 1469-1472.