Supplementary Information

高质量的材料科学文本挖掘数据集构建方法*

刘悦 [1)4)] 刘大晖 [1)] 葛献远 [1)] 杨正伟 [1)] 马舒畅 [1)] 邹喆义 [5)] 施思齐 [2)3)]

1) (上海大学计算机工程与科学学院，上海 200444)

2) (上海大学材料科学与工程学院，上海 200444)

3) (上海大学材料基因组工程研究院，上海 200444)

4) (上海市智能计算系统工程技术研究中心，上海 200444)

5) (湘潭大学材料科学与工程学院，湘潭 411105)

表 S1 NASICON 实体识别数据集的注释示例

Table S1. Examples of annotation for entity recognition dataset of NASICON.

实体类型	示例
Composition	The bond valence sum (BVS) of 1.08 is in good agreement with the value expected for **Na+**.
Structure	The results from both measurements can be correlated only, if at room temperature Na+ is located in the interstitial sites and if the transition of these ions to the **Na2 site** occurs at high temperature.
Property	Conductivity measurements in the range 30 − 350 <sYm> reveal the **activation energy** of 0.3 eV for Na+ conduction but **conductivity** values were found to change with temperature of sample preparation.
Processing	**Regrinding** and **reheating** the mix results in very slow incorporation of the free ZrO_2, probably by replacement of Na^+ from the zirconium sites.
Application	It is shown that $Na_3TiP_3O_9N$ can reversibly cycle Na-ions in a manner suitable for **secondary batteries**, and that the volume changes on Na removal are remarkably small (<1%) with respect to other known Na–ion.
Characterization	The **XRD diffraction pattern** of $Na_3MnTi(PO_4)_3$ can be indexed into a rhombohedral NASICON type unit cell with the R3c space group.
Feature	**Single crystal** x-ray analysis is used to identify the composition NaZr2P3012 and to refine its structure, which is rhombohedral.
Condition	The anisotropy of the thermal vibrations of sodium atoms in $NaSn_2(PO_4)_3$ at **room temperature** is described by two different flattened ellipsoids.

表 S2 NASICON 关系抽取数据集的注释示例

Table S2. Example of annotation for relational extraction dataset of NASICON.

关系类型	示例
Cause-Effect	1. (*Composition-Property*, *Property-Property*) In the system Na(1+*x*)Zr2Si*x* P(3-*x*)O12 , on the other hand, the introduction of excess **Na ions**(1) causes electrostatic **Na+-Na+ interactions**(2) that can lower the **activation energy**(3) even though transport must be via a Na1 site. 2. (*Structure-Property*) The increase of the **M1 site size**(1) in Na2SnFe(PO4)3 is accompanied by oxygen displacements perpendicular to the c axis which gives rise to **rotation**(2) of the PO4 tetrahedra and leads to a **distortion**(3) of the Sn(Fe)(1-*x*)(PO4)3 framework.
Component-Whole	1. (*Composition-Composition*) The original Na super ionic conductors NASICON materials are solid solutions derived from **NaZr2P3O12**(1) by partial replacement of **P**(2) by **Si**(3) with extra **Na**(4) to balance the charges.
Feature-Of	1. (*Composition-Property*) The calculated BVS value 5.22 shows that the **As5 cation**(1) is also slightly **over bonded**(2). 2. (*Composition-Structure*) The present paper reports on the **crystal structure** and vibrational spectra of **NaZr2(AsO4)3**.
Located-Of	1. (*Composition-Structure*) The **Na cation**(1) is located on the **6b position**(2) with a trigonal antiprismatic coordination and enhanced anisotropic displacement parameters.
Instance-Of	1. (*Property-Property*) The monochromator is a crystal of Ge that selects a **wavelength**(1) of **1.594 A**(2). 2. (*Feature-Feature*) X-ray powder diffraction shows that the **phosphates**(1) belong to the **NZP type**(2). 3. (*Structure-Structure*) The four **P-O bonds**(1) in the near regular **tetrahedron**(2) (point symmetry2) range from 1.524 to 1.525 A, with O-P-O angles deviating by no more than 1.5 <sYm> from the ideal 109.5 <sYm> tetrahedral angle (Table III).
Condition-On	1. (*Processing-Condition*) For each crystal structure determination, data are collected using **Mo Ka radiation**(1) up to **29-650**(2). 2. (*Property-Condition*) The **disorder**(1) is larger at **100 K**(2) than at **295 K**(3).
Method-Of	1. (*Characterization-Property*) The **Rietveld plots**(1) represent a good structure fit between observed and calculated intensity with satisfactory ***R*-factors**(2). 2. (*Characterization-Property*) Its **anisotropic thermal expansion**(1) has been calculated from high temperature **X-ray diffraction**(2), and it is linear ina range from room temperature up to 800 <sYm>. 3. (*Processing-Condition*) **Good crystals**(1) can, however, be obtained after **tempering**(2) in platinum crucible for several weeks at 11000 <sYm>.
Other	1. (*Condition-Property*) At **room temperature**(1) no **diffuse intensity**(2) is observed.

表 S3 MatBERT-BiLSTM-CRF 模型参数

Table S3. Parameters of MatBERT-BiLSTM-CRF.

参数名称	值
批量大小（Batch size）	32
迭代周期（Epoch）	100
单词向量维度（Word vector dimension）	768
LSTM 单元维度（LSTM unit dimension）	128
丢包率（Dropout rate）	0.1
学习率（Learning rate）	3×10^{-5}
优化器（Optimizer）	AdamW
提前停止耐力值（Early stopping patience）	3
最大句子长度（Max sentence length）	75

表 S4 不同机器学习模型的激活能预测结果[33]

Table S4. Results of different machine learning models[33].

Models	$\mathrm{Dataset}_{31}$			$\mathrm{Dataset}_{45}$		
	RMSE	MAPE	R^2	RMSE	MAPE	R^2
LASSO	0.09	0.06	0.86	0.06	0.04	0.94
GPR	0.09	0.06	0.86	0.05	0.04	0.96
Ridge	0.09	0.06	0.86	0.05	0.04	0.95
SVR	0.10	0.07	0.84	0.07	0.06	0.92
KNN	0.11	0.07	0.80	0.09	0.06	0.87
RF	0.10	0.06	0.83	0.05	0.04	0.96

