Predicting Reaction Success Using A Transformer Model Pretrained on Reaction SMILES Data

November 2023

Eric Gilbert, PhD

Outline

- Motivation and challenges.
- Workflow for creating a fine-tuned BERT model for yield prediction.
- Benchmarking
- Multi-modal learning: experimental text and reaction SMILES.
- Use cases

Predicting Reaction Yield

Motivation

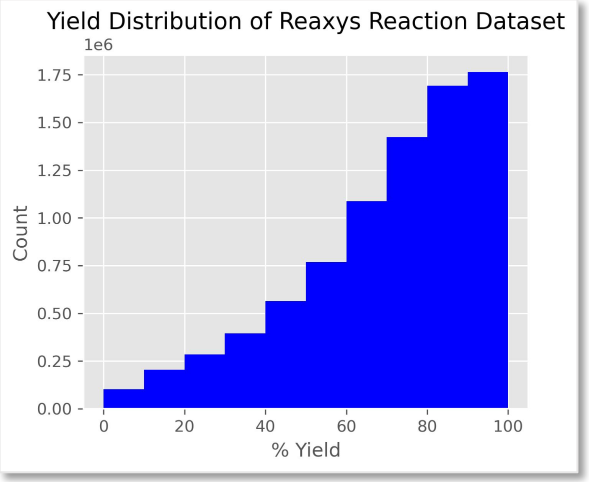
- ~20% of reactions fail or yield too low[1]
- wasteful- materials, human resources, opportunity cost, time

Challenges

- Literature and patent data biased toward higher yielding reactions.
- Models need to learn from failed reactions.

Strategy

- Pretrain a model from scratch using Reaxys reaction data.
- Fine-tune model on ELN data for predicting reaction success.
- Binary classification task- yield >5% or <5%



[1]Neves, P., McClure, K., Verhoeven, J. *et al.* Global reactivity models are impactful in industrial synthesis applications. *J Cheminform* **15**, 20 (2023). https://doi.org/10.1186/s13321-023-00685-0

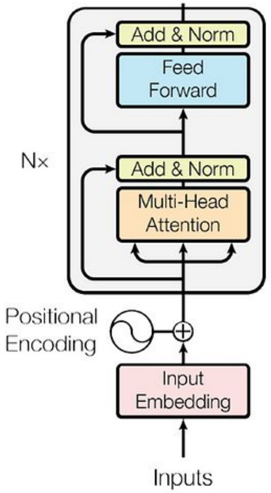
Workflow

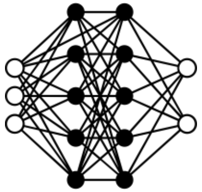
pre-training

preprocessed reaction SMILES from Reaxys

reaction_smiles
C1=CC=C(C=C1)NN=C2C(=O)C=CC3=C2C(=CC(=C3)S(=O)(=O)[O-])S(=O)(=O)[O-].[Na+].[Na+].Cl.CN(C)C=O.O=C...
c1c[nH]cn1.CCOC(=O)[C@H](C)O.CC(C)(C)[Si](C)(C)Cl.ClCCl>>CCOC(=O)[C@H](C)O[Si](C)(C)C(C)(C)C
FC(Cl)(Cl)Cl.FF.CC[C@H]1C[C@@H](CC(=O)OC)C1(C)C.ClC(Cl)Cl>>CC[C@]1(F)C[C@H](CC(=O)OC)C1(C)C
Cc1nc2c(Br)cc(Br)cc2c(=O)n1N.O=C(c1ccccc1)c1ccccc1.CCO.Cl>>Cc1nc2c(Br)cc(Br)cc2c(=O)n1N=C(c1cccc...
O=[N+]([O-])c1cccc(Cl)c1Cl.OCC(F)(F)F.CN(C)C=O.[HH].[NaH]>>FC(F)(F)COc1cccc(Cl)c1OCC(F)(F)F

- pretraining with ~9M reactions, 4 GPUs, 30 epochs, ~10 days





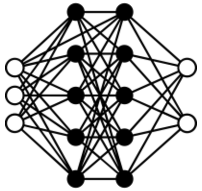
BERT Masked Language Model

fine tuning

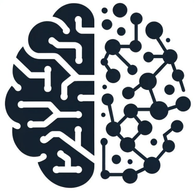
ELN training data*

Rxn smiles	Yield class
...CCBr.CC#N>>O=C1CC...	1
…(O)C(F)(F)>>c1cc2…	0
…cc(Cl)c1>>COC(=O)….	1

+

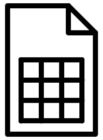


pretrained BERT MLM

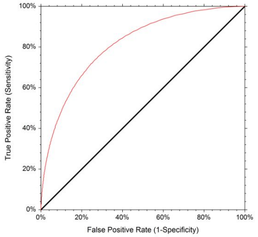


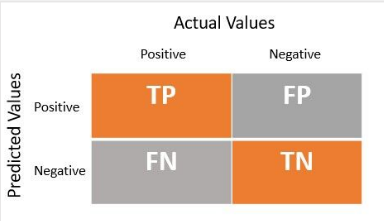
fine-tuned BERT MLM

inference



test set

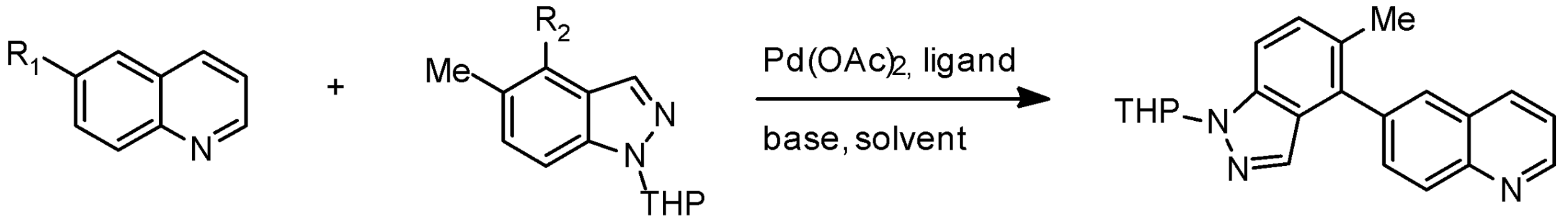




model performance metrics

*ELN data, model fine tuning, and inference on secure Amazon Workspaces hosted by Janssen.

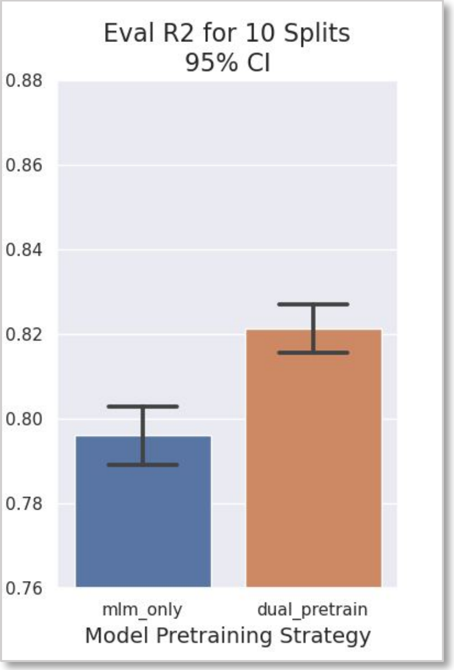
Suzuki Reaction Benchmark

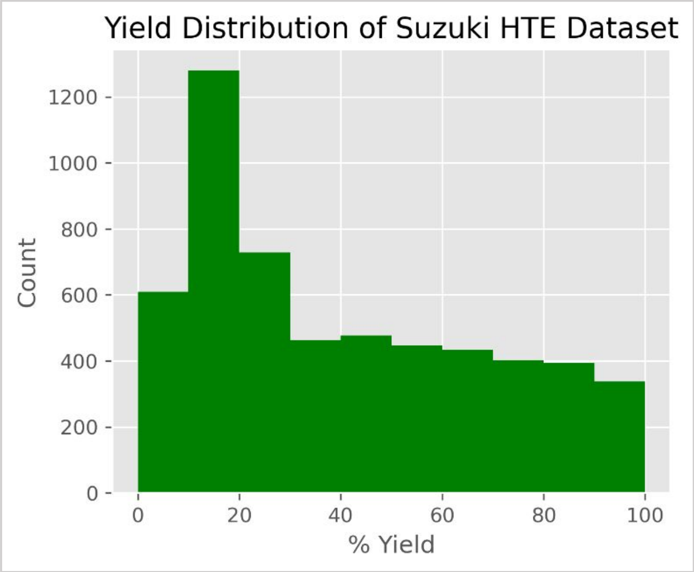


5760 reactions[1]

11 ligands
6 boronic acids
4 aryl halides
7 bases
4 solvents

Base model	R^2
MLM-only pretrain	0.80 ± 0.01
Dual pretrain	0.82 ± 0.01
[2]Schwaller rxnfp	0.79 ± 0.01



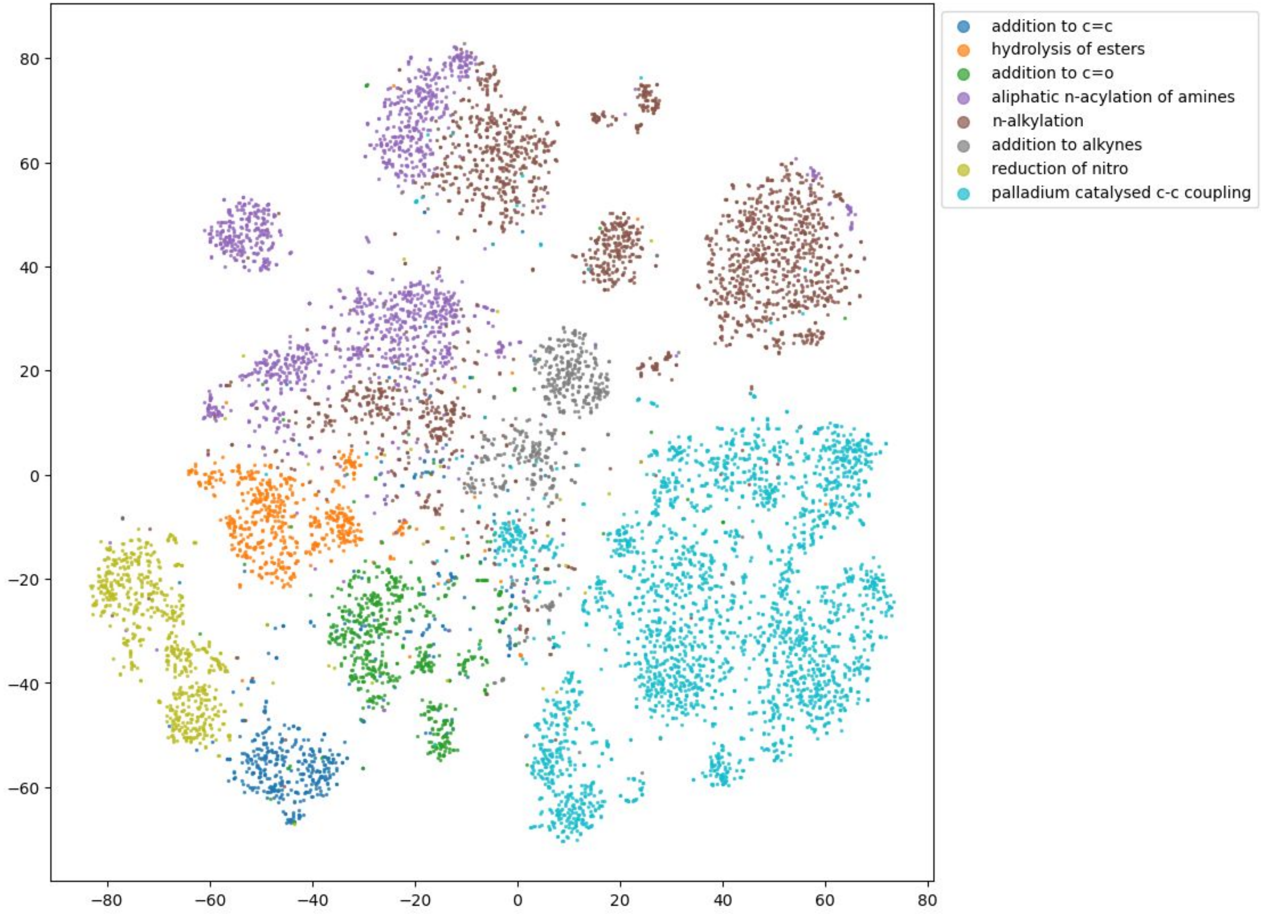


Dual pretraining leads to statistically significant improvement in model performance.

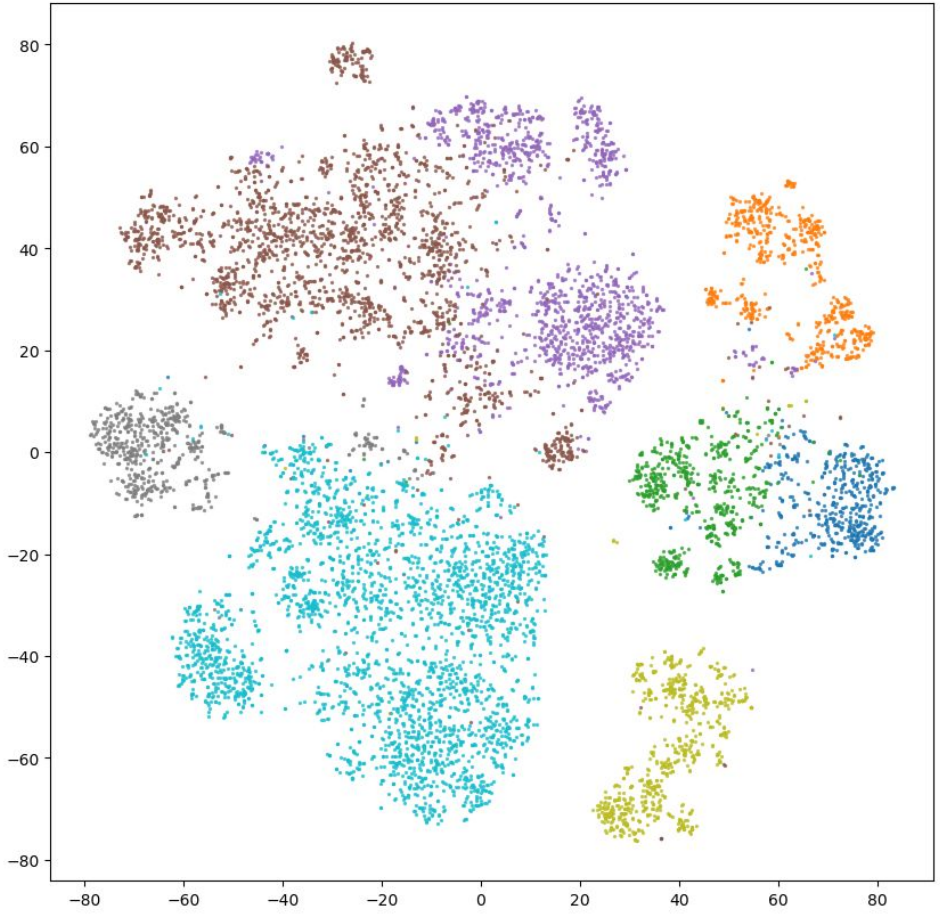
1) Perera et al., Science 359, 429–434 (2018)
2) Philippe Schwaller *et al* 2021 *Mach. Learn.: Sci. Technol.* **2** 015016

Comparison of Embedding Projections for Pretrained Models

MLM only pretrain



Dual pretrain

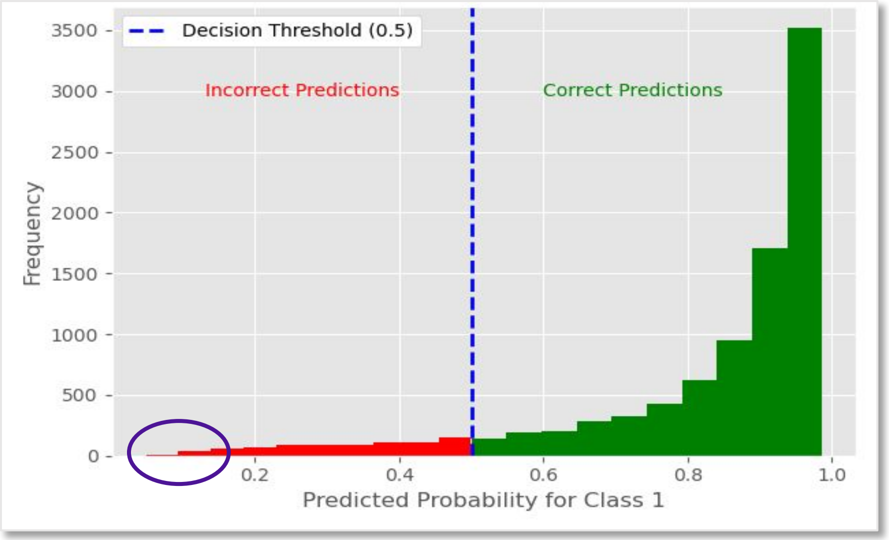


- dual pretrained model shows improved clustering of reaction classes

ELSEVIER

Use Cases – Data Quality

- Data quality insights from test set inference.
  - can be instructive to look at what the model got ‘most wrong’
  - helpful if slice data by reaction class when interrogating



- missing reagents?
- missing catalysts?
- erroneous reactants?
- etc.

- Data quality issues may not be obvious when looking at model metrics by class.
  - models are surprisingly resilient at learning from messy data.
- Use insights to inform anomaly detection in training data embedding clusters.

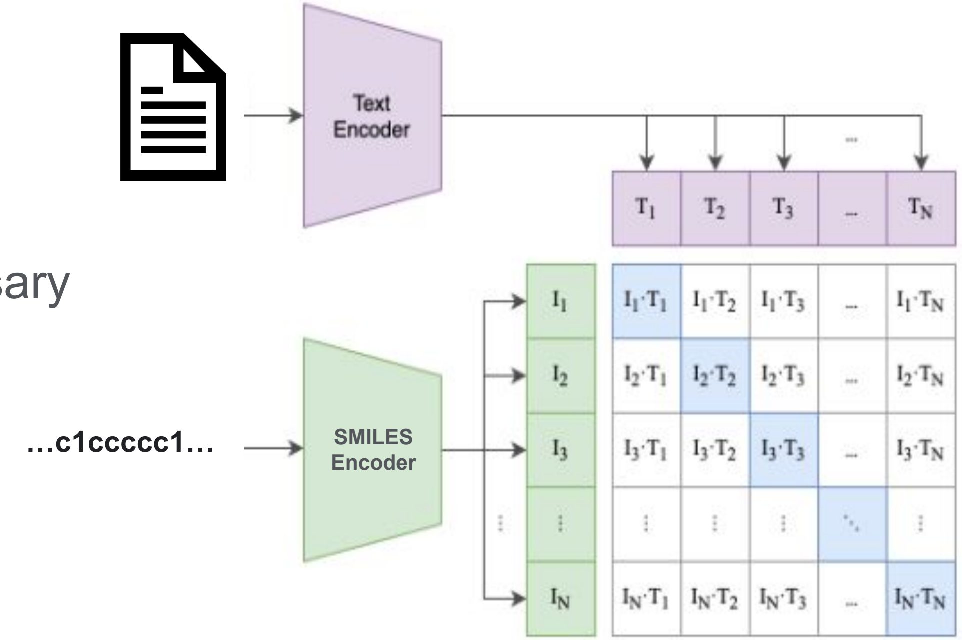
ELSEVIER

Use Cases - Synthesis

- Focus high throughput experimentation (HTE) efforts.
  - can create a combinatorial combination of potential reagents
  - rank order probabilities of reaction success
- Aid medicinal chemists on more focused synthetic queries:
  - which solvent is predicted to be best for this transformation?
  - rank order potential targets based on predicted probability of success
- Incorporate into multi-modal model.

Multi-modal Deep Learning

- Contrastive learning on procedure text and reaction SMILES
  - ‘foundation model’
- Text features are associated with reaction SMILES
  - domain adaptation between literature and patent text necessary
- Enables new applications:
  - Zero / Few-shot learning
  - Cross-modality search
- Example of association of text and reaction SMILES:
  - low temperature in procedure correctly associated with solvent in SMILES:



Attribution Label	Attribution Score	Word Importance
heating	-4.02	[CLS] C C (C) (C) O C (= O) N 1 C C N (C 2 C C C 3 (c 4 c c c 5 c (c 4) O C O 5) C C 3 C 2) C C 1 . Cl C Cl . O = C (O) C (F) (F) F >> c 1 c c 2 c (c c 1 C 1 3 C C C (N 4 C C N C C 4) C C 1 C 3) O C O 2 [EOS]
heating	-2.77	[CLS] C O C (= O) C S c 1 n n n (- c 2 c c c c c 2) n 1 . Cl C Cl . O = C (O O) c 1 c c c c (Cl) c 1 >> C O C (= O) C S (= O) c 1 n n n (- c 2 c c c c c 2) n 1 [EOS]
heating	-0.84	[CLS] C N (C) C (n 1 n [n+] ([O-]) c 2 n c c c c 2 1) = [N+] (C) C . C S (= O) (= O) N 1 C C N C C 1 . Cl C Cl . F [P] (F) (F) (F) (F) F . O = C (O) c 1 c c c (- n 2 n c (- c 3 c c c c c 3) c c 2 O) c c 1 >> C S (= O) (= O) N 1 C C N (C (= O) c 2 c c c (N 3 N = C (c 4 c c c c c 4) C C 3 = O) c c 2) C C 1 [EOS]

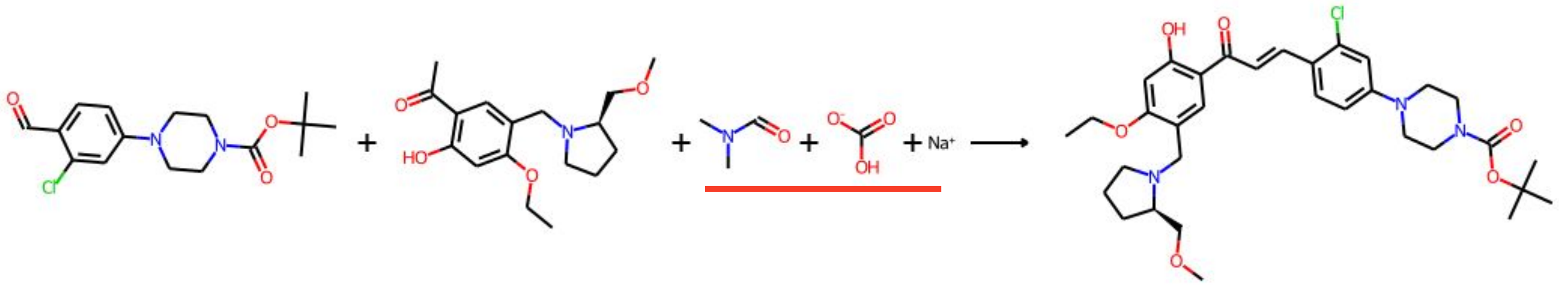
ELSEVIER

Use Cases – Data Quality

- Example from Reaxys with data inconsistency:

  Procedure text:

  ‘Add 303mg B-122 to 4ml EtOH, then add 221mg KOH, 640mg B-10, stir at RT, and monitor LC-MS until there is no B-122 left.’



- Inconsistency identified using cosine similarity between vector embeddings of text and SMILES.
- Potential application:
  - Prevent errors in ELN data entry improve data quality reproducibility, better AI models.

Use Cases – Search

- SMILES-to-Text search

  - can we suggest conditions / workup from existing procedures using reaction SMILES as input?
  - use similarity between vector embeddings of text and SMILES
  - Advantage: Suggest applicable existing procedures rather than black box prediction of conditions.

Summary

- Pretrained a BERT Masked Language Model from scratch using Reaxys reaction SMILES.
  - investigated impact of adding reaction classification task to pretraining
- Fine tuned models on Janssen ELN data to predict reaction success (>5% yield).
- Benchmarked MLM-only & dual-pretrained models on Suzuki benchmark.
  - use cases- data quality, anomaly detection & synthesis.
- Multi-modal model- contrastive training with reaction SMILES & procedure text.
  - use cases- data quality / consistency, SMILES-to-Text search
- Demonstrated using proprietary ELN data hosted on server by pharma partner.
- Pretrained model can readily be used with other companies and their ELN data.

Acknowledgements

ELSEVIER

- Frederik van den Broek
- Kathleen Maffei
- David Wöhlert
- Kinga Szarkowska
- Timur Madzhidov
- Ralph Hössel
- Thomas Böttjer
- Matthew Clark
- Tugrul Kaynak

janssen | PHARMACEUTICAL COMPANIES OF Johnson & Johnson

Janssen

- Jörg Wegner
- Paulo Neves
- Ramil Nugmanov
- Jonas Verhoeven
- Kostiantyn Chernichenko