S+pKa Predictor of Ionization Constants – Recent Progress and Results

Robert Fraczkiewicz, PhD

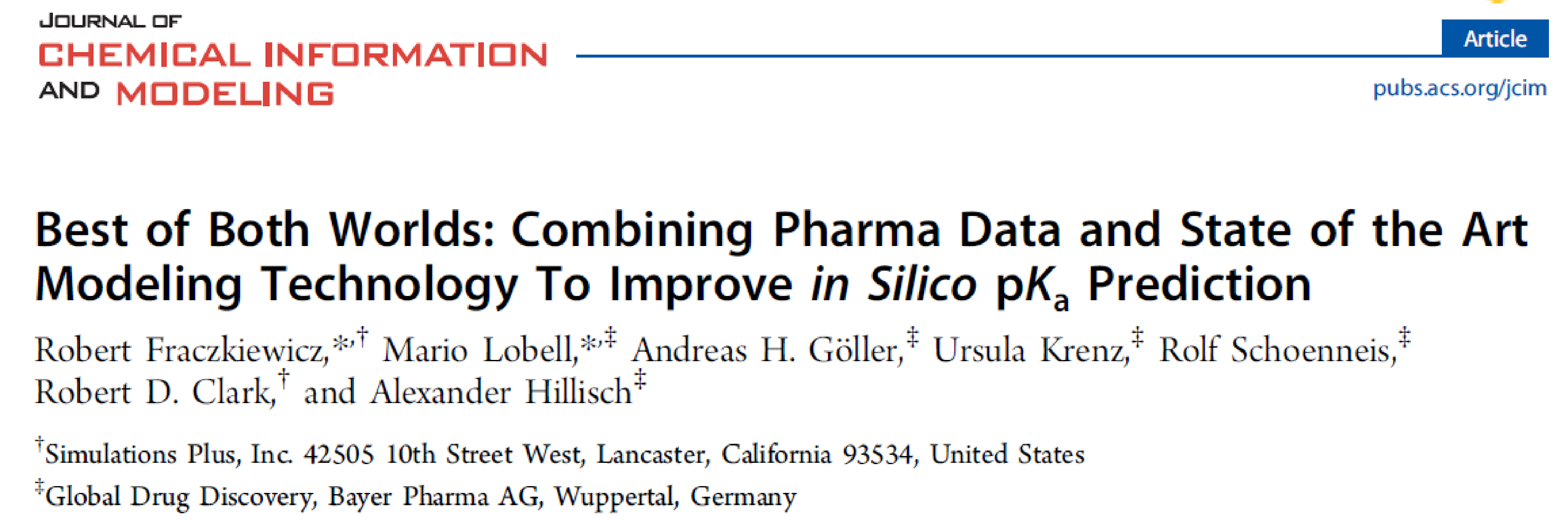
Research Fellow

• REC

Please note: this presentation, including questions from the audience, is being recorded and may be made available.

A bit of history

- Until 2012 the S+pKa model was exclusively trained on ~11,000 compounds from published literature. This model will be labeled as "v 6.0".
- In 2012 Bayer Pharma AG had shared with us an additional set of ~16,000 compounds with measured pK_a. The resulting "v 7.0" model was trained on combined data and its prediction results were published in 2015.



JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve *in Silico* pK_a Prediction

Robert Fraczkiewicz,*[†] Mario Lobell,*[‡] Andreas H. Göller,[‡] Ursula Krenz,[‡] Rolf Schoenneis,[‡] Robert D. Clark,[†] and Alexander Hillisch[‡]

[†] Simulations Plus, Inc. 42505 10th Street West, Lancaster, California 93534, United States

[‡] Global Drug Discovery, Bayer Pharma AG, Wuppertal, Germany

Fraczkiewicz, R., et al. (2015). Journal of Chemical Information and Modeling **55**(2): 389-397.

A bit of history

- S+pKa "v 7.0" has shown dramatic improvements in prediction quality as evaluated in the *Bayer chemical space*. All test sets were external.

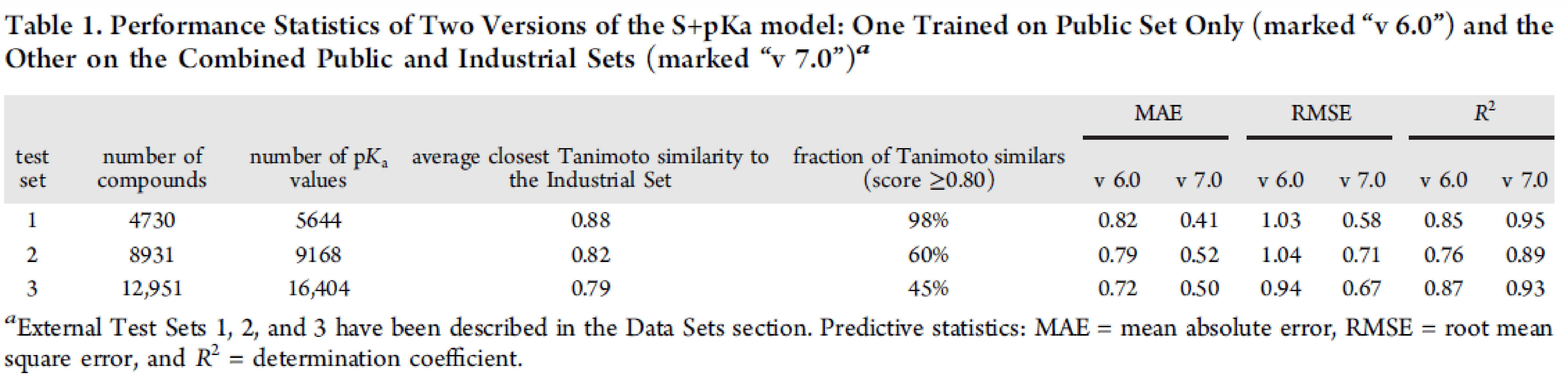


Table 1. Performance Statistics of Two Versions of the S+pKa model: One Trained on Public Set Only (marked "v 6.0") and the Other on the Combined Public and Industrial Sets (marked "v 7.0")[a]

					MAE		RMSE		R^2	
test set	number of compounds	number of pK_a values	average closest Tanimoto similarity to the Industrial Set	fraction of Tanimoto similars (score ≥0.80)	v 6.0	v 7.0	v 6.0	v 7.0	v 6.0	v 7.0
1	4730	5644	0.88	98%	0.82	0.41	1.03	0.58	0.85	0.95
2	8931	9168	0.82	60%	0.79	0.52	1.04	0.71	0.76	0.89
3	12,951	16,404	0.79	45%	0.72	0.50	0.94	0.67	0.87	0.93

[a] External Test Sets 1, 2, and 3 have been described in the Data Sets section. Predictive statistics: MAE = mean absolute error, RMSE = root mean square error, and R^2 = determination coefficient.

A bit of history

- It outperformed competiton, too.

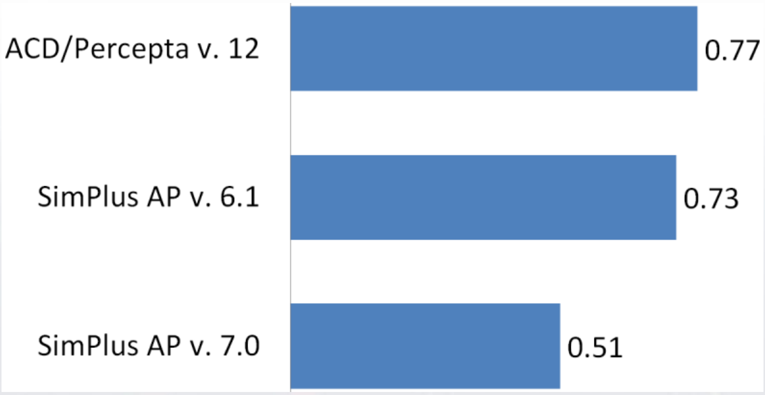
ACD/Percepta v. 12 and ADMET Predictor™ v 6.1 show comparable pK_a prediction accuracy

ADMET Predictor™ v 7.0 (after retraining with BTr) shows significantly improved pK_a prediction accuracy

Prediction statistics for 981-compound Bayer test set with 981 exp. pK_a values (subset of newest measurements on 12951 Bayer compounds):

Predicted by	Trained with	MAE	RMSE	R^2
ACD/Percepta v 12	15932 lit pK_a	0.77	1.05	0.84
ADMET Predictor v 6.1	14147 lit pK_a	0.73	0.95	0.86
ADMET Predictor v 7.0	14149 lit pK_a + 19467 Bayer pK_a	0.51	0.67	0.93

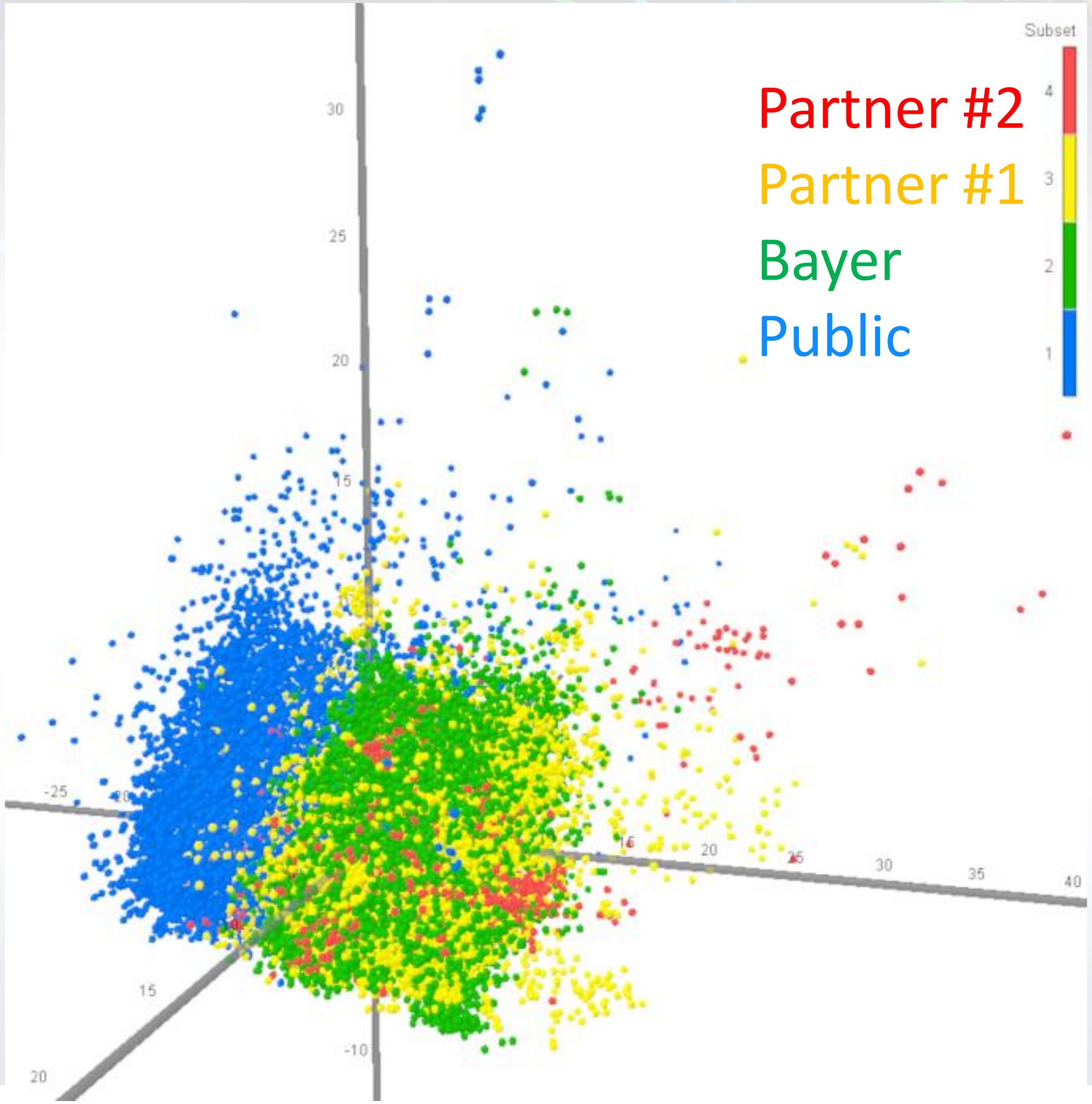
MAE:

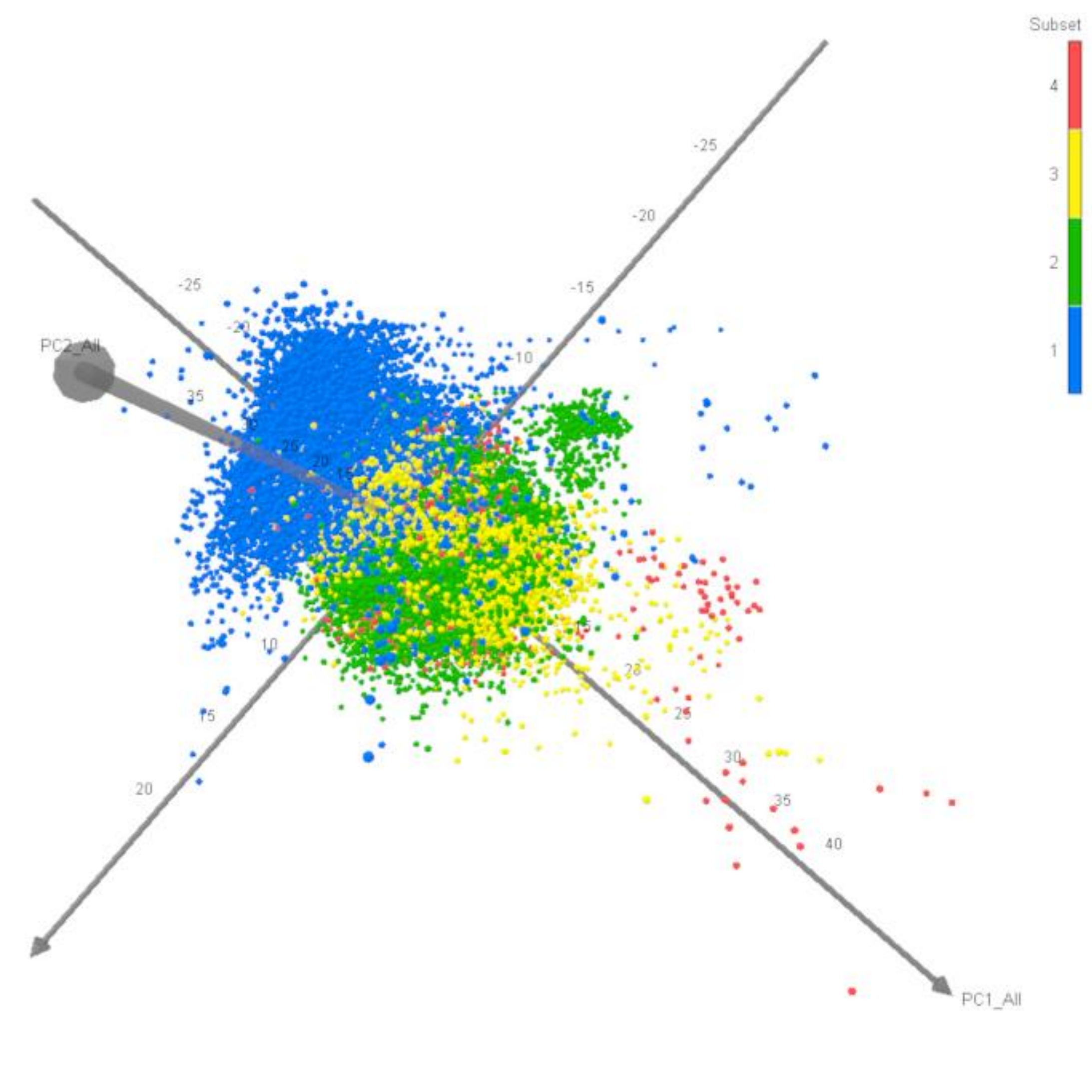


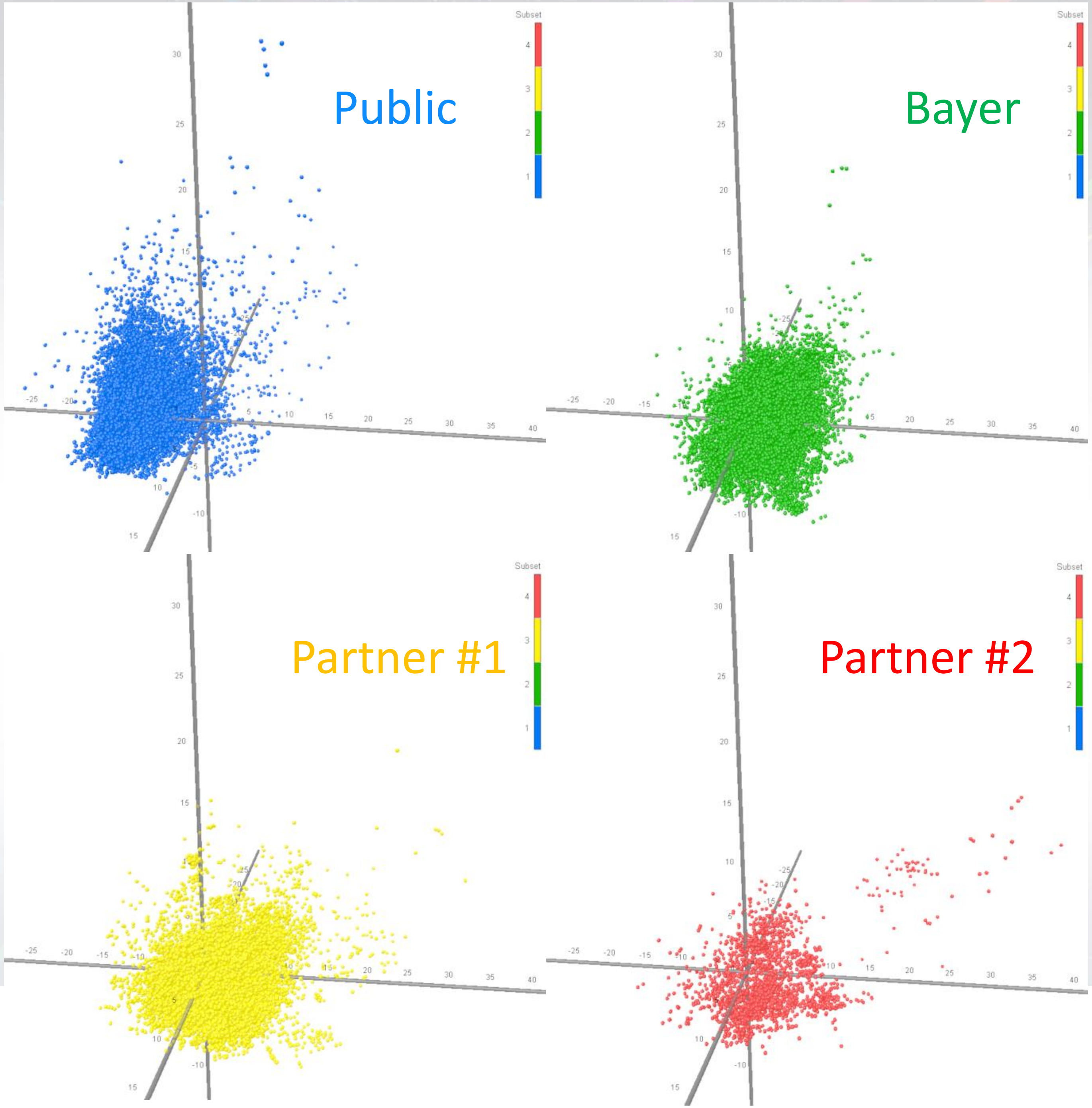
2022

- Two new industrial partners (large pharmaceutical companies; further labeled as "Partner #1" and "Partner #2") have indicated inadequate coverage of their chemical space by the "v 7.0" S+pKa.
- Instead of complaining both partners have shared with us significant amount of experimental pK_a data extracted from their corporate databases.
  - Partner #1 has provided ~19,000 compounds
  - Partner #2 has provided ~2,400 compounds
- From August 2022 until November 2022 we were busy rebuilding the S+pKa model with the new data appended to public+Bayer set. The resulting newest version carries the "v 10.5" label.

Chemical space projected on the first 3 principal components of the ADMET Predictor molecular descriptors matrix



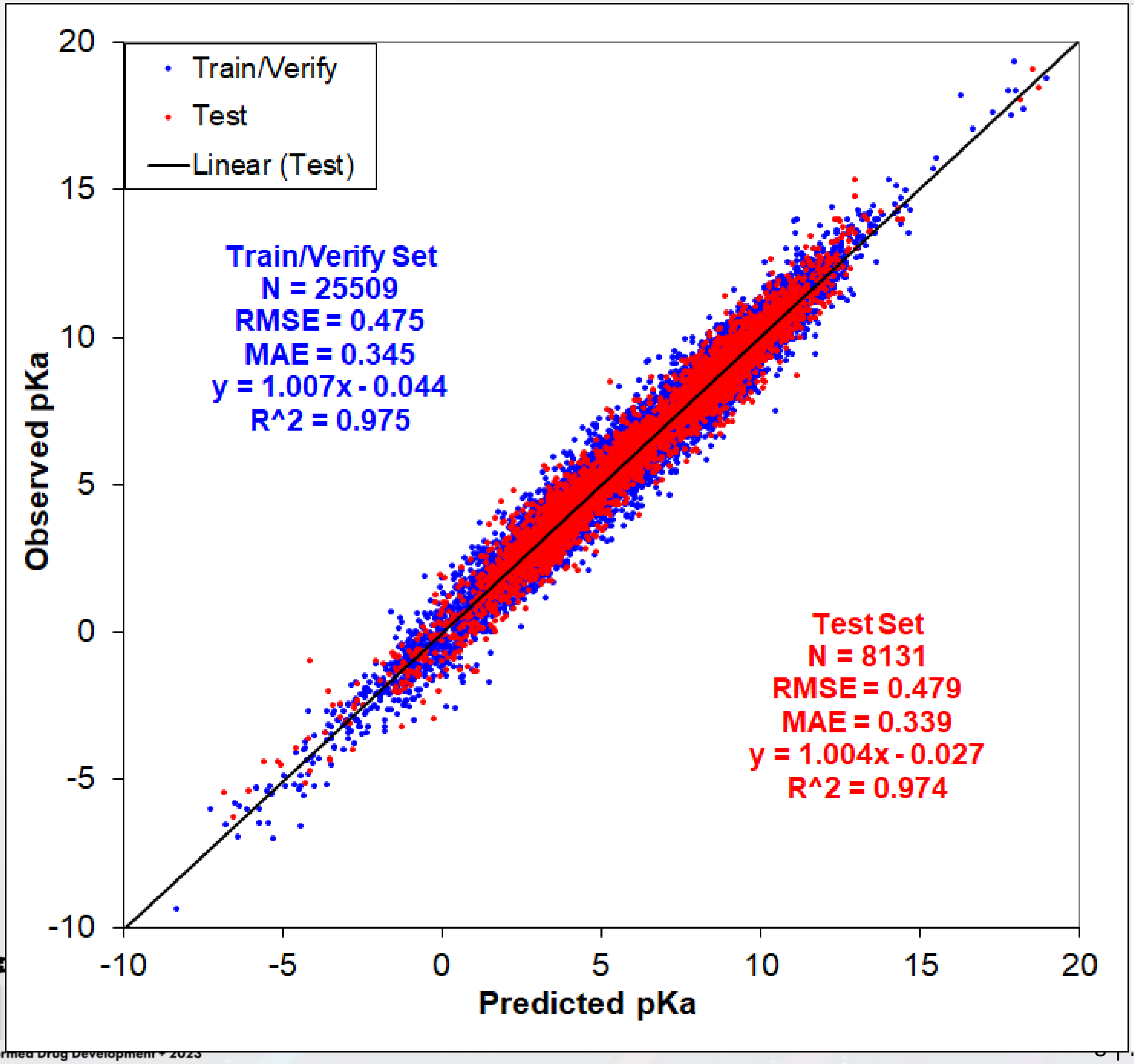



Public
Bayer
Partner #1
Partner #2
Subset
4
3
2
1
30
25
20
15
10
5
-10
-25
-20
-15
10
15
20
25
30
35
40

"v 7.0" vs. "v 10.5" performance

- It’s apples vs. oranges, but the relative improvement is welcome

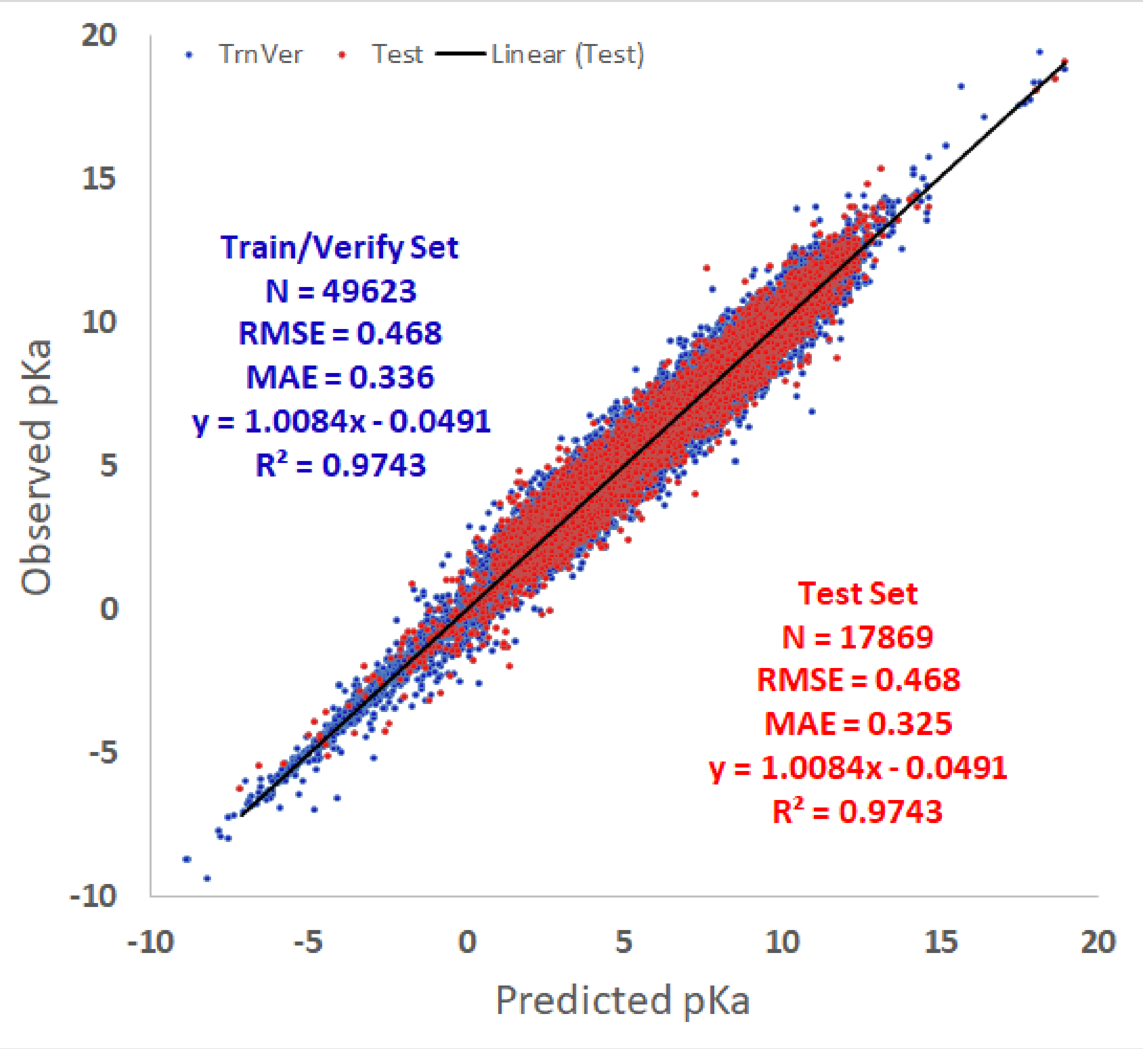
v 7.0



Train/Verify Set
N = 25509
RMSE = 0.475
MAE = 0.345
y = 1.007x - 0.044
R^2 = 0.975

Test Set
N = 8131
RMSE = 0.479
MAE = 0.339
y = 1.004x - 0.027
R^2 = 0.974

v 10.5

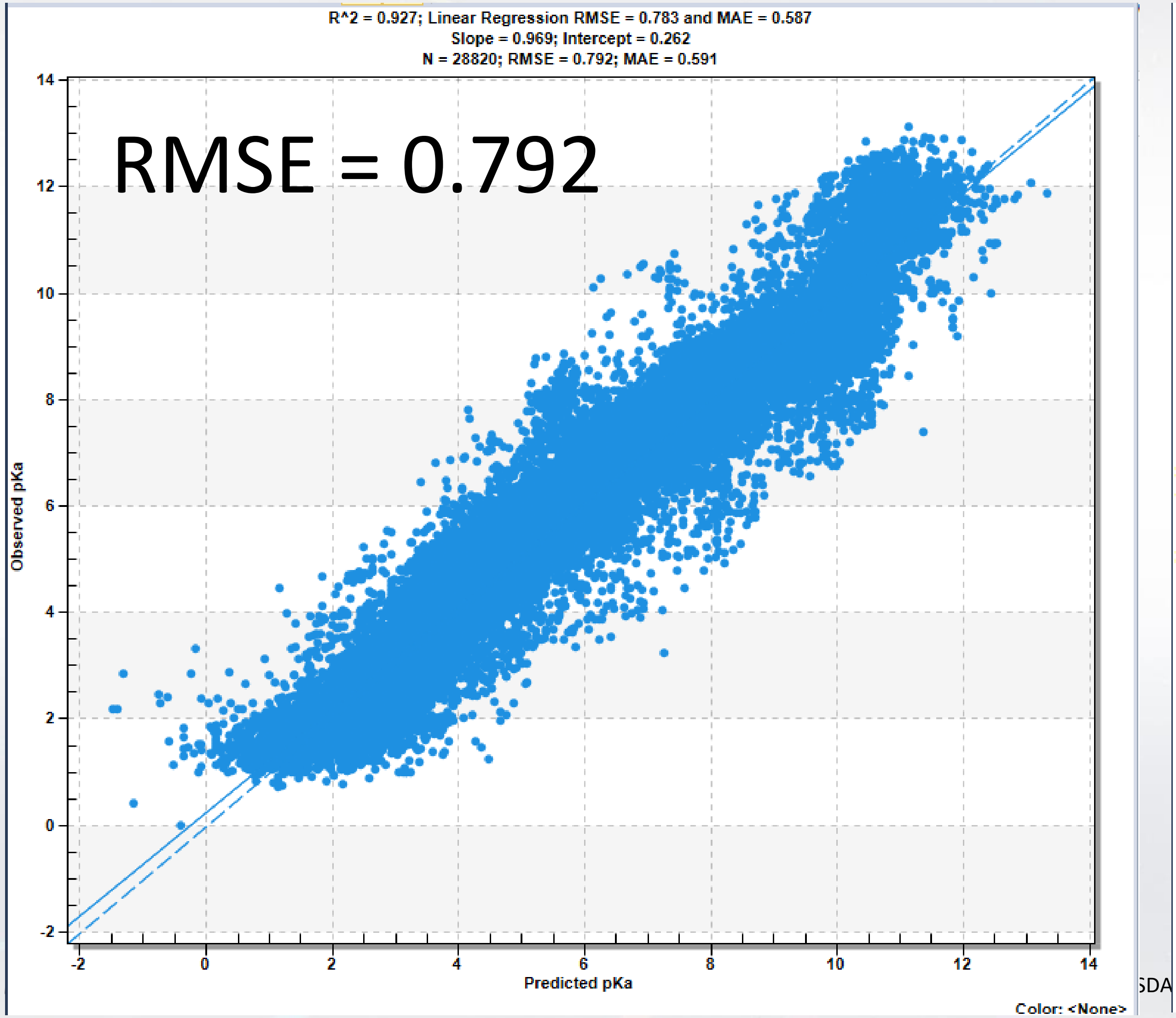


Train/Verify Set
N = 49623
RMSE = 0.468
MAE = 0.336
y = 1.0084x - 0.0491
R^2 = 0.9743

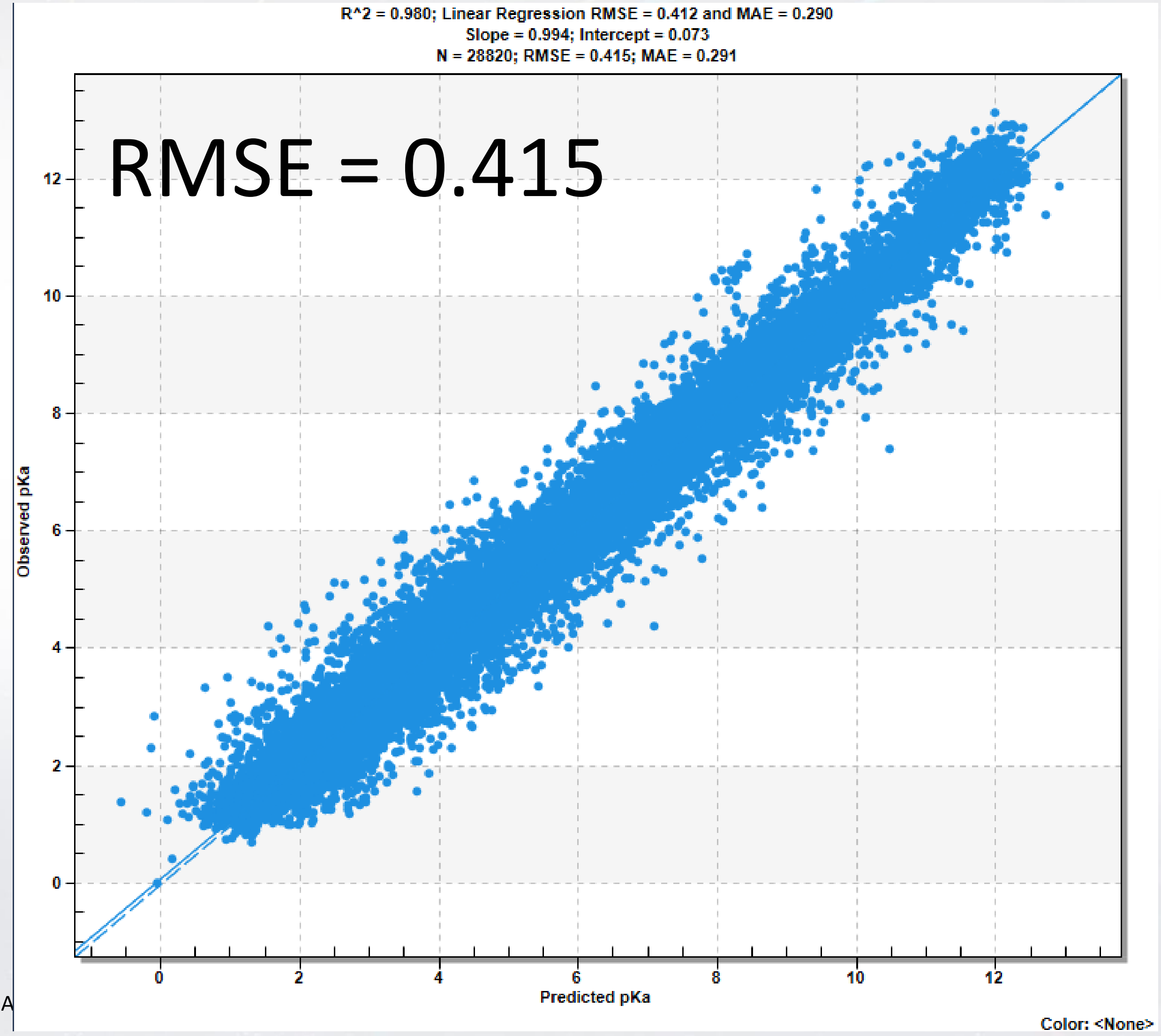
Test Set
N = 17869
RMSE = 0.468
MAE = 0.325
y = 1.0084x - 0.0491
R^2 = 0.9743

"v 7.0" and "v 10.5" models vs. received data

- Data from Partner #1. "v 7.0" RMSE = 0.792, "v 10.5" RMSE = 0.415



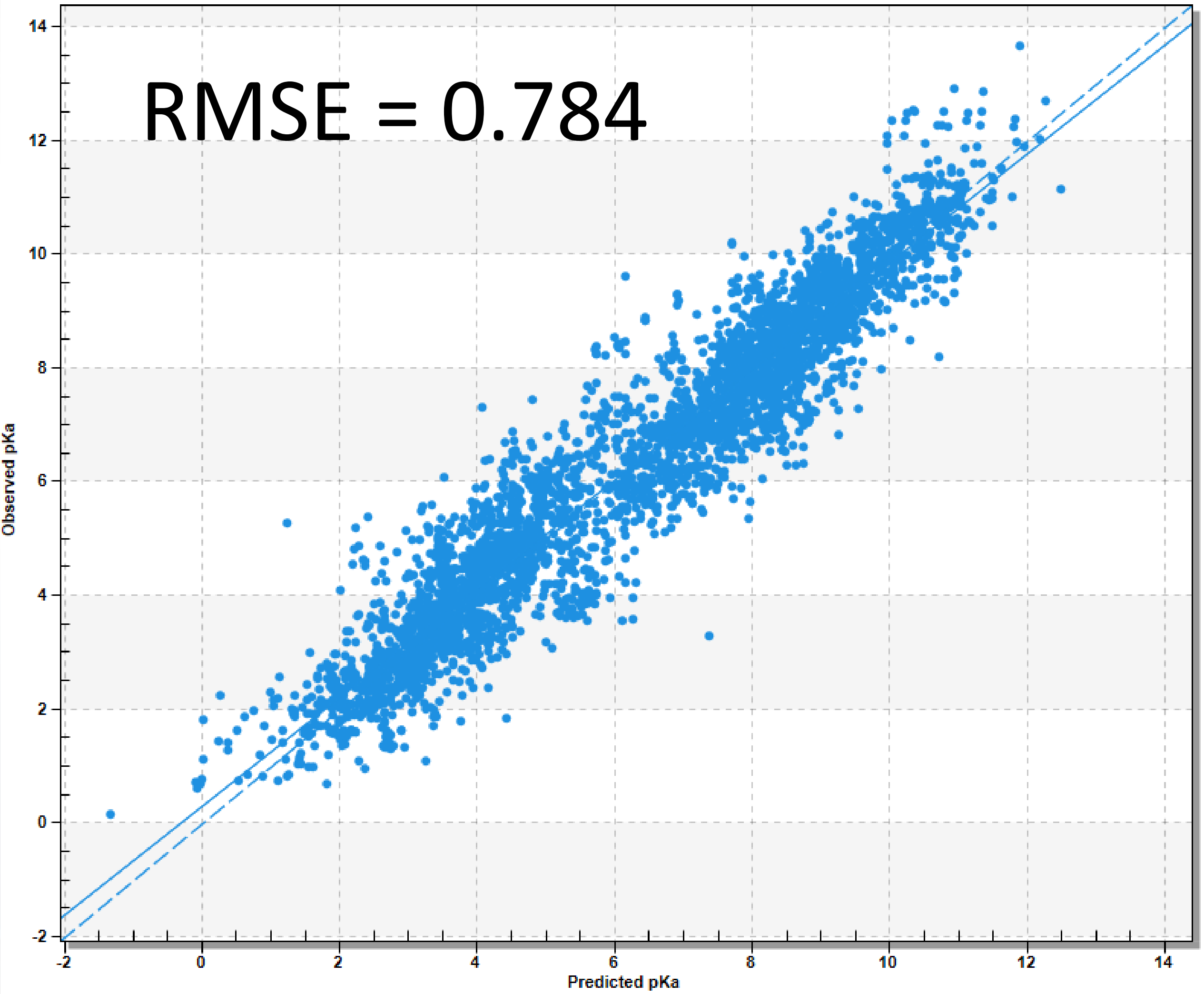
RMSE = 0.792



RMSE = 0.415

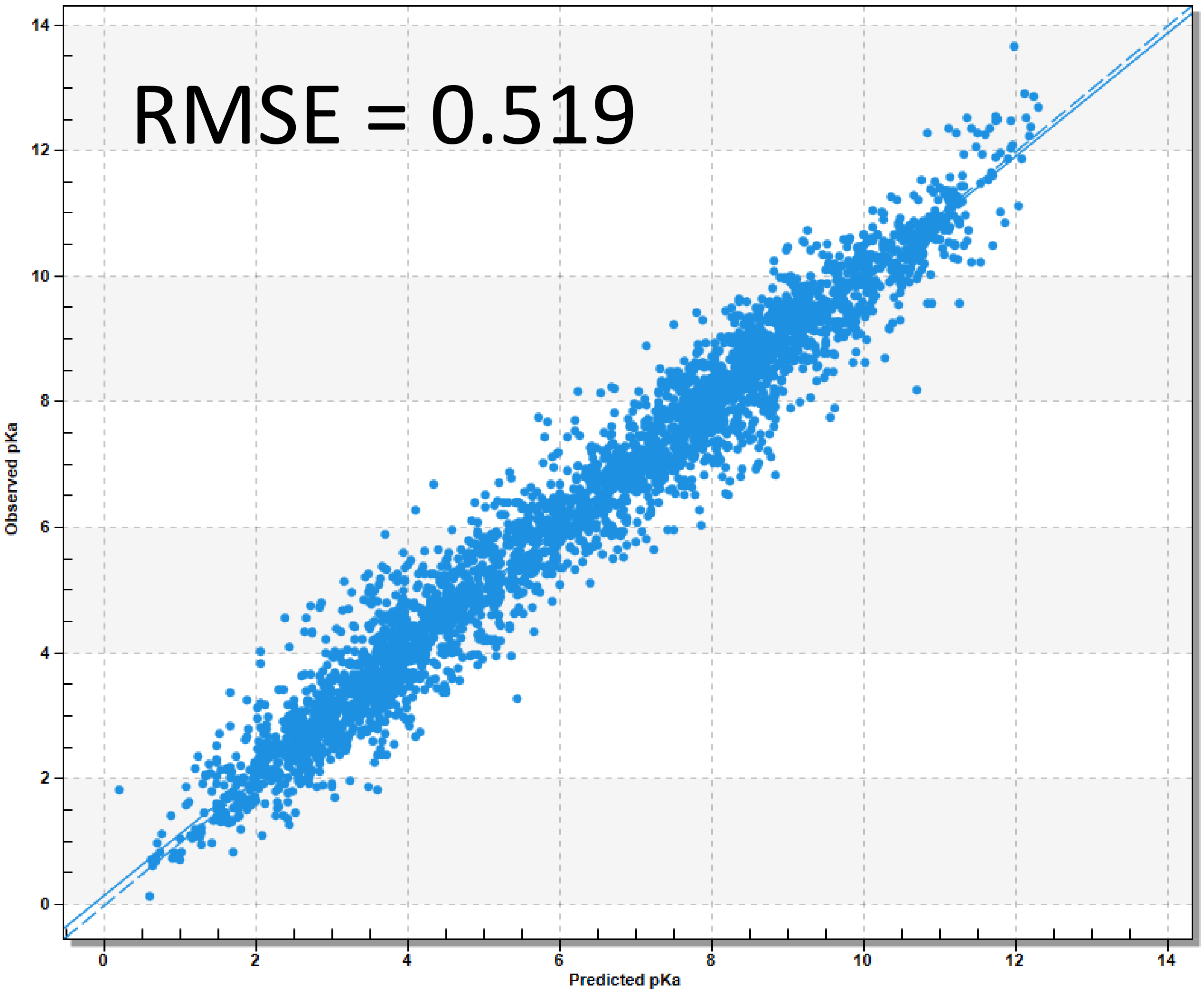
"v 7.0" and "v 10.5" models vs. received data

- Data from Partner #2. "v 7.0" RMSE = 0.784, "v 10.5" RMSE = 0.519



R^2 = 0.915; Linear Regression RMSE = 0.774 and MAE = 0.588
Slope = 0.955; Intercept = 0.309
N = 4034; RMSE = 0.784; MAE = 0.597

RMSE = 0.784

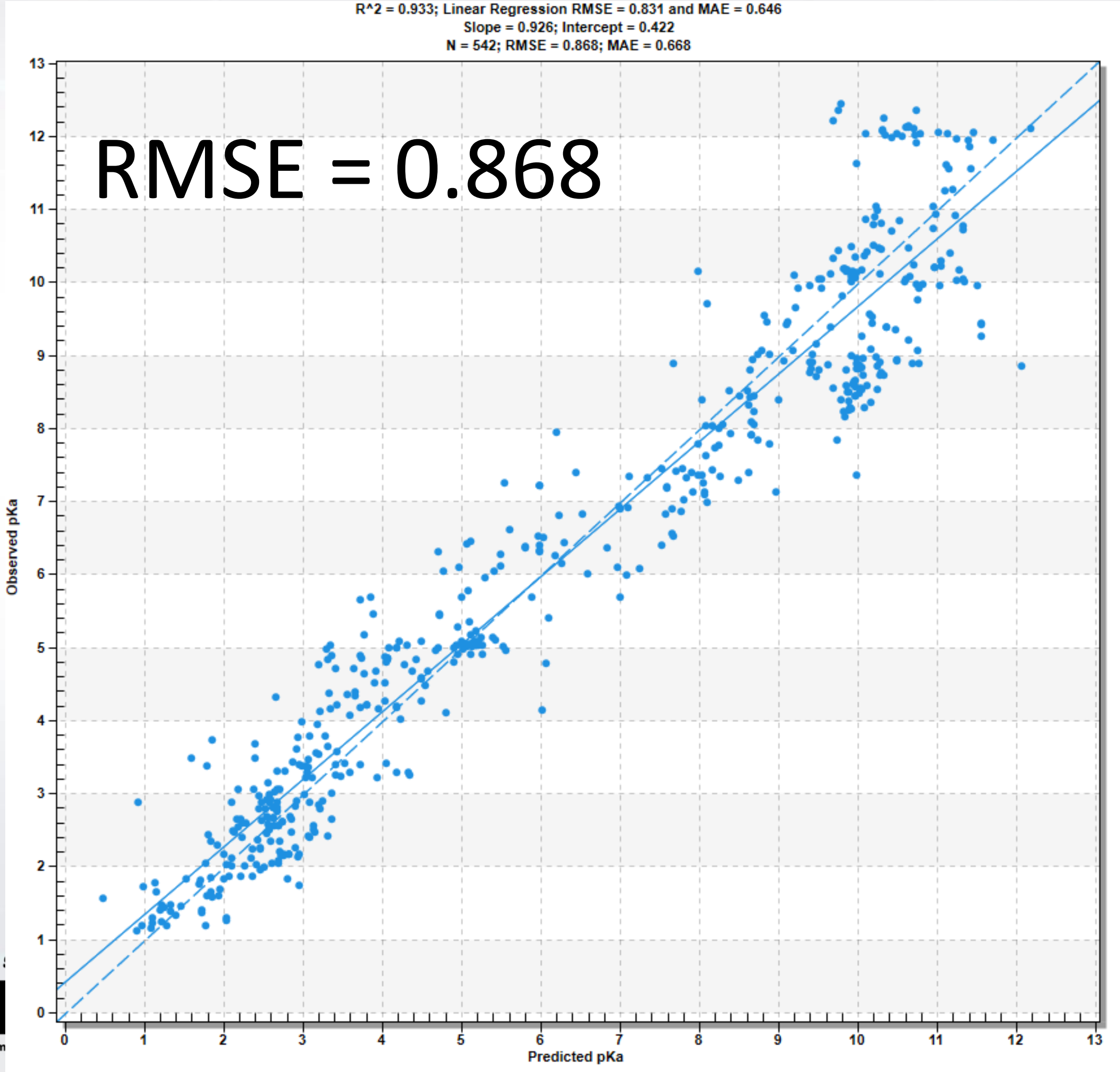


R^2 = 0.962; Linear Regression RMSE = 0.515 and MAE = 0.383
Slope = 0.981; Intercept = 0.152
N = 4034; RMSE = 0.519; MAE = 0.383

RMSE = 0.519

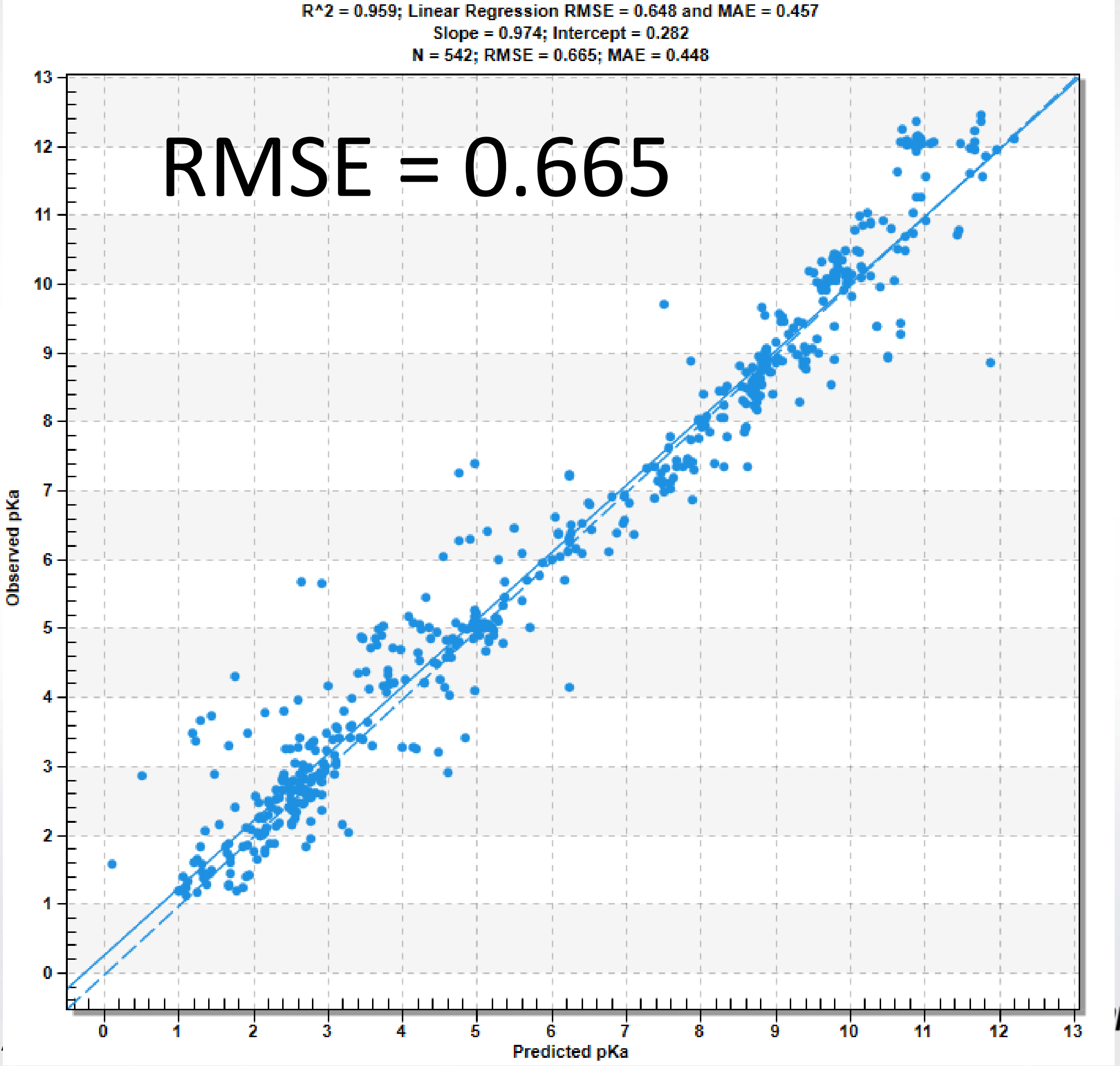
"v 7.0" and "v 10.5" models in external testing

- At Partner #1 site. "v 7.0" RMSE = 0.868, "v 10.5" RMSE = 0.665



R^2 = 0.933; Linear Regression RMSE = 0.831 and MAE = 0.646
Slope = 0.926; Intercept = 0.422
N = 542; RMSE = 0.868; MAE = 0.668

RMSE = 0.868

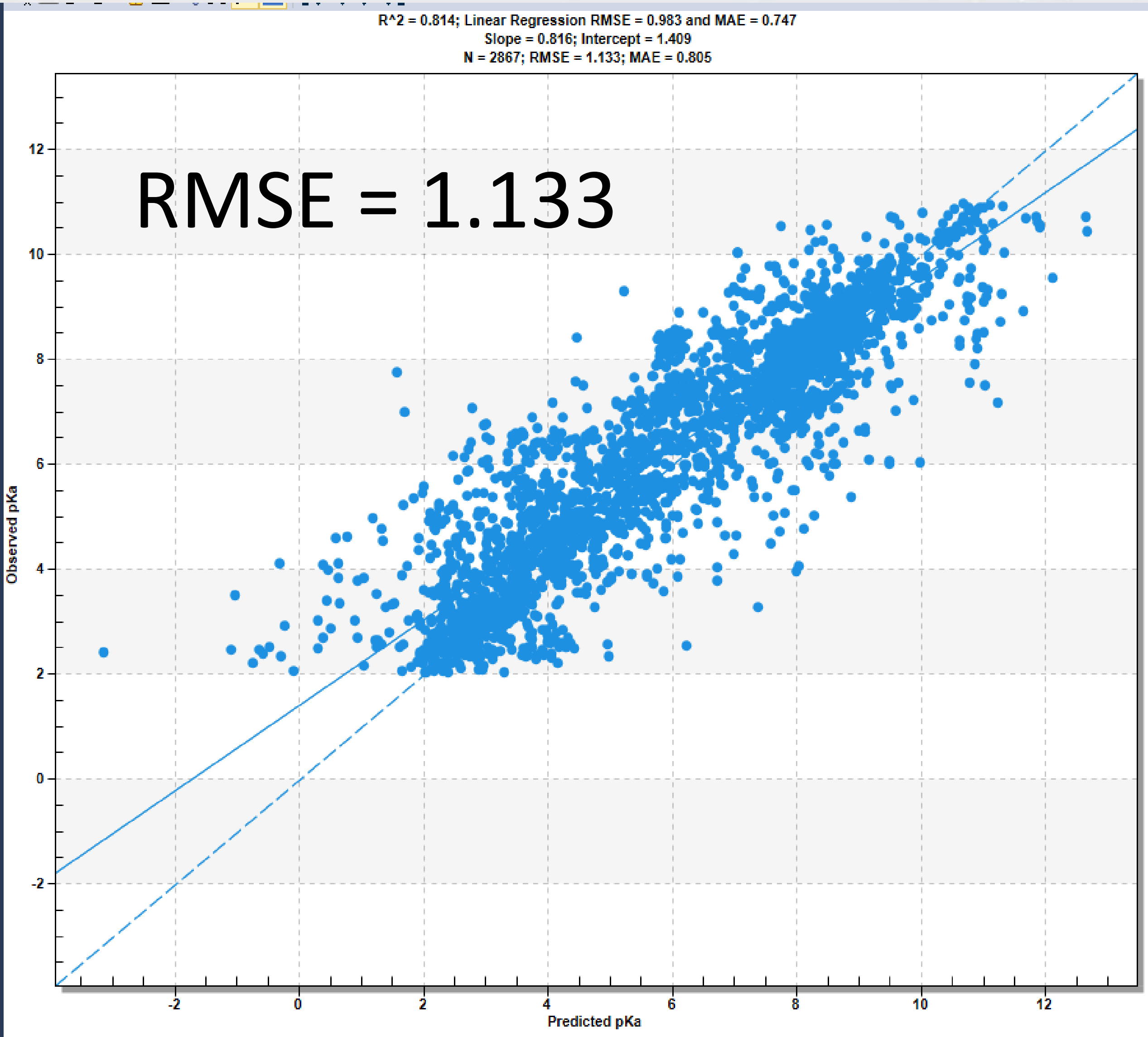


R^2 = 0.959; Linear Regression RMSE = 0.648 and MAE = 0.457
Slope = 0.974; Intercept = 0.282
N = 542; RMSE = 0.665; MAE = 0.448

RMSE = 0.665

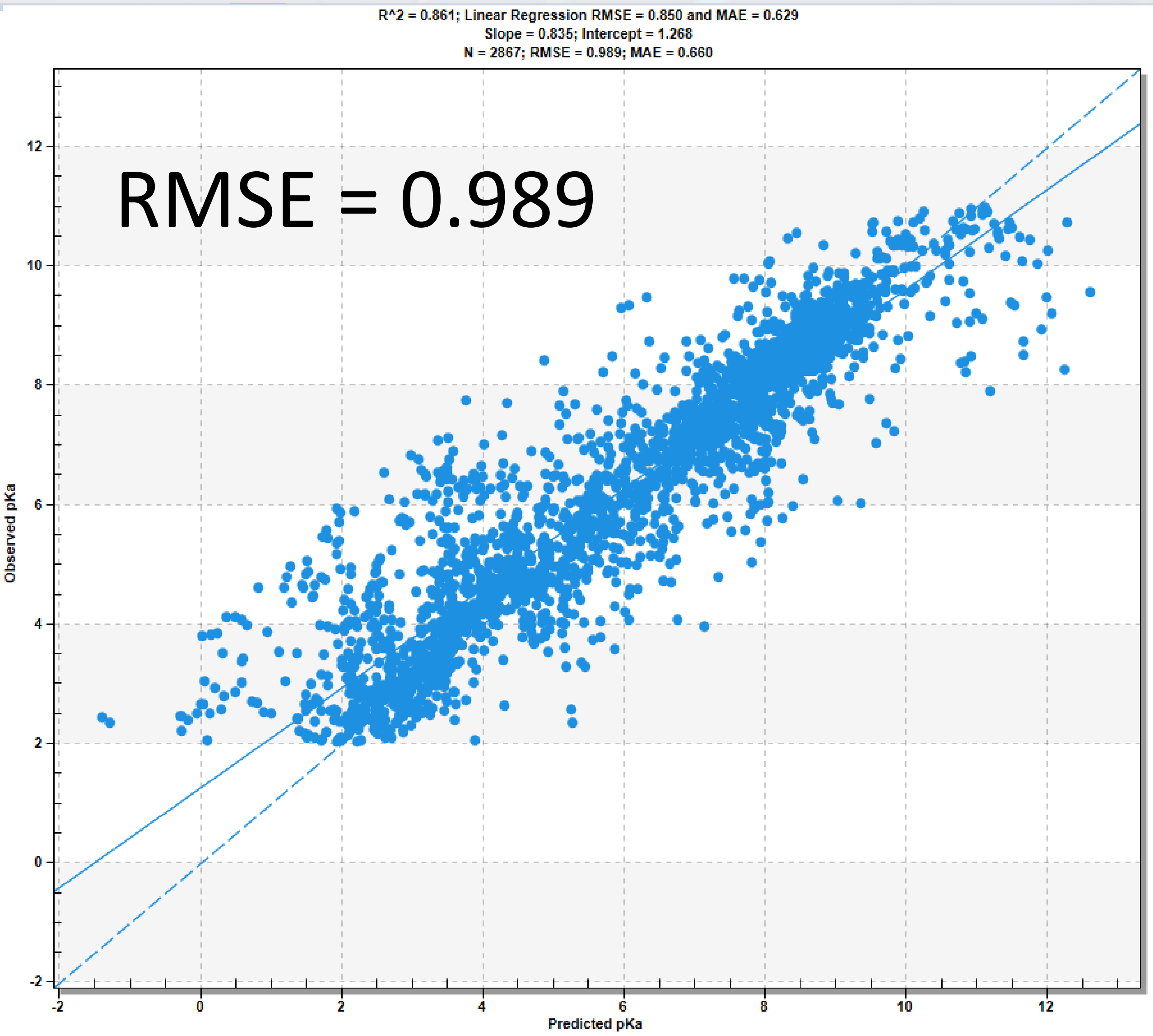
"v 7.0" and "v 10.5" models in external testing

- At Partner #2 site. "v 7.0" RMSE = 1.133, "v 10.5" RMSE = 0.989



R^2 = 0.814; Linear Regression RMSE = 0.983 and MAE = 0.747
Slope = 0.816; Intercept = 1.409
N = 2867; RMSE = 1.133; MAE = 0.805

RMSE = 1.133



R^2 = 0.861; Linear Regression RMSE = 0.850 and MAE = 0.629
Slope = 0.835; Intercept = 1.268
N = 2867; RMSE = 0.989; MAE = 0.660

RMSE = 0.989

Conclusions

- The chemical space covered by the new S+pKa model has been significantly expanded.
- Prediction accuracy has been improved.
- Partner #1 was very much forthcoming with a data set representative of their chemical space and reaped sizable rewards. Moreover, "changing input tautomer for some biggest outliers improved predictions".
- Partner #2 delivered much less data and the set’s place in their chemical space remains uknown.
- From our side questions were raised regarding validity of some spectrophotometric measurements. We are awaiting answers from both partners.

<u>And now a new message from the Little Bird</u>:

In January 2023 we have received thousands of pK_a data points from a new industrial Partner #3.

Stay tuned...