Supplementary Information

for

IPECnet: ML model for predicting the area of water solubility of interpolyelectrolyte complexes

*Ilya V. Grigoryan[1,2], Liubov A. Antiufrieva[3], Anna P. Grigoryan[1,4], Vladislava A. Pigareva[,6], Generalov A. Evgenii[1], Gennady B. Khomutov[1,2], Andrey V. Sybachin[5]**

[1]Physics Department of Lomonosov Moscow State University, Russia, Moscow, 199991, Leninskie Gory, 1-2

[2]Kotelnikov Institute of Radioengineering and Electronics, Russian Academy of Sciences, Moscow, 125009 Russia

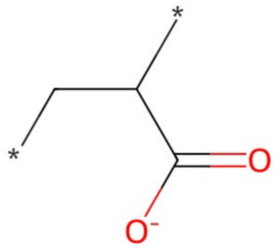
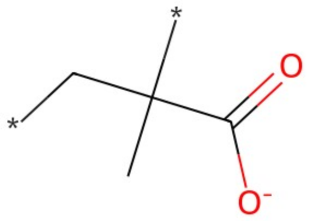
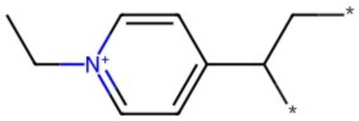
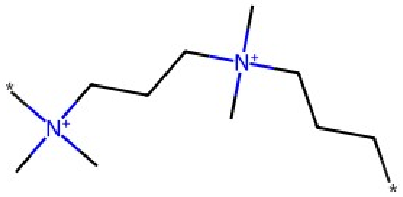
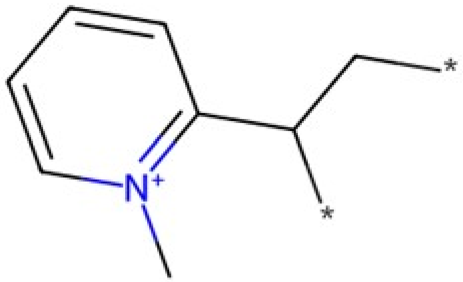
[3]Skolkovo Institute of Science and Technology, the territory of the Skolkovo Innovation Center, Bolshoy Boulevard, 30, bld. 1, Moscow 121205, Russia

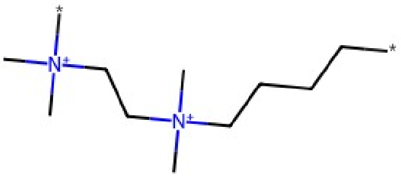
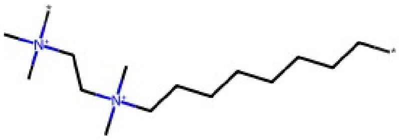
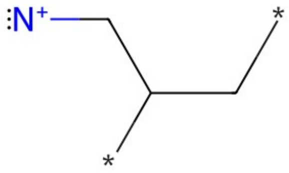
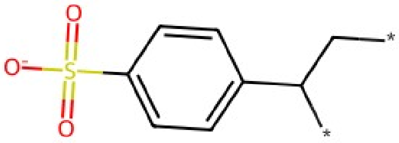
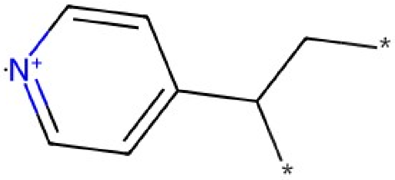
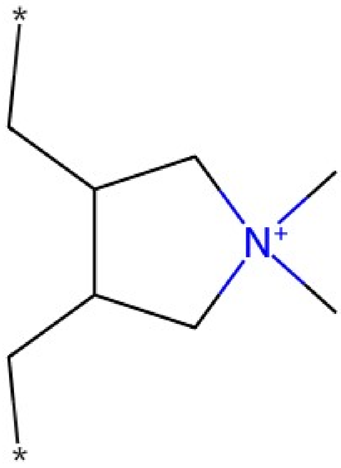
[4]Faculty of Space Research of Lomonosov Moscow State University, Moscow, Russia, 119991, Leninskiye Gory, 1-52

[5]Chemistry Department of Lomonosov Moscow State University, Russia, Moscow, 199991, Leninskie Gory, 1-2

[6]A. N. Nesmeyanov Institute of Organoelement compounds Russian Academy of Sciences, Russia, Moscow, +119334, Vavilova St., 28, bld. 1.

Table S1. The structures of monomers of polymers studied in the work. Visualizations are obtained using the RDKit library. The chemical structure of the monomer is presented in the pSMILES format

PC/PA index	Polymer	
1	Sodium poly(acrylate) [*]C(C[*])C(=O)[O-]	
2	Sodium poly(methacrylate) [*]C(C)(C[*])C(=O)[O-]	
3	Poly-N-ethyl-4-vinylpyridinum bromide [*]CC([*])c1cc[n+](CC)cc1	
4	3,3-ionene bromide [*][N+](C)(C)CCC[N+](C)(C)CCC[*]	
5	Poly-N-methyl-2-vinylpyridinum [*]CC([*])c1[n+](C)c(C)ccc1	

PC/PA index	Polymer	
6	2,4-ionene bromide [*][N+](C)(C)CC[N+](C)(C)CCCC[*]	
7	2,8-ionene bromide [*][N+](C)(C)CC[N+](C)(C)CCCCCCCC[*]	
8	Poly(allylamine) hydrochloride [*]CC([*])C[N+]	
9	Sodium poly(styrenesulfonate) [*]CC([*])c1ccc(S(=O)(=O)[O-])cc1	
10	Poly(4-vinylpyridinium) [*]CC([*])c1cc[n+]cc1	
11	Poly(diallyldimethylammonium chloride) solution == PDADMAC C1[n+](C)(C)CC(C[*])C1C[*]	

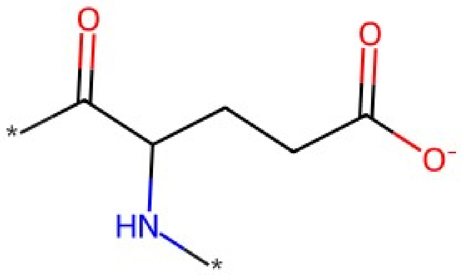
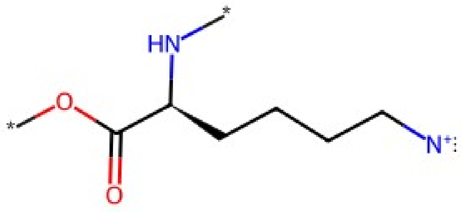
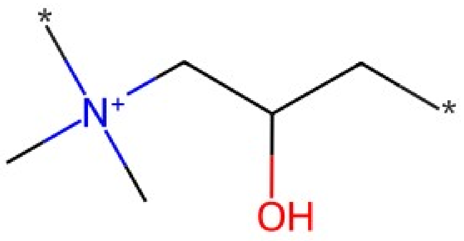
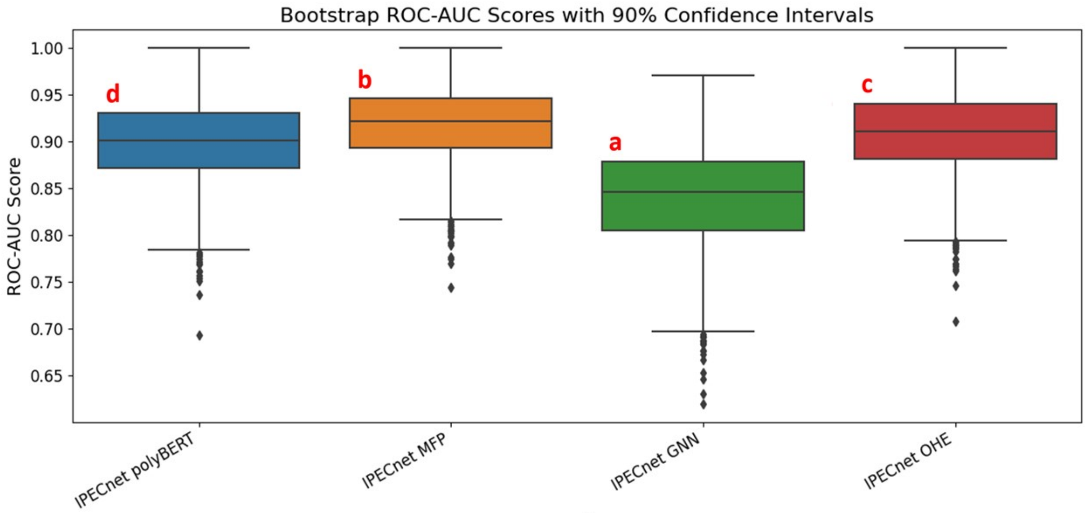
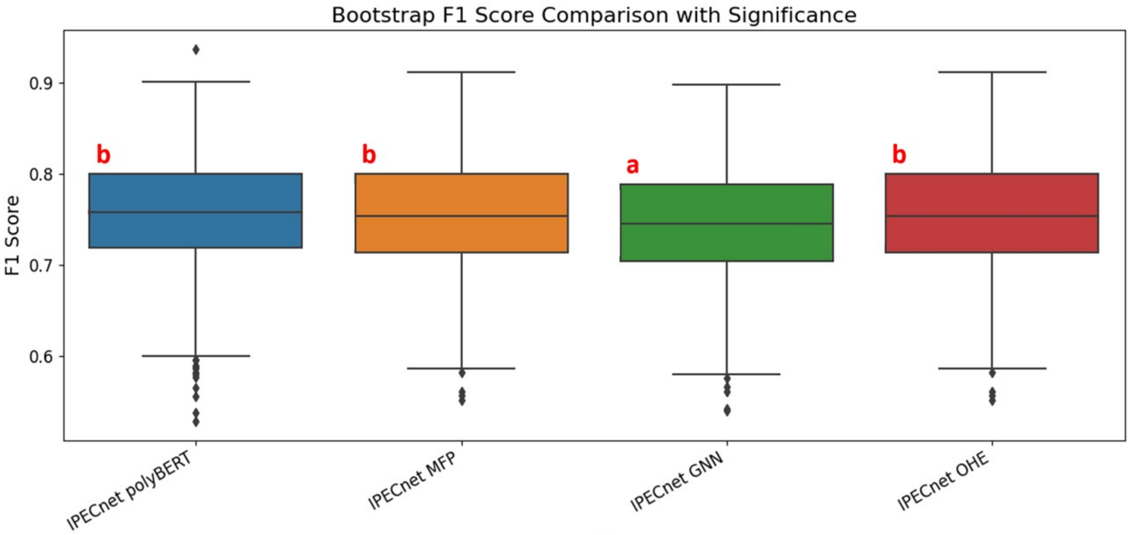
PC/PA index	Polymer	
12	Poly-L-glutamic acid sodium salt == PGA-Na [*]C(C(CCC([O-])=O)N[*])=O	
13	Poly-L-lysine hydrobromide == polylysHBr C(CC[N+])C[C@@H](C(=O)O[*])N[*]	
14	Hyperbranched Kaustamin == FL [N+]([*])(CC(O)C[*])(C)C	

Figure S1. The Bootstrap AUC-score, F1-score and Accuracy with 90% confidence intervals for the models described above. Statistical significance according to the nonparametric Mann-Whitney U test is indicated in small Latin letters directly next to the graphs. "IPECnet polyBERT" – original "IPECnet" model described in article with polyBERT as molecular embedding, "IPECnet MFP" – "IPECnet" with Morgan Fingerprint as molecular embedding instead of polyBERT, "IPECnet GNN" - "IPECnet" with graph neural network as molecular embedding instead of polyBERT, "IPECnet OHE" - "IPECnet" with one-hot encoding as molecular embedding instead of polyBERT


Bootstrap ROC-AUC Scores with 90% Confidence Intervals
ROC-AUC Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
d
b
a
c
IPECnet polyBERT
IPECnet MFP
IPECnet GNN
IPECnet OHE


Bootstrap F1 Score Comparison with Significance
F1 Score
0.9
0.8
0.7
0.6
b
b
a
b
IPECnet polyBERT
IPECnet MFP
IPECnet GNN
IPECnet OHE

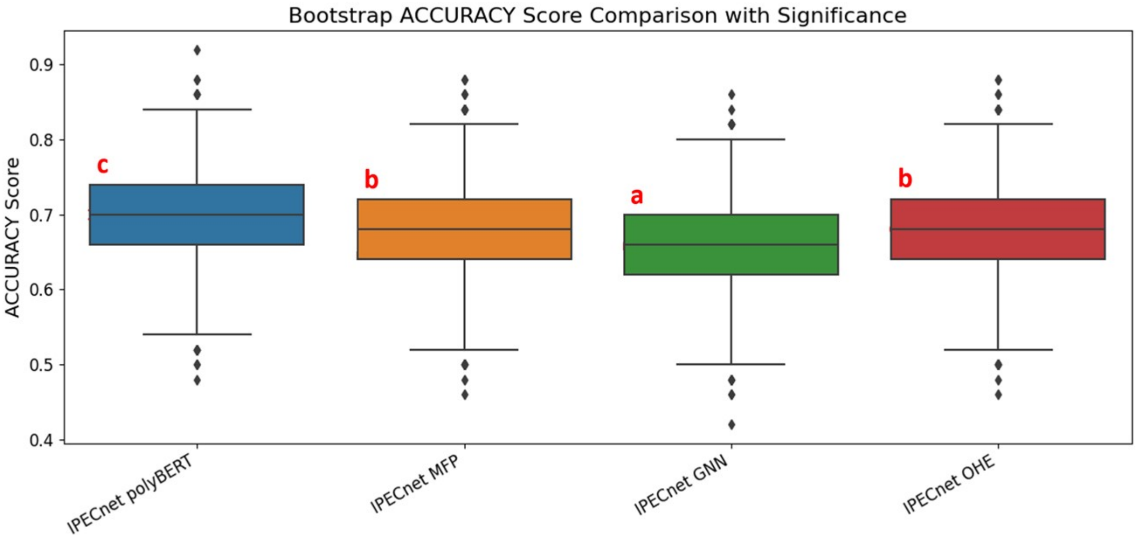

Bootstrap ACCURACY Score Comparison with Significance
ACCURACY Score
0.9
0.8
0.7
0.6
0.5
0.4
c
b
a
b
IPECnet polyBERT
IPECnet MFP
IPECnet GNN
IPECnet OHE

Figure S2. Trade-offs analysis between model complexity performance for “No chemistry model”, “IPECnet light”, “IPECnet” and “IPECnet without chemistry”

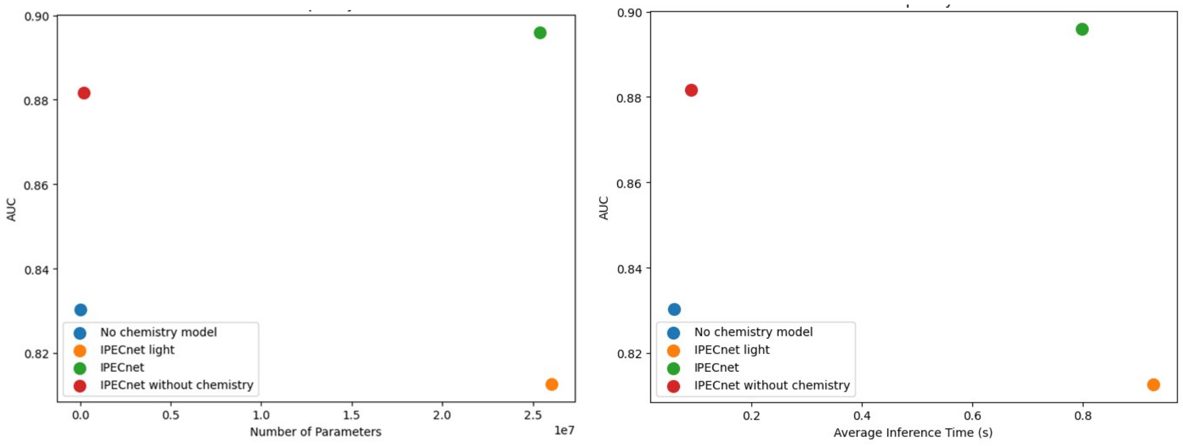


Figure S3. The results of comparing the “IPECnet” architecture model with various chemical embeddings (polyBERT, MorganFingerptint, GNN, OHE) trained on the initial and augmented datasets. As part of the application of the augmentation technique in our task, we examined samples increased in size by 4, 16 and 36 times. The results of the models trained on such datasets are marked on the graphs below as x4, x16 and x36, respectively. The method of compiling datasets using augmentation techniques is such that x2 means that each SMILES representation of GPE and HPE has been doubled, thereby increasing the total size of the dataset to x4. The rest of the changes were made according to the same rule. Augmentation technology was implemented based on [https://arxiv.org/abs/1703.07076].

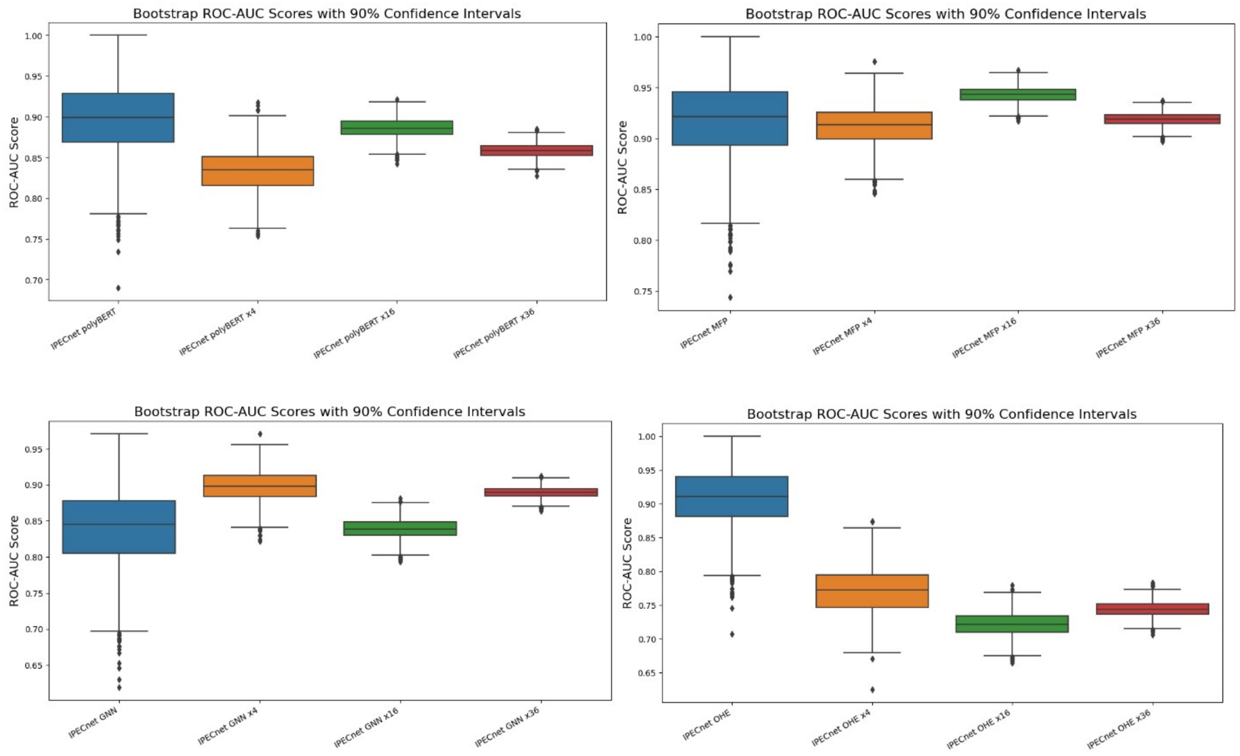
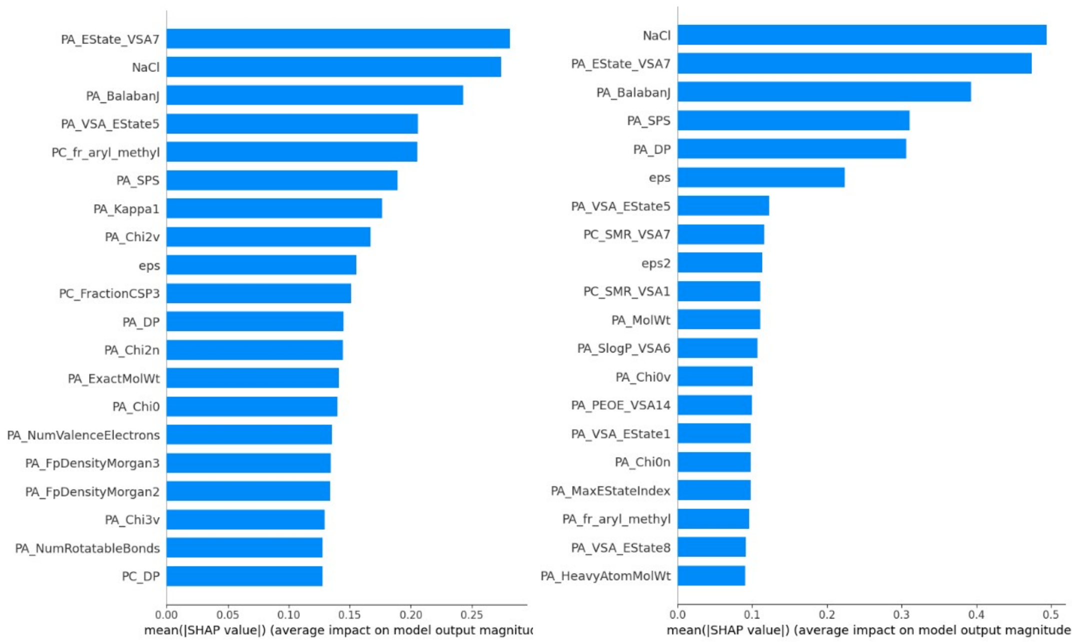


Figure S4. Average impact on model output magnitude for 20 most important (by mean (|SHAP value|) features for “No chemistry model” (left) and “IPECnet without chemistry” (right). “PA_” features are properties related to the HPE polymer present in excess in the system. “PC_” features are properties related to the GPE polymer present in the defect in the system. Features designations are available on GitHub.


PA_EState_VSA7
NaCl
PA_BalabanJ
PA_VSA_EState5
PC_fr_aryl_methyl
PA_SPS
PA_Kappa1
PA_Chi2v
eps
PC_FractionCSP3
PA_DP
PA_Chi2n
PA_ExactMolWt
PA_Chi0
PA_NumValenceElectrons
PA_FpDensityMorgan3
PA_FpDensityMorgan2
PA_Chi3v
PA_NumRotatableBonds
PC_DP
0.00
0.05
0.10
0.15
0.20
0.25
mean(|SHAP value|) (average impact on model output magnitu
NaCl
PA_EState_VSA7
PA_BalabanJ
PA_SPS
PA_DP
eps
PA_VSA_EState5
PC_SMR_VSA7
eps2
PC_SMR_VSA1
PA_MolWt
PA_SlogP_VSA6
PA_Chi0v
PA_PEOE_VSA14
PA_VSA_EState1
PA_Chi0n
PA_MaxEStateIndex
PA_fr_aryl_methyl
PA_VSA_EState8
PA_HeavyAtomMolWt
0.0
0.1
0.2
0.3
0.4
0.5
mean(|SHAP value|) (average impact on model output magnitude