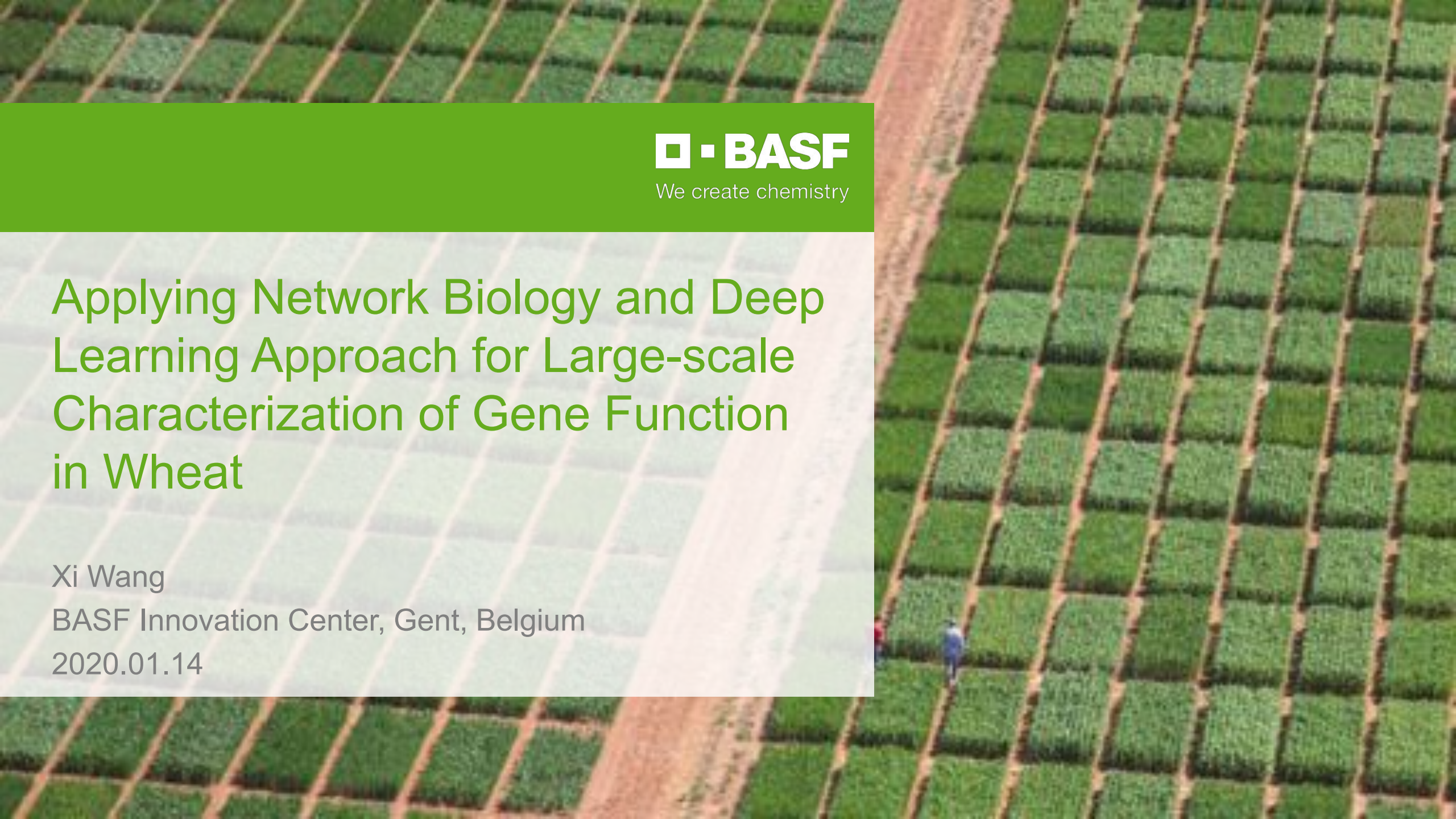


BASF
We create chemistry

Applying Network Biology and Deep Learning Approach for Large-scale Characterization of Gene Function in Wheat

Xi Wang
BASF Innovation Center, Gent, Belgium
2020.01.14

Outlines

- Who are we – organization and mission statement
- Functional annotation platform – motivation and overview
- Two ongoing projects
  - Integrative network analysis for gene discovery in wheat
  - Protein sequence based functional annotation using deep learning
- Summary and future work

Trait Research within BASF Seeds & Traits R&D

Mission: Create **gene-based solutions** to improve traits in crops that benefit farmers and society

The entire BAYER Trait Research has been merged to BASF since 1 August, 2018

- ❑ BASF commits to seeds and traits business which complements chemical crop protection
- ❑ ~350 people, Morrisville US and Gent Belgium

Trait Research activities in **Gent**

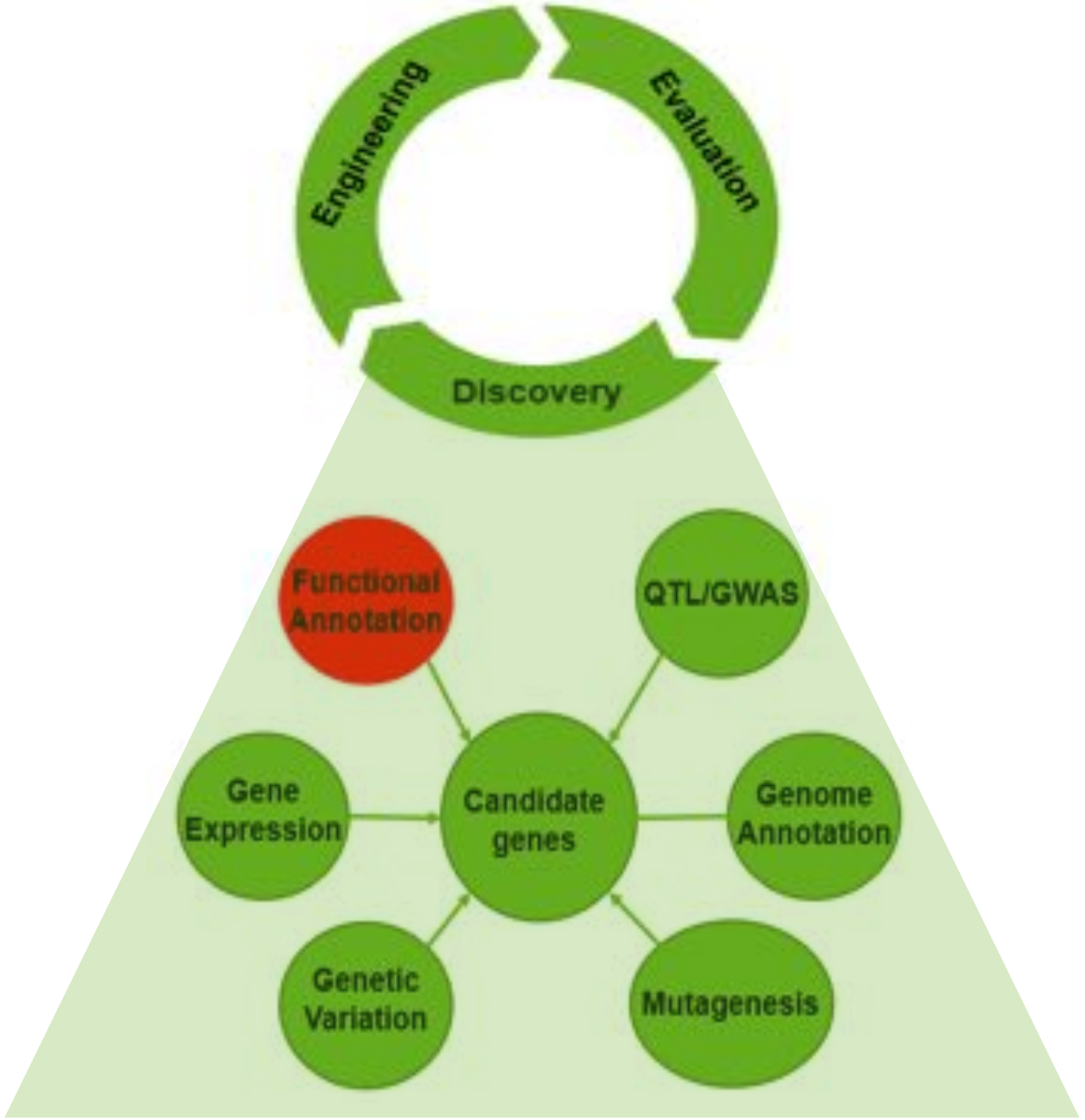
- ❑ **Discover** and understand new leads and optimize existing ones
- ❑ **Engineer and evaluate** new leads in-planta
- ❑ Novel technology development and partnership with external innovators
- ❑ Crops such as wheat and canola

BASF
We create chemistry

Functional annotation platform

Part of discovery pipeline to identify and understand gene function

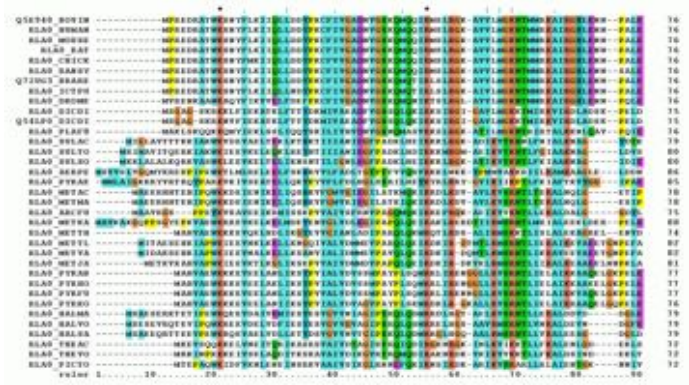
- Motivation
  - Lack of functional characterization for crop genes including wheat
  - Molecular experiments are reliable, but low-throughput → Computational prediction rapidly generate hypothesis about roles of candidate/unknown genes
  - Homology based approach is error prone and suffers from complex many-to-many homology relationsip → Requirments of additional data sets and methologies



BASF
We create chemistry

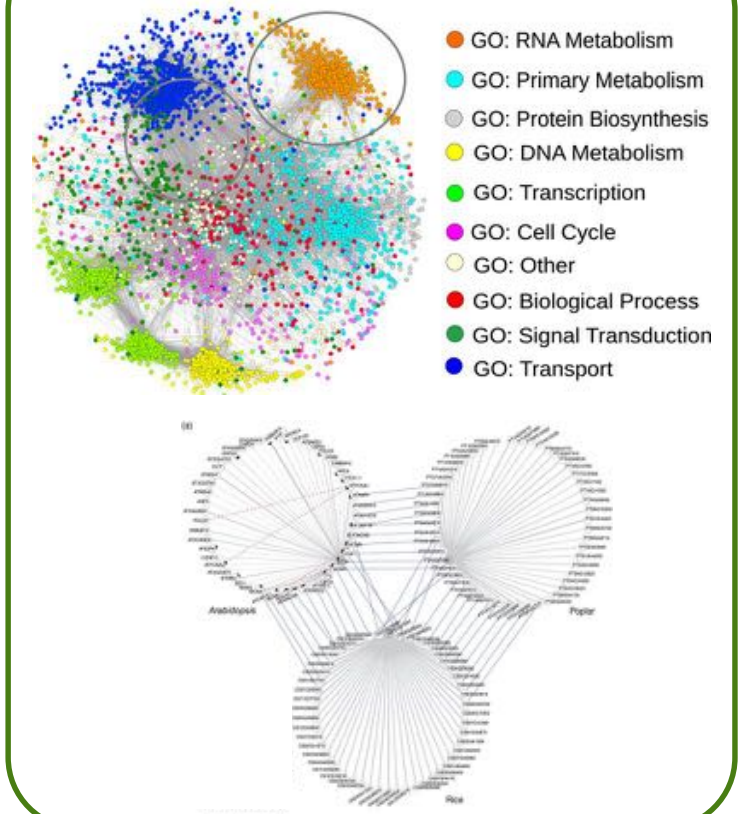
Functional annotation platform – components

Sequence Homology

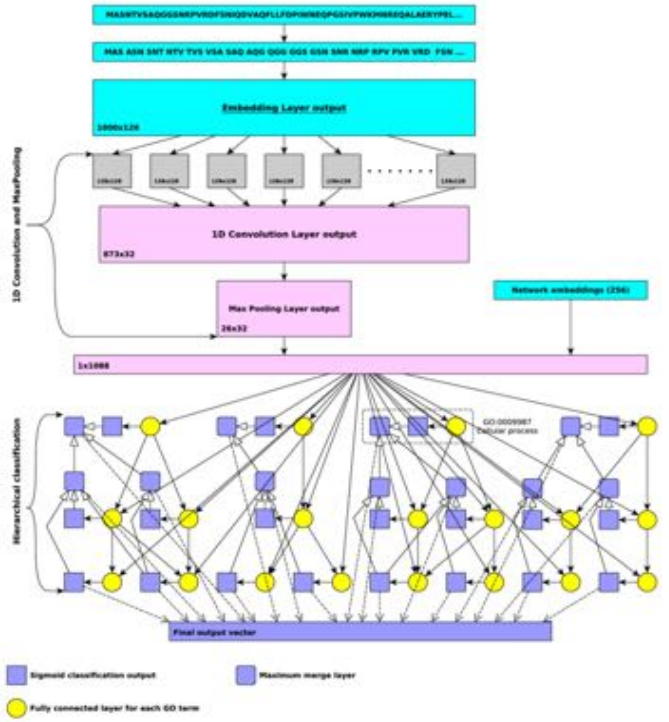


- BLAST2GO
- InterproScan
- Integrative orthology
- Gene family enrichment
- External resource

Network/Systems Biology

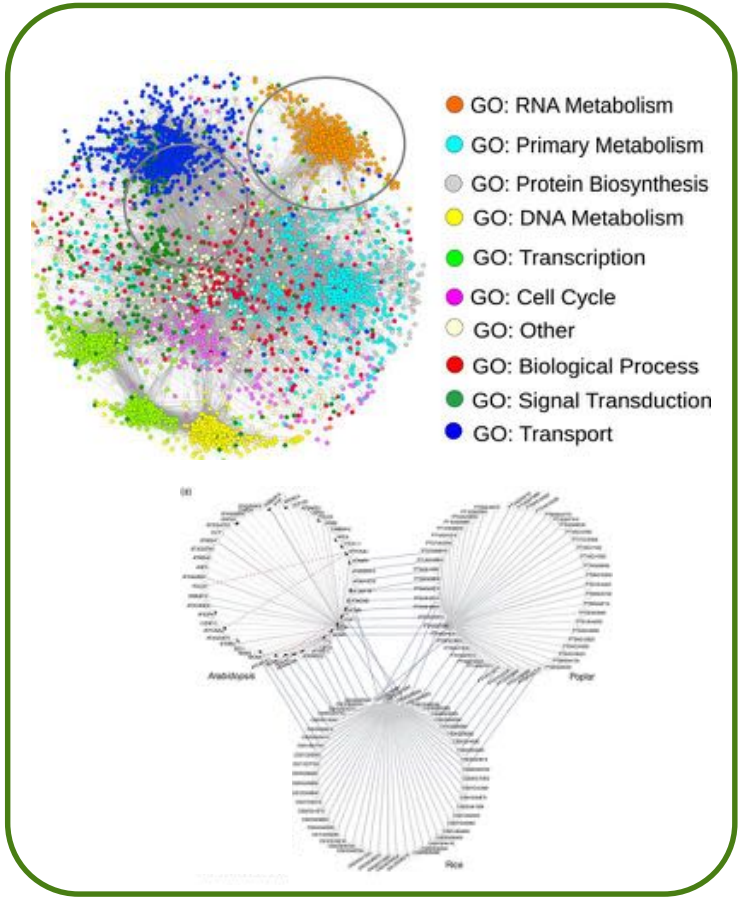


Deep/Machine Learning

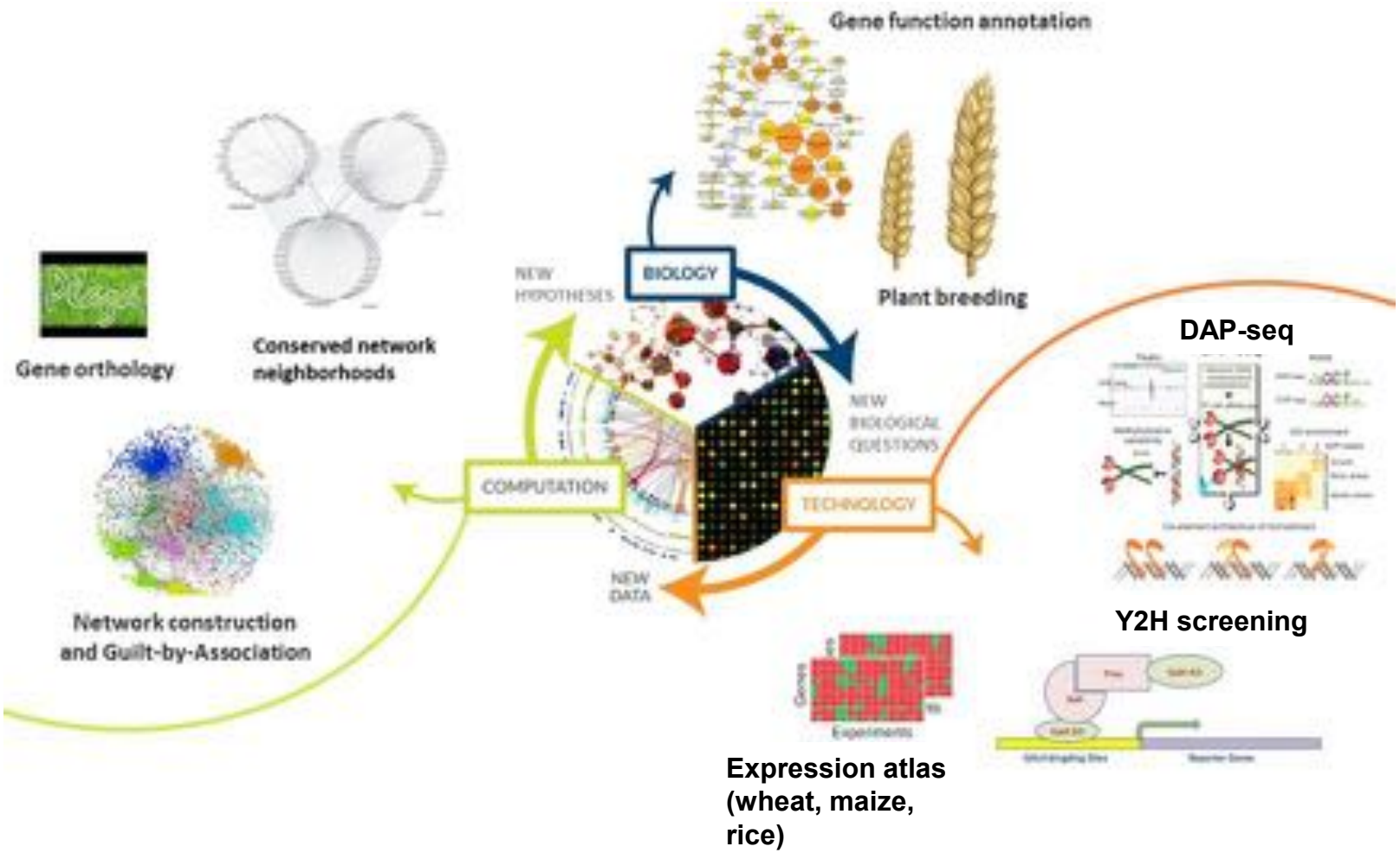


Integrative network analysis – project setup

- 3 year project funded by Flemish government
- Academic-industrial collaboration to combine novel wheat datasets and technologies/methodologies
- Require both wet-lab and bioinformatics expertise
  - Bioinformatics Postdoc performs integrative network analysis (Klaas Vandepoele @ VIB PSB)
  - Technician supports DAP-seq and Y2H experiments (Phillipa Borril's lab @ Birmingham Univ.)



Network biology project overview



- Cross-species co-expression and gene regulatory network analysis
- Generate wheat specific protein-protein (Y2H) and protein-DNA (DAP-seq) interaction datasets on selected TFs and regulators
- Integrative network analysis for functional annotation in wheat
  - Exploit gene-gene interactions
  - Gene function discovery
  - Gene prioritization for trait of interest

BASF
We create chemistry

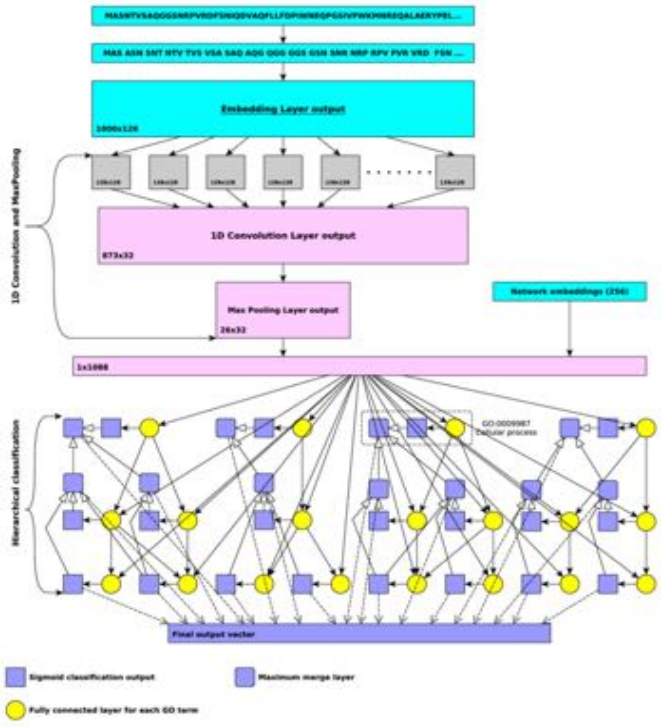
Deep learning based GO prediction

- Deep learning based GO prediction using
  - Protein sequences
  - Protein-protein interaction
  - Regulatory sequences

- Technology proof-of-concept
  - For protein classification problem
  - Currently on public dataset, to be tested in wheat

- Academic-industrial collaboration with Wesley De Neve's group @ Ghent Univ.
  - 1.5 year project
  - Jasper Zuallaert, PhD student
  - Xiaoyong Pan, Postdoc

Deep Learning



Challenges for DL-based GO prediction

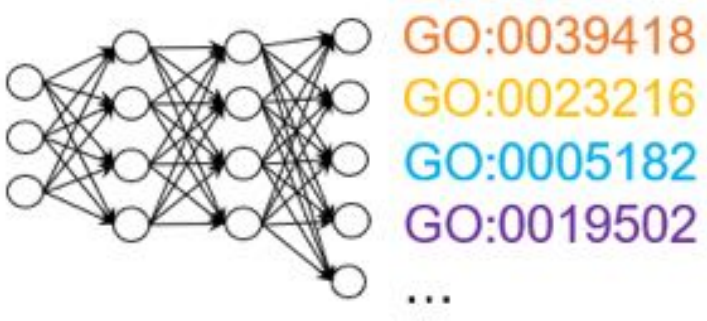
(1) variable input length

MLALFKFKTUJETJLLALCAHUE
MKJLAIEIIWACALCLAHT
MKEIPTIAXYYENTJAETHULACJLIAL
MTEIMZLLELALCJTL

(2) amino acid sequences → sparse input vector

M	L	A	L	F	K	F	K	T	Q
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	1	0	0
0	1	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

(3) hierarchical multi-label classification



BASF
We create chemistry

Improved DL architecture to tackle the challenges

(1) variable input length

Generate fixed-length input vectors:
- (basic) zero-padding
- Gated Recurrent Units (GRUs)
- dynamic length max pooling
- *K*-max pooling

(2) amino acid sequences → sparse input vector

Encoding strategies:
- (basic) one-hot encoding
- pre-trained embeddings
- ad hoc trainable embeddings

(3) hierarchical multi-label classification

Output strategies:
- (basic) non-hierarchical subnetwork for each term
- hierarchically structured subnetworks

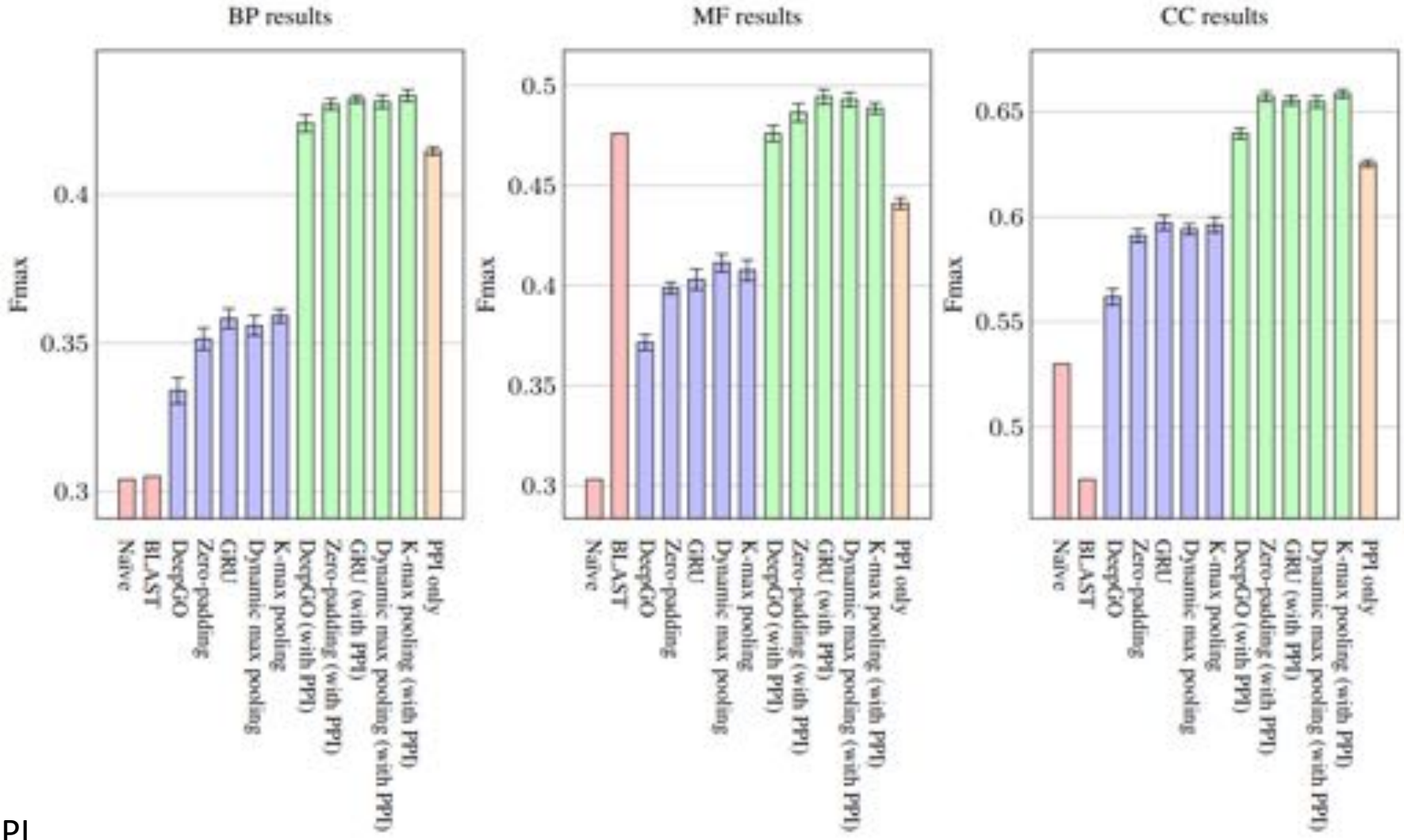
(4) external factors

Use of extra data sources, next to sequence information:
- `STRING`: protein-protein interaction (PPI) information
- `EggNOG`: orthology network information

starting point: DeepGO (developed @ KAUST)
http://deepgo.bio2vec.net/deepgo/

Performance of DL based GO prediction

- DL based GO prediction outperforms BLAST and native approach, expect for MF
- PPI adds significant values
- Better results using improved DL network architecture



Blue: DL using protein sequences
Green: DL using protein sequences + PPI
BP = Biological Process; MF = Molecular Function; CC = Cellular Components

Summary and future work

- Integrative network analysis for gene discovery in wheat
  - Cross-speices co-expression network analysis
  - Wheat specific protein-protein and gene regulatory networks
  - Network based functional annotation
  - Trait-associated candidate gene identification and prioritization

- Deep learning for functional anntation in wheat
  - Technology proof-of-concept: outperform naive sequence homology apporach
  - Transferability of model trained from model (plant) speices to wheat
  - Beyond sequence – combining e.g. biological network and phenotypic datasets

- Additional collaboration partners and external funding to futher explore novel technologies and methodologies

Acknowledgements

VIB-PSB	Klaas Vandepoele Pasquale Luca Curci
Birmingham University	Philippa Borrill
Ghent University	Wesley De Neve Jasper Zuallaert Xiaoyong Pan
BASF Innovation Center	Mark Davey Ralf Schmidt Christophe Liseron-Monfils Denys Marushchak